# International Journal of Computer Science & Emerging Technologies

ISSN 2044-6004 (Online)

Volume 1 Issue 4

December, 2010

www.ijcset.excelingtech.co.uk

ijcset@gmail.com

# IJCSET BOARD MEMBERS

- **Nafiz Imtiaz Bin Hamid**, Bangladesh
- **Jasvir Singh**, India
- **Manas Ranjan Biswal**, India
- **Ratnadeep R. Deshmukh**, India
- **Sujni Paul**, India
- **Mousa Demba**, Saudi Arabia
- **Yasser Alginahi**, Saudi Arabia
- **Tarun Kumar**, India
- **Alessandro Agostini**, Saudi Arabia

# TABLE OF CONTENTS

# Improved Fuzzy C-Means Clusters With Ant Colony Optimization

[1]C.Immaculate Mary, [2]Dr. S.V. Kasmir Raja

[1]Associate professor of Computer Science ,Sri Sarada College, Salem-636016, Tamil Nadu, India.

[2]Dean (Research), S.R.M. University , Chennai, Tamil Nadu, India.

E-mail: cimmaculatemary@gmail.com

***Abstract:*** Cluster analysis aims at identifying groups of similar objects and, therefore helps to discover distribution of patterns and interesting correlations in large data sets. These methods are not only major tools to uncover the underlying structures of a given data set, but also promising tools to uncover local input-output relations of a complex system. Fuzzy C-means (FCM) is one of the most widely used fuzzy clustering algorithms in real world applications. However there are two major limitations that exist in this method. The first is that a predefined number of clusters must be given in advance. The second is that the FCM technique can get stuck in sub-optimal solutions. In this paper, we have proposed an ant colony algorithm to improve the clusters obtained from fuzzy c-means clustering. The proposed algorithm is tested in medical domain and the results show that post processing refinement of clusters improves the cluster quality.

***Keywords***: Fuzzy C-Means, Ant Colony Optimization, Cluster Refinement

## 1. Introduction

Clustering is one of the most useful tasks in data mining process for discovering groups and identifying interesting distributions and patterns in the underlying data. Clustering problem is about partitioning a given data set into groups (clusters) such that the data points in a cluster are more similar to each other than points in different clusters (Guha, et al., 1998). For example, consider a retail database records containing items purchased by customers. A clustering procedure could group the customers in such a way that customers with similar buying patterns are in the same cluster. Thus, the main concern in the clustering process is to reveal the organization of patterns into “sensible” groups, which allow us to discover similarities and differences, as well as to derive useful conclusions about them. This idea is applicable in many fields, such as life sciences, medical sciences and engineering. Clustering may be found under different names in different contexts, such as unsupervised learning (in pattern recognition), numerical taxonomy (in biology, ecology), typology (in social sciences) and partition (in graph theory) (Theodoridis & Koutroubas, 1999). In the clustering process, there are no predefined classes and no examples that would show what kind of desirable relations should be valid among the data that is why it is perceived as an unsupervised process (Berry & Linoff, 1996). On the other hand, classification is a procedure of assigning a data item to a predefined set of categories (Fayyad, et al., 1996). Clustering produces initial categories in which values of a data set are classified during the classification process.

Clustering analysis is the main component of unsupervised techniques. Recently various algorithms for clustering large data sets and streaming data sets have been proposed (Pal and Bezdek 2002, Ramakrishnan and Livny 1996, Bradley et al., 1998, Farnstrom et al., 2000, Guha et al., 1998, Ng and Han 2002, Gupta and Grossman 2004, O’Callaghan et al., 2002). The focus has been primarily either on sampling (Pal and Bezdek 2002, Guha et al., 1998, Ng and Han 2002, Hathaway and Bezdek, 2006) or incrementally loading partial data, as much as can fit into memory at one time. The incremental approach (Bradley et al., 1998, Farnstrom et al., 2000, Gupta and Grossman 2004, O’Callaghan et al., 2002) generally keeps sufficient statistics or past knowledge of clusters from a previous run of a clustering algorithm in some data structures and uses them in improving the model for the future.

Clustering can also be performed in two different modes: crisp and fuzzy. In crisp clustering, the clusters are disjoint and non-overlapping in nature. Any pattern may belong to one and only one class in this case. In case of fuzzy clustering, a pattern may belong to all the classes with a certain fuzzy membership grade (Jain et al., 1999). A common fuzzy clustering algorithm is the Fuzzy C-Means (FCM), an extension of classical C Means algorithm for fuzzy applications (Bezdeck et al., 1984). The FCM method (Canno et al., 1986, Kamel and Selim, 1994), suffer several difficulties: a) sensitive to the initialization; b) inability to find a global minimum and; c) difficulty of deciding how many clusters exist. Since FCM's performance depends on selected metrics, it will depend on the feature- weights which are incorporated into the Euclidean distance. Wang et al., (2004) try to adjust these feature weights to improve FCM's performance. Wang and Garibaldi (2005) proposed an alternative fuzzy clustering algorithm, Simulated Annealing Fuzzy Clustering (SAFC) that improves the cluster quality. Various algorithms (Cheng et al., 1998, Eschrich et al., 2003, Altman 1999, Kolen and Hutcheson, 2002, Borgelt and Kruse, 2003), for speeding up clustering have also been proposed.

As seen in the literature, the researchers contributed only to reduce the time complexity or to accelerate the algorithm ; there is no contribution in cluster refinement. In this study, we propose a new algorithm to improve the fuzzy c-means. In this proposed algorithm, an ant colony optimization algorithm is applied to refine the cluster to improve the quality. The paper is organized as follows: section 2 presents the general fuzzy c-means algorithm. Section 3 discusses the proposed cluster refinement algorithm with ant colony optimization. Section 4 presents the results and the work is concluded in section 5.

## 2. Standard Fuzzy C-Means Clustering

The FCM algorithm, also known as Fuzzy ISODATA, is one of the most frequently used methods in pattern recognition. It is based on minimization of the given objective function to achieve good classifications.

$$J(U,V)=\sum_{i=1}^{n}\sum_{j=1}^{c}(\mu_{ij})^m \| x_i - v_j \|^2$$

J(U,V) is a squared error clustering criterion, and solutions of minimization of (1) are least-squared error stationary points of J(U,V). The expression, X = {$x_1, x_2, \ldots, x_n$} is a collection of data, where n is the number of data points. V = { $v_1, v_2, \ldots, v_c$} is a set of corresponding cluster centers in the data set **X**, where c is the number of clusters. $\mu_{ij}$ is the membership degree of data $x_i$ to the cluster centre $v_j$. Meanwhile, $\mu_{ij}$ has to satisfy the following conditions:

$$\mu_{ij}\in[0,1], \forall i=1,\ldots,n, \forall j=1,\ldots,c$$

$$\sum_{j=1}^{c}\mu_{ij}=1, \forall i=1,\ldots,n$$

Where U = $(\mu_{ij})_{n*c}$ is a fuzzy partition matrix, $\| x_i - v_j \|$ represents the Euclidean distance between $x_i$ and $v_j$, parameter m is the "fuzziness index" and is used to control the fuzziness of membership of each datum in the range m ∈ [1,∞] . In this experimentation the value of m=2.0 was chosen. Although there is no theoretical basis for the optimal selection of m, this has been chosen because the value has been commonly applied within the literature. The FCM algorithm is described in, for example, and can be performed by the following steps:

1. Initialize the cluster centers V = { $v_1, v_2, \ldots, v_c$}, or initialize the membership matrix $\mu_{ij}$ with random value and make sure it satisfies the above conditions and then calculate the centers.
2. Calculate the fuzzy membership $\mu_{ij}$ using

$$\mu_{ij}=\frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{ik}}\right)^{\frac{2}{m-1}}}$$

$$where, d_{ij}=\| x_i - v_j \|, \forall i=1,\ldots,n,$$
$$\forall j=1,\ldots c$$

3. Compute the fuzzy centers $v_j$ using

$$v_j=\frac{\sum_{i=1}^{n}(\mu_{ij})^m x_i}{\sum_{i=1}^{n}(\mu_{ij})^m}, \forall j=1,\ldots,c$$

4. Repeat steps (2) and (3) until the minimum J value is achieved.
5. Finally, defuzzification is necessary to assign each data point to a specific cluster (i.e. by setting a data point to a cluster for which the degree of the membership is maximal).

## 3. Aco Based Cluster Refinement

Ant-based clustering and sorting was originally introduced for tasks in robotics by Deneubourg et al. (1991). Lumer and Faieta (1994) modified the algorithm to be applicable to numerical data analysis, and it has subsequently been used for data-mining (Lumer and Faieta (1994), graph-partitioning (Kuntz and Snyers 1994, Kuntz and Snyers, 1999, Kuntz et al., 1998) and text-mining (Handl and Meyer 2002, Hoe et al., 2002, Ramos V., and Merelo, 2002).

Such ant-based methods have shown their effectiveness and efficiency in some test cases (Handl et al., 2003). However, the ant-based clustering approach is in general immature and leaves big space for improvements. With these considerations, however, the standard ant-based clustering performs well; the algorithm consists of lot of parameters like pheromone, agent memory, number of agents, number of iterations and cluster retrieval etc. For these parameters more assumptions have been made in the previous works. So far, ants are used to cluster the data points. Here is the first time; we have used ants to refine the clusters. The clusters from the above section are considered as input to this ACO based refinement step.

The basic reason for our refinement is, in any clustering algorithm the obtained clusters will never gives us 100% quality. There will be some errors known as misclustered. That is, a data item can be wrongly clustered. These kinds of errors can be avoided by using our refinement algorithm.

In our proposed method, three ants are used to refine the clusters. These ants are allowed to go for a random walk on the clusters. Whenever it crosses a cluster, it will pick an item from the cluster and drop it into another cluster while moving. And then the quality of the clusters is compared with the drop probability calculated from two cluster validity indexes; Partition Coefficient (PC) and Partition Entropy (PE) are defined as (Bezdek 1981):

$$PC = \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{c}\mu_{ij}^{2}$$

$$PE = -\frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{c}\mu_{ij}\log_{a}(\mu_{ij})$$

PC and PE is used to measure the fuzziness of the fuzzy partition matrix, the lower the fuzziness of a partition is, the larger the PC value (or the smaller the PE value). From these validity indexes the drop probability is calculated as:

$$P_d = PE / PC$$

If $P_d$ is smaller than the previous iteration, then the drop is made permanent and next iteration is continued with the changed cluster indexes. Otherwise, the next iteration is continued with the old cluster indexes.

This random walk is repeated for N number of times. From the following section, it is shown that our refinement algorithm improves the cluster quality. The algorithm is given as:

1. Initialize the cluster centers V = { $v_1$, $v_2$,…, $v_c$}, or initialize the membership matrix $\mu_{ij}$ □with random value and make sure it satisfies the above conditions and then calculate the centers.
2. Calculate the fuzzy membership $\mu_{ij}$
3. Compute the fuzzy centers $v_j$
4. Repeat steps (2) and (3) until the minimum J value is achieved.
5. Finally, defuzzification is necessary to assign each data point to a specific cluster.
6. Ant based refinement
    a. Input the clusters from fuzzy c-means.
    b. For i = 1 to N do
        i. Let the ants go for a random walk to pick the items
        ii. Drop the items into some other cluster.
        iii. Check whether the quality improving or not by calculating PE and PC.
        iv. If it improves then drop the items permanently.
    c. Repeat

## 4. Results

Clustering validity is a concept that is used to evaluate the quality of clustering results. If the number of clusters is not known prior to commencing an algorithm, the clustering validity index may be used to find the optimal number of clusters (Rezaee et al., 1998). This can be achieved by evaluating all of the possible clusters with the validity index and then the optimal number of clusters can be determined by selecting the minimum value of the index. Many clusters validation indices have been developed in the past. In the context of fuzzy methods, some of them only use the membership values of a fuzzy cluster of the data, such as the partition coefficient and partition entropy. The advantage of this type of index is that it is easy to compute but it is only useful for the small number of well-separated clusters. Furthermore, it also lacks direct connection to the geometrical properties of the data. In order to overcome this problem Xie and Beni defined a validity index which measures the compactness and separation of clusters (Xie and Beni, 1991). In this paper, the Xie-Beni index has been chosen as the cluster validity measure because it has been shown to be able to detect the correct number of clusters in several experiments (Pal and Bezdek, 1995). Xie-Beni validity is the combination of two functions. The first calculates the compactness of data in the same cluster and the second computes the separateness of data in different clusters. Let S represent the overall validity index, $\pi$ □be the compactness and s be the separation of the fuzzy c partition of the data set. The Xie-Beni validity can now be expressed as:

$$S = \pi / s$$

Where, $$\pi = \frac{\sum_{j=1}^{c}\sum_{i=1}^{n}\mu_{ij}^{2} \| x_i - v_j \|^2}{n}$$

and $$s = (d_{min})^2$$

$d_{min}$ is the minimum distance between cluster centres, given by $\min_{ij} \| v_i - v_j \|$. Smaller values of $\pi$ indicate that the clusters are more compact and larger values of s indicate the clusters are well separated. Thus a smaller S reflects that the clusters have greater separation from each other and are more compact. The following tables present the results, shows that our proposed method outperforms than the standard method.

**Table 1. Performance of Clustering for Wisconsin Breast Cancer Dataset**

| | **Fuzzy C-Means** | **Refined Fuzzy C-Means with ACO** |
|---|---|---|
| No. of Classes | 2 | 2 |
| No. of Clusters | 2 | 2 |
| Partition Coefficient | 0.8268 | 0.9861 |
| Partition Entropy | 0.2985 | 0.0392 |
| Xie-Beni Index | 3.3862 | 2.9562 |

**Table 2. Performance of Clustering for Dermatology Dataset**

| | **Fuzzy C-Means** | **Refined Fuzzy C-Means with ACO** |
|---|---|---|
| No. of Classes | 6 | 6 |
| No. of Clusters | 6 | 6 |
| Partition Coefficient | 0.9433 | 0.9847 |
| Partition Entropy | 0.1412 | 0.0554 |
| Xie-Beni Index | 1.1803 | 3.4420 |

## 5. Conclusion

Cluster analysis is one of the major tasks in various research areas. However, it may be found under different names in different contexts such as unsupervised learning in pattern recognition, taxonomy in biology, partition in graph theory. The clustering aims at identifying and extract significant groups in underlying data. Thus based on a certain clustering criterion the data are grouped so that data points in a cluster are more similar to each other than points in different clusters. Since clustering is applied in many fields, a number of clustering techniques and algorithms have been proposed and are available in literature. In this paper, an ant colony algorithm is presented to improve the cluster from fuzzy c-means clustering. The performance is compared with the standard fuzzy c-means clustering; the result shows the proposed method performs better than the standard method.

### REFERENCES

1. Altman, D., 1999. Efficient Fuzzy Clustering of Multi-spectral Images, FUZZ-IEEE.
2. Bensaid, A., J. Bezdek, L.O. Hall, and L.P. Clarke, 1996. Partially Supervised

Clustering for Image Segmentation, Pattern Recognition, V. 29, No. 5, pp. 859-871.

1. Berry, Gordon Linoff, 1996. Data Mining Techniques For marketing, Sales and Customer Support. John Willey & Sons, Inc.
2. Bezdek, J. C., Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, 1981.
3. Bezdeck J.C, Ehrlich R., Full W., 1984. FCM:Fuzzy C-Means Algorithm. Computers and Geoscience 1984.
4. Borgelt, C., and R Kruse, 2003. Speeding Up Fuzzy Clustering with Neural Network Techniques. Fuzzy Systems. V. 2, pp. 852–856.
5. Bradley, P.S., U Fayyad, and C Reina, 1998. Scaling Clustering Algorithms to Large Databases. In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, KDD-1998. pp., 9–15.
6. Canno, R.L., Dave, J.V., Bezdek, J.C., 1986. Efficient Implementation of the Fuzzy C-Means Clustering Algorithm. IEEE Trans. Pattern Anal. Mach. Intel. vol. 8, pp. 248-255.
7. Cheng, T.W., Dmitry B. Goldgof, and Lawrence O. Hall, 1998. Fast Fuzzy clustering. Fuzzy Sets and Systems. V. 93, pp. 49–56.
8. Deneubourg J.L., Goss S., Franks, N. Sendova-Franks A., Detrain C., and Chétien L. 1991. The Dynamics of Collective Sorting: Robot-like Ants and Ant-like Robots. In Proceedings of the 1st International Conference on Simulation of Adaptive Behavior: From Animals to Animats,. MIT Press, Cambridge, MA, USA. Vol. 1, pp. 356-363.
9. Eschrich, S., J Ke, LO. Hall and DB. Goldgof, 2003. Fast Accurate Fuzzy Clustering through Data Reduction. IEEE Transactions on Fuzzy Systems, vol. 11, no. 2, pp. 262–270.
10. Farnstrom, F., J Lewis, and C Elkan, 2000. Scalability for Clustering Algorithms Revisited. ACM SIGKDD Explorations. vol. 2, pp. 51–57.
11. Fayyad, M. U., Piatesky-Shapiro, G., Smuth P., Uthurusamy, R. 1996. Advances in Knowledge Discovery and Data Mining. AAAI Press.
12. Guha, S., R Rastogi, and K Shim, 1998. CURE: An Efficient Clustering Algorithm for Large Databases. In Proceedings of ACM SIGMOD International Conference on Management of Data, pp. 73–84.
13. Guha, S., Rastogi, R., Shim K. 1998. CURE: An Efficient Clustering Algorithm for Large Databases, Published in the Proceedings of the ACM SIGMOD Conference.
14. Gupta, C., and R Grossman, 2004. GenIc: A Single Pass Generalized Incremental Algorithm for Clustering. Proceedings of the Fourth SIAM_ International Conference on Data Mining (SDM 04), pp. 22–24.
15. Handl J., Knowles J., and Dorigo M. 2003. On the Performance of Ant-based Clustering. In Design and Application of Hybrid Intelligent Systems, Frontiers in Artificial Intelligence and Applications,. Netherlands: IOS Press, vol. 104, pp. 204-213.
16. Handl J., and Meyer B. 2002. Improved ant-based clustering and sorting in a document retrieval interface. In Proceedings of the Seventh International Conference on Parallel Problem Solving from Nature, Springer-Verlag, Berlin, Germany, vol. 2439, pp. 913–923.
17. Hathaway, R. J., and JC. Bezdek, 2006. Extending Fuzzy and Probabilistic Clustering to Very Large Data Sets, Journal of Computational Statistics and Data Analysis.
18. Hoe K., Lai W., and Tai T. 2002. Homogeneous ants for web document similarity modeling and categorization. In Proceedings of the Third International Workshop on Ant Algorithms, Springer-Verlag, Heidelberg, Germany, vol. 2463, pp. 256–261.
19. Jain, A.K, Murty MN and Flynn PJ, 1999. Data clustering: a review. ACM Computing Surveys, vol. 31, no.3, pp. 264|323.
20. Kamel, M.S., Selim,S.Z. 1994. New Algorithms for Solving the Fuzzy Clustering Problem. Pattern Recognition. Vol. 27, pp. 421-428.
21. Kolen, J.F., and T Hutcheson, 2002. Reducing the Time Complexity of the Fuzzy C-Means Algorithm. IEEE Transactions on Fuzzy Systems. vol. 10, pp. 263–267.
22. Kuntz P., and Snyers D. 1994. Emergent colonization and graph partitioning. In Proceedings of the Third International

Conference on Simulation of Adaptive Behaviour: From Animals to Animats. MIT Press, Cambridge, MA, vol. 3, pp. 494–500.

23. Kuntz P., and Snyers D. 1999. New results on an ant-based heuristic for highlighting the organization of large graphs. In Proceedings of the 1999 Congress on Evolutionary Computation, IEEE Press, Piscataway, NJ, pp. 1451–1458.
24. Kuntz P., Snyers D., and Layzell P. 1998. A stochastic heuristic for visualizing graph clusters in a bi-dimensional space prior to partitioning. Journal of Heuristics. vol. 5, no. 3, pp. 327–351.
25. Kowalski, G., 1997. Information Retrieval Systems – Theory and Implementation. Kluwer Academic Publishers.
26. Larsen B., and Aone C. 1999. Fast and Effective Text Mining Using Linear-time Document Clustering. KDD-99, San Diego, California.
27. Lumer, E., and Faieta B. 1994. Diversity and adaptation in populations of clustering ants. In Proceedings of the Third International Conference on Simulation of Adaptive Behaviour: From Animals to Animats. MIT Press, Cambridge, MA, vol. 3, pp. 501–508.
28. Lumer, E., and Faieta B. 1995. Exploratory database analysis via self-organization.
29. Ng, R.T., and J Han, 2002. CLARANS: A Method for Clustering Objects for Spatial Data Mining. IEEE Transactions on Knowledge and Data Engineering. Vol. 14, no. 5, pp. 1003–1016.
30. O'Callaghan, L., N. Mishra, A. Meyerson, S. Guha, and R. Motwani, 2002. Streaming-Data Algorithms for High-Quality Clustering, Proceedings of IEEE International Conference on Data Engineering.
31. Pal, N.R., and JC. Bezdek, 1995. On cluster validity for the fuzzy c-means model. IEEE Trans. Fuzzy Systs., Vol. 3, pp. 370-379.
32. Pal, N.R., and JC. Bezdek, 2002. Complexity Reduction for "Large Image" Processing. IEEE Trans on Systems, Man, and Cyber., Part B vol. 32, no. 5, pp. 598–611.
33. Ramakrishnan, Z. R., M Livny, 1996. BIRCH: An Efficient Data Clustering Method for Very Large Databases. Tian ACM SIGMOD International Conference on Management of Data. pp. 103–114.
34. Ramos V., and Merelo JJ. 2002. Self-organized stigmergic document maps: Environments as a mechanism for context learning. In Proceedings of the First Spanish Conference on Evolutionary and Bio-Inspired Algorithms, Centro Univ. M´erida, M´erida, Spain, pp. 284–293.
35. Rezaee, M.R., BPF. Leieveldt, and JHC. Reiber, 1998. A New Cluster Validity Index for the Fuzzy C-Means. Pattern Recognition Letters. Vol. 19, pp. 237-246.
36. Theodoridis, S., Koutroubas, K. 1999. Pattern recognition, Academic Press.
37. Wang, X., J M. Garibaldi, 2005. Simulated Annealing Fuzzy Clustering in Cancer Diagnosis. Informatica, vol. 29, pp. 61–70.
38. Wang, X., Y Wang, L Wang, 2004. Improving fuzzy c-means clustering based on feature-weight learning. Pattern Recognition Letters. vol. 25, pp. 1123–1132.
39. Xie, X. L., and G. Beni, 1991. A validity measure for fuzzy clustering. IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 13, pp. 841-847.

# Blood Vessel Extraction Using Wiener Filter and Morphological Operation

V. Vijaya Kumari [1], Dr. N. Suriyanarayanan [2]

[1]Department of ECE, V.L.B. Janakiammal College of Engineering and Technology
Coimbatore, India
Email: ebinviji@rediffmail.com
[2]Department of Physics, Government College of Technology
Coimbatore, India ,
Email: esnsuri@yahoo.co.in

***Abstract:*** Diabetic retinopathy (DR) is a common retinal complication associated with diabetes. Along with optic disk and blood vessel of normal patients, the diabetic patient's retinal image has exudates. Depending on the severity of diabetics micro aneurysms and hemorrhages may also present. So in the diagnosis of diabetic retinopathy, the detection of exudates plays the fundamental role. Sometimes exudates and optic disk are similar in brightness, color and contrast. It is very important to differentiate them. To detect exudates correctly, optic disk should be detected first and then it should be masked. The blood vessel is extracted and the meeting point of the blood vessel is the center of the optic disk. In this the blood vessel is extracted using Wiener filter and morphological operation opening and closing. The peak signal to noise ratio is calculated for both the methods and are compared. The edge of the blood vessels are clearly detected by applying Laplacian and Gaussian operators and the thinning of blood vessel is done using morphological operator and smoothened for better clarity in the extracted blood vessel.

***Keywords:*** Diabetic retinopathy, micro aneurysms, optic disk, exudates

## 1. INTRODUCTION

The prevalence of Diabetic Retinopathy is high and the incidence is growing in step with worldwide increases in Diabetic Maculopathy. Diabetic screening programmes are necessary in addressing all of these factors when working to eradicate preventable vision loss in diabetic patients. When performing retinal screening for Diabetic Retinopathy, some of these clinical presentations are expected to be imaged. Diabetic retinopathy is globally the primary cause of blindness not because it has the highest incidence because it often remains undetected until severe vision loss occurs.

Advances in shape analysis, the development of strategies for the detection and quantitative characterization of blood vessel changes in the retina are therefore of great importance. Automated early detection of the presence of exudates can assist the ophthalmologists to prevent the spread of the disease more efficiently.

Direct digital image acquisition using fundus cameras combined with image processing and analysis techniques has the potential to enable automated diabetic retinopathy screening. The normal features of fundus images include optic disk, fovea and blood vessels. Exudates and hemorrhages are the main abnormal features which is the leading cause of blindness in the working age population.

Optic disk is the brightest part in the normal fundus images which can be seen as a pale, round or vertically slightly oval disk. The change in the shape, color or depth of the optic disk is an indicator of various ophthalmic pathologies especially for glaucoma.

Exudates are one of the most common occurring lesions in diabetic retinopathy. Exudates can be identified as areas with hard white or yellowish colors and varying sizes, shapes and locations near the leaking capillaries within the retina. The shape, brightness and location of exudates vary a lot among different patients.

Retinal images of human plays an important role in the detection and diagnosis of many eye diseases for ophthalmologists. Exudates can be identified by segmenting the blood vessels and removing it. Swelling of blood vessel can also be the symptom of diabetes. Chwialkowski et al [2] accomplish segmentation of blood vessels using multi resolution analysis based on wavelet transform.

The segmentation process is applied to the magnitude image and the velocity information from the phase difference image is integrated on the resulting vessel area to get the blood flow measurement. Vessel boundaries are localized by employing a multivariate scoring criterion to minimize the effect of imaging artifacts such as partial volume averaging and flow turbulence. Niki et al [7] describe their 3D blood vessel reconstruction and analysis method.

Vessel reconstruction is achieved on short scan cone-beam filtered back propagation reconstruction algorithm based on Gulberg and Zeng's work. Schmitt et al [10] combine thresholding with region growing technique to segment vessel tree in 3D in their work of determination of the contrast agent propagation in 3D rotational XRA image volumes. Poli and Valli [8] develop an algorithm to enhance and detect vessels in real time.

The algorithm is based on a set of multiple oriented linear filters obtained as linear combination of properly shifted Gaussian kernels. Figueiredo and Leitao [4] describe their non smoothing approach in estimating vessel contours in angiograms. This technique has two key features. First it does not smooth the image to avoid the distortions introduced by smoothing. Second it does not assume a constant background which makes the technique well suited for the non subtracted angiograms. Donizelli [3] combines mathematical morphology and region growing algorithms to segment large vessels from digital subtracted angiography images. Krissian et al [5] develop a multi scale model to extract and reconstruct 3D vessels from medical images.

The method uses a new response function which measures the contours of the vessels around the centerlines. It consists of three main steps. First the multi scale responses from discrete set of scales are computed. Second, the local extreme in multi scale response is extracted. Finally the skeleton of the local extreme is created and the result is visualized. Aylward et al [1] utilize intensity ridges to approximate the medial axes of tubular objects such as vessels. Fuzzy clustering is another approach to identify vessel segments. It uses linguistic descriptions like "vessel" and "non vessel" to track vessels in retinal angiogram images. One disadvantage of the vessel tracking approaches is that they are not fully automatic.

Rost et al [9] describe their knowledge-based system, called SOLUTION and designed to automatically adopt low-level image processing algorithms to the needs of the application. Smets et al [11] present a knowledge-based system for the delineation of blood vessels on subtracted angiograms. The system encodes general knowledge about appearance of blood vessels in these images in the form of 11 rules (e.g. that vessel have high intensity center lines, comprise high intensity regions bordered by parallel edges etc.).

The system is successful where the image contains high contrast between the vessel and the background and that the system has considerable problems at vessel bifurcations and self-occlusions. Nekovei and Sun [6] describe their back-propagation network for the detection of blood vessels in X-ray angiography. This system does not extract the vascular structure. Its purpose is to label the pixels as vessel or non-vessel.

## 2. BLOOD VESSEL EXTRACTION USING WIENERFILTER

Filters are commonly used to extract a desired signal from a background of random noise or deterministic interference. The most design techniques of filters are based firmly on frequency domain concepts. By contrast, Wiener filters are developed using time-domain concepts. They are designed to minimize the mean-square error between their output and a desired or required output. The performance of the wiener filter may be evaluated by listening to signals and noise.

In this method, the retinal image is taken as the input image. Then the input retinal image is pre-processed. In pre-processing stage, the input image is resized to [576,576] and the green channel image is separated as the blood vessel appears brighter in the green channel image. Then filter is used to remove the noise in the input image. Then histogram equalization is applied to the filtered image. Then bottom hat transform is applied to the equalized image. Figure 2.3 shows the results of blood vessel extraction using wiener filter.

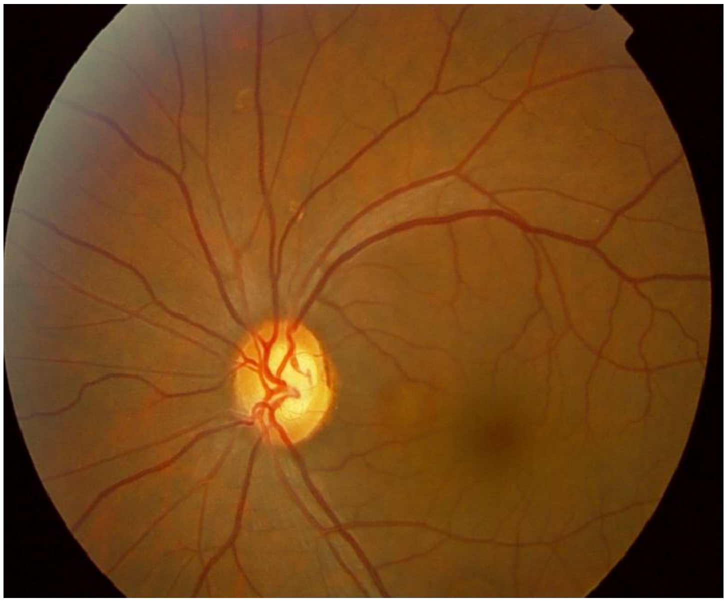


Figure.2.1 Input Retinal Original image

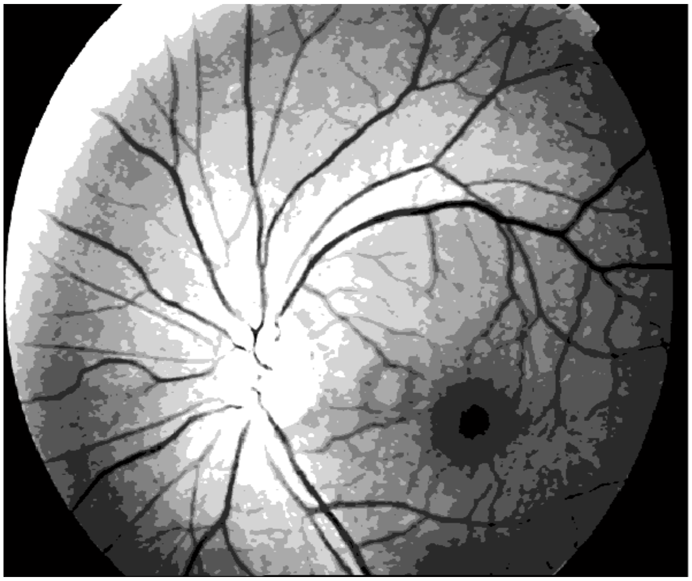


Figure 2.2 Histogram equalized image

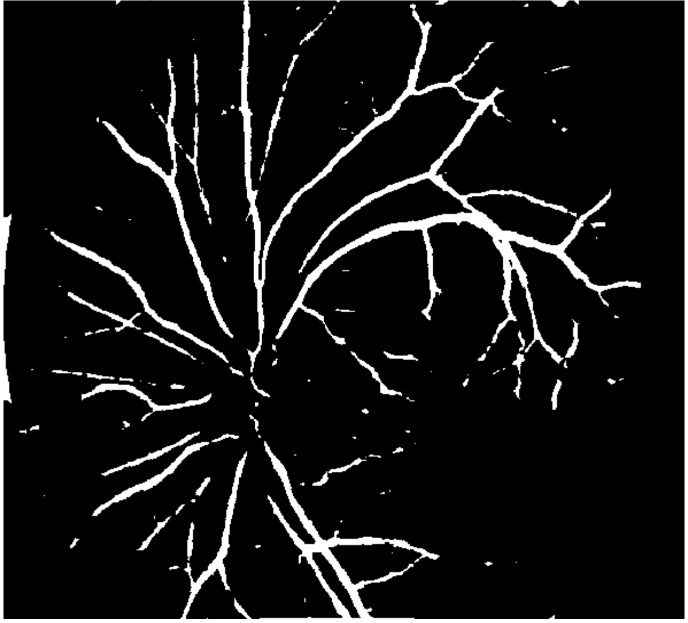


Figure 2.3 out put of Extracted blood vessels

## 3. BLOOD VESSEL EXTRACTION USING MORPHOLOGICAL OPERATION

In this method, the retinal image is taken as the input image. Then the input retinal image is pre-processed. In pre-processing stage, the input image is resized to [576,576] and the green channel image is separated as the blood vessel appears brighter in the green channel image. Then morphological operation is performed on the green channel image.

The primary morphological operations are dilation and erosion. The more complex morphological operations are opening and closing. Dilation is an operation that grows or thickens objects in a binary image. The specific manner and extent of this thickening is controlled by shape referred to a structuring element. Dilation is defined in terms of set operation. Erosion shrinks or thins objects in a binary image. The manner and extent of shrinking is controlled by a structuring element.

Subtractions of closed images across two different scales (S1 and S2 be size of structuring elements) give the blood vessel segments of image. Disk shaped structuring element is used. S2 is set as high value so that the main blood vessels get closed. S1 is chosen as 1 or 2 pixels below S2 to obtain thicker blood vessels or S1 is chosen as at least 4 pixels below S2 to obtain entire blood vessels.

Then edge detection is performed on the morphologically operated image. Laplacian and Gaussian operator detects the blood vessels accurately. Then thresholding is performed on the edge detected image. The blood vessel edges are thinned to a single line width. Then the blood vessels are smoothed. Smoothing function is used to smooth the thinned image for the betterment of blood vessel extraction. The smoothing is performed using box method with window size 5. Figure 3(c) shows the blood vessel extraction using morphological operation.

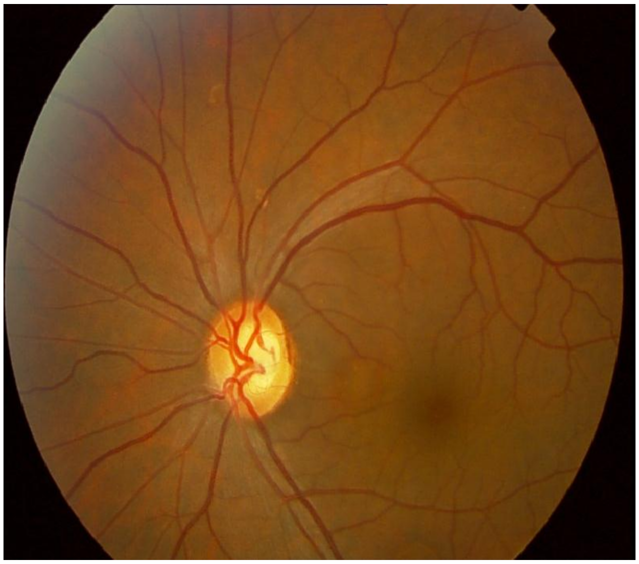


Figure 3.1 Original image

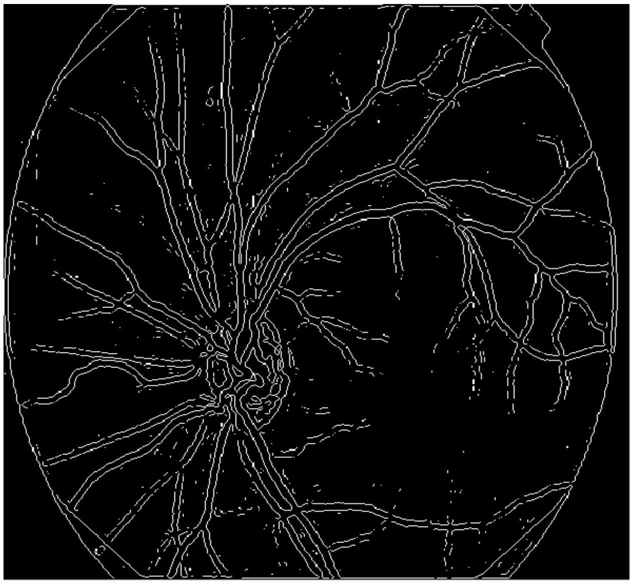


Figure 3.2 morphologically thinned image

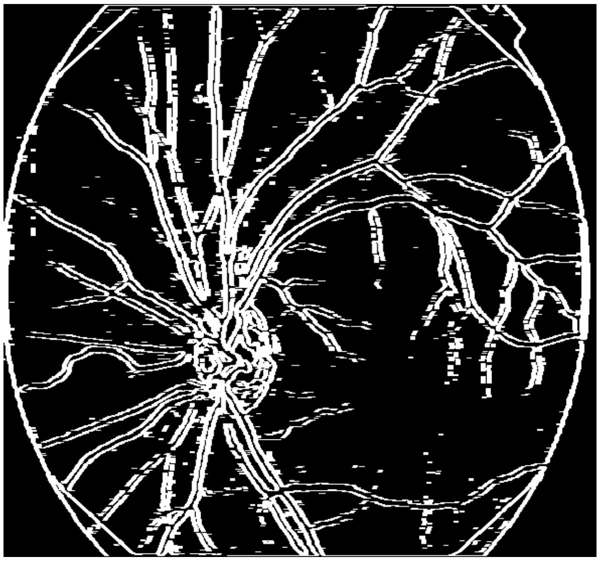


Figure 3.3 Extracted blood vessels

The PSNR values for 50 images are calculated and the average is shown in the table 1.

**Table 1**

| Method | Wiener filter | Morphological method |
|---|---|---|
| PSNR(average) | 5.6861 | 5.8025 |

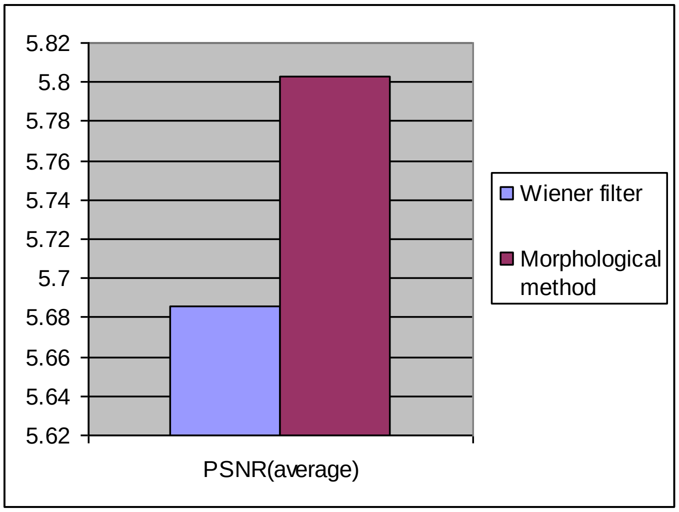


Figure 3.4 Extracted blood vessel and Wiener filter

## 4. CONCLUSION

The green channel image from RGB image is taken and the blood vessel is extracted from the retinal image. The extraction is done using wiener filter and morphological operation. The average PSNR obtained from the 50 retinal images shows that the performance is better in morphological operation. This extraction is important in the diagnosis of diabetic retinopathy. Once the blood vessel is extracted and segmented then the exudates can be easily detected. The optic disk and exudates can be detected in future. Also the thickness of the vessel and other anatomical features can be measured.

## REFERENCES

1. Aylward, S., Pizer, S., Bullitt, E. and Eberl, D. (1996) “Intensity ridge and widths for tabular object segmentation and registration”, in Wksp on Math. Methods in Biomedical. Image Analysis, pp.131–138.
2. Chwialkowski, M.P., Ibrahim, Y.M., Hong, F.L. and Peshock, R.M. (1996) “A method for fully automated quantitative analysis of arterial flow using flow- sensitized images”, Comp. Med. Imaging and Graphics, vol. 20, pp. 365–378.
3. Donizelli, M. (1998) “Region-oriented segmentation of vascular structures from dsa images using mathematical morphology and binary region growing”, vol. 2.
4. Figueiredo, M.T. and Leitao, J.M.N. (1995) “A non smoothing approach to the estimation of vessel contours in angiograms”, IEEE Trans. on Med. Image., vol. 14, pp. 162–172.
5. Krissian, K., Malandain, G. and Ayache, N. (1998) “Model-based multi scale detection and reconstruction of 3d vessels”, Technical Report 3442, INDIA.
6. Nekovei, R. and Sun, Y. (1995) “Back-propagation network and its configuration for blood vessel

detection in angiograms", IEEE Trans. On Neural Nets, vol. 6, pp. 64–72.
7. Niki, N., Kawata, Y., Sato, H. and Kumazaki, T. (1993) "3d imaging of blood vessels using x-ray rotational angiographic system", IEEE Med. Imaging Conf., vol. 3, pp. 1873–1877.
8. Poli, R. and Valli, G. (1997)"An algorithm for real-time vessel enhancement and detection", Comp. Methods and Prog. In Biomed. vol. 52, pp. 1–22.
9. Rost, U., Munkel, H. and Liedtke, C.E. (1998) "A knowledge based system for the configuration of image processing algorithms", Fachtagung Information's and Microsystem Technik.
10. Schmitt, H., Grass, M., Rasche, V., Schramm, O., Haehnel, S. and Sartor, K. (2002) "An x-ray-based method for the determination of the contrast agent propagation in 3-d vessel structures", IEEE Trans. on Med Img., vol. 21, pp. 251–262.
11. Smets, C., Verbeeck, G., Suetens, P. and Oosterlinck, A. (1988) "A knowledge- based system for the delineation of blood vessels on subtraction angiograms", Pattern Rec. Let., vol. 8, pp.113–121.

## Author Biography

**First Author:** I have graduated my B.E degree from the Bharathiar University in the year 1993, I had been working as a Lecturer in Maharaja Engineering College, Avinashi and Dr. Mahalingam College of Engineering and Technology, Pollachi. In the year 2005, I completed my post graduated degree in Applied Electronics under Anna University. Presently I am working as Assistant Professor at VLB Janakiammal College of Engineering and Technology, Coimbatore. My area of interest includes image processing, Biomedical Engineering, Soft Computing and medical Electronics.

# Analysis of Newton's Forward Interpolation Formula

**Nasrin Akter Ripa**
**Daffodil International University,**
**Department of Electronics and Telecommunication Engineering**
**nar@daffodilvarsity.edu.bd**

***Abstract:*** This work presents a theoretical analysis Newton's Forward Interpolation Formula. In order to analyze the method, unit step, unit ramp and sinusoidal signals are chosen. Also to check the performance of the considered method an increasing function and a decreasing function has considered. Some sampled values of a signal are calculated and then the Newton's forward Interpolation formula is used to reconstruct the signal such as image resizing. Errors are analyzed by comparing the actual sampled values with the values obtained by Newton's Interpolation formula.

**Keywords:** Interpolation, forward or interpolation formula, forward difference table, increasing and decreasing function, unit step, unit ramp.

## 1. Introduction

From very ancient time Interpolation is being used for various purposes. Sir Edmund Whittaker, a professor of Numerical Mathematics at the University of Edinburgh from 1913 to 1923, observed "the most common form of interpolation occurs when we seek data from a table which does not have the exact values we want." Liu Zhuo used the equivalent of second order Gregory-Newton interpolation to construct an "Imperial Standard Calendar". In 625 AD, Indian astronomer and mathematician Brahmagupta introduced a method for second order interpolation of the sine function and, later on, a method for interpolation of unequal-interval data. Numerous researchers study the possibility of Interpolation based on the fourier transformer, the Hartley transformer and the discrete cosine transform. In 1983, Parker et al. published a first comparison of Interpolation techniques in medical image processing. They failed, however to implement cubic B-spline Interpolation correctly and arrive at erroneous conclusions concerning this technique [1]. In present days, several algorithms are used for image resizing [2] based on Newton's Interpolation Formula [3].

In this paper, Newton's forward Interpolation formula for equal interval is used for reconstructing a signal defined by a function to analyze its performance. This paper is organized as follows: In section 2, I explain the mathematical principal of Newton's Interpolation method. The implementation is devised in section 3. In section 4 the experimental results are given. The conclusion is summarized in section 5.

## 2. Theory of Interpolation and the considered functions

### 2.1 Forward Interpolation Formula:

Newton's forward difference formula is a finite difference identity giving an interpolated value between tabulated points in terms of the first value and the powers of the forward difference Δ. Let $y = f(x)$ denotes a function which takes the values $y_0, y_1, y_3, \ldots\ldots, y_n$ for the equidistant values $x_0, x_1, x_2, \ldots\ldots, x_n$ respectively of the independent variable $x$. According to Newton's forward interpolation formula the value of $y$ at a particular point can be found as follows:

$$f(x) = g(u) = y_0 + u\Delta y_0 + \frac{u(u-1)}{2!}\Delta^2 y_0 + \frac{u(u-1)(u-2)}{3!}\Delta^3 y_0 + .\frac{u(u-1)(u-2)\ldots\ldots(u-n+1)}{n!}\Delta^n y_0$$

… … … … (1)

where $u = \dfrac{x - x_0}{h}$ and $h = x_1 - x_0$ . Also $\Delta y_0, \Delta^2 y_0, \Delta^3 y_0, \ldots\ldots\ldots$ are the 1st, 2nd, 3rd forward difference respectively and so on.

The forward difference is a finite difference defined by

$$\Delta a_n = a_{n+1} - a_n$$

Higher order differences are obtained by repeated operations of the forward difference operator,

$$\Delta^k a_n = \Delta^{k-1} a_{n+1} - \Delta^{k-1} a_{n,}$$

so

$$\begin{aligned}\Delta^2 a_n &= \Delta_n^2 \\ &= \Delta(\Delta_n)\end{aligned}$$

$$= \Delta(a_{n+1} - a_n)$$
$$= \Delta_{n+1} - \Delta_n$$
$$= a_{n+2} - 2a_{n+1} + a_n$$

In general,

$$\Delta_n^k \equiv \Delta^k a_n \equiv \sum_{i=0}^{k}(-1)^i \binom{k}{i} a_{n+k-i} \ldots \ldots \ldots (2)$$

### 2.2 Unit step function:

The unit step function is defined as:

$$U(t - t_0) = \begin{cases} 1 & t \geq t_0 \\ 0 & t < t_0 \end{cases} \quad \ldots. \ldots. \ldots \ldots \ldots\ldots. (3)$$

### 2.3 Unit ramp function:

The *ramp function*, denoted by *r* (*t*) is a signal whose amplitude increases proportionally as time increases. The mathematical definition of a ramp signal is

$$r(t) = \begin{cases} kt & t \geq 0 \\ 0 & t < 0 \end{cases} \ldots. \ldots. \ldots\ldots\ldots.. \ldots. \ldots (4)$$

## 3. Analysis and implementation

To analyze the effect of the considered method and unit step [4] function is considered first. According to the definition the calculated value of this function using equation (3) is shown in the tabular form:

| Time (T) | Amplitude (x) |
|---|---|
| 0 | 1 |
| 2 | 1 |
| 4 | 1 |
| 6 | 1 |
| 8 | 1 |
| 10 | 1 |

Table 1: Value of an unit step function according to the independent variable T.

The unknown values are also calculated by using equation (1) to reconstruct the signal more accurately and are shown below:

| T | x | $\Delta x$ | $\Delta^2 x$ | $\Delta^3 x$ | $\Delta^4 x$ | $\Delta^5 x$ |
|---|---|---|---|---|---|---|
| 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 2 | 1 | 0 | 0 | 0 | 0 | |
| 4 | 1 | 0 | 0 | 0 | | |
| 6 | 1 | 0 | 0 | | | |
| 8 | 1 | 0 | | | | |
| 10 | 1 | | | | | |

Table 2: Forward Difference table for calculating the unknown values of the function

| Time (T) | Amplitude (x) |
|---|---|
| 1 | 1 |
| 3 | 1 |
| 5 | 1 |
| 7 | 1 |
| 9 | 1 |

Table 3: Calculated values of unit step function using Newton's Forward Interpolation Formula

With these two sets of data a graph is plotted using MATLAB to compare those sets of values:

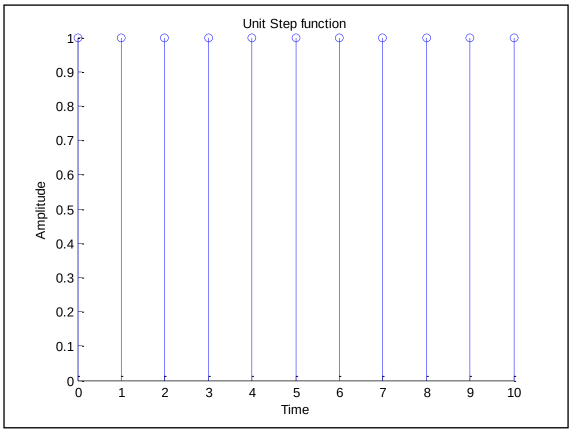


Figure 1: Unit step function

From this graph it is observed that in this case the considered method gives 100% accuracy as expected.
For a unit ramp function the calculated using equation (4) values using equation 4 are tabulated below:

| Time (T) | Amplitude (x) |
|---|---|
| 1 | 1 |
| 3 | 3 |
| 5 | 5 |
| 7 | 7 |
| 9 | 9 |

Table 4: Values for unit ramp function obtained from equation (4)

The values obtained by necessary calculations are given below in tabular form in Table 4 and Table 5 and a graph is also plotted using MATLAB shown in Figure 2:

| T | x | $\Delta x$ | $\Delta^2 x$ | $\Delta^3 x$ | $\Delta^4 x$ | $\Delta^5 x$ |
|---|---|---|---|---|---|---|
| 0 | 0 | 2 | 0 | 0 | 0 | 0 |
| 2 | 2 | 2 | 0 | 0 | 0 | |
| 4 | 4 | 2 | 0 | 0 | | |
| 6 | 6 | 2 | 0 | | | |
| 8 | 8 | 2 | | | | |
| 10 | 10 | | | | | |

Table 4: Forward Difference table for unit ramp function

| Time (T) | Amplitude (x) |
|---|---|
| 1 | 1 |
| 3 | 3 |
| 5 | 5 |
| 7 | 7 |
| 9 | 9 |

Table 5: Calculated values of unit ramp function using Newton's Forward Interpolation Formula

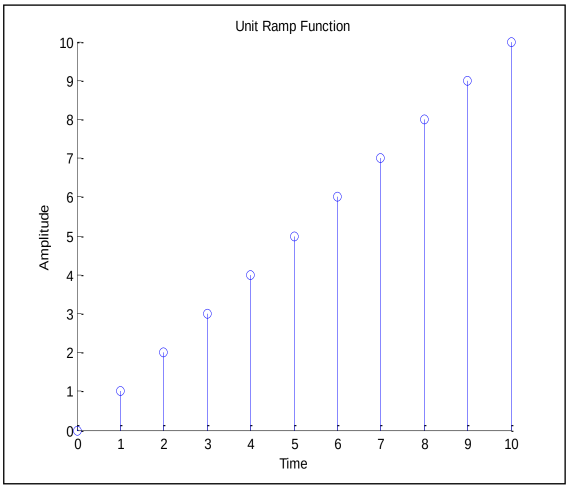


Figure 2: Graph of a unit ramp function

Doing the same procedure in the case of an increasing function (in this case, unit ramp function) it is clear that this method gives 100% accuracy.

But what will happen if we consider a decreasing function? The forward difference table is given below (Table 6):

| T | x | $\Delta x$ | $\Delta^2 x$ | $\Delta^3 x$ | $\Delta^4 x$ | $\Delta^5 x$ |
|---|---|---|---|---|---|---|
| 0 | 10 | -2 | 0 | 0 | 0 | 0 |
| 2 | 8 | -2 | 0 | 0 | 0 | |
| 4 | 6 | -2 | 0 | 0 | | |
| 6 | 4 | -2 | 0 | | | |
| 8 | 2 | -2 | | | | |
| 10 | 0 | | | | | |

Table 6: Forward Difference Table for a linearly decreasing function

Unknown values are obtained from Newton's polynomial. Those are tabulated below in Table 7:

| Time (T) | Amplitude (x) |
|---|---|
| 1 | 19 |
| 3 | 17 |
| 5 | 15 |
| 7 | 13 |
| 9 | 11 |

Table 7: calculated values of the function

To see the effect we can draw a graph with two sets of data.

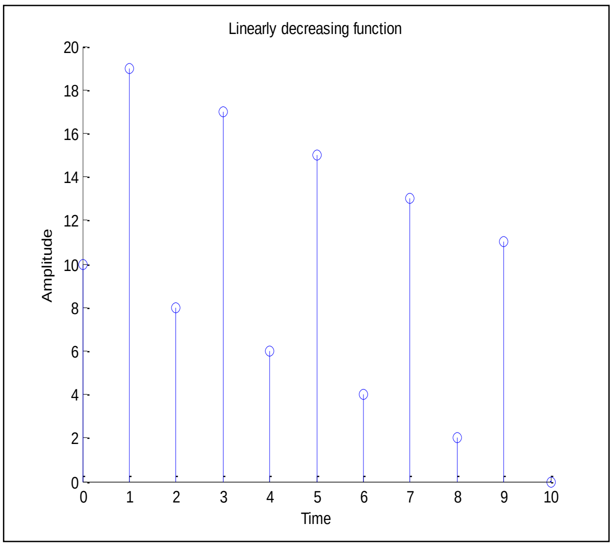


Figure 3: Graph of a linearly decreasing function

In this case we see a large variation between the expected value and the obtained value. But a little bit modification for the decreasing function can make the result perfect. The modified equation is:

$$f(x) = g(u) = u\Delta y_0 + \frac{u(u-1)}{2!}\Delta^2 y_0 + \frac{u(u-1)(u-2)}{3!}\Delta^3 y_0 + \dots\dots\dots + \frac{u(u-1)(u-2)\dots\dots(u-n+1)}{n!}\Delta^n y_0 \dots\dots\dots\dots (5)$$

Using equation (5) the calculated values are shown in Table 8:

| Time (T) | Amplitude(x) |
|---|---|
| 1 | 9 |
| 3 | 7 |
| 5 | 5 |
| 7 | 3 |
| 9 | 1 |

Table 8: Calculated values of a linearly decreasing function

Now, the result can be observed by the following graph:

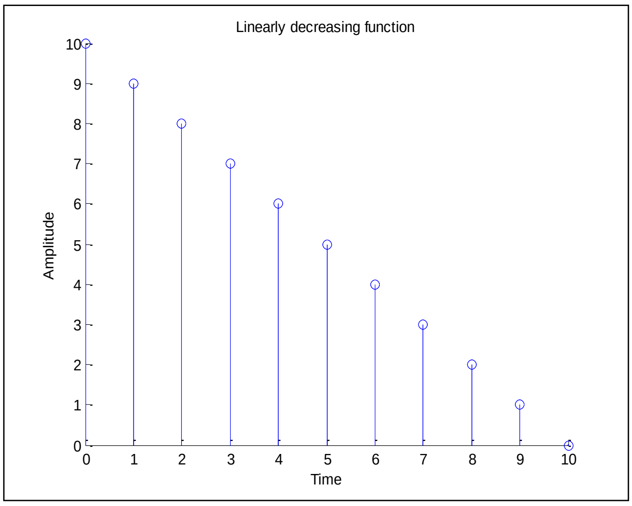


Figure 4: Graph of a linearly decreasing function

Finally I've considered a sinusoid. To reduce the calculation we only calculate the points for one quarter of a cycle. The obtained values are shown below:

| T | x | $\Delta x$ | $\Delta^2 x$ | $\Delta^3 x$ | $\Delta^4 x$ | $\Delta^5 x$ | $\Delta^6 x$ | $\Delta^7 x$ | $\Delta^8 x$ | $\Delta^9 x$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | -0.309 | 0.0302 | 0.0274 | -0.0059 | -0.0015 | -0.0006 | 0.003 | -0.006 | 0.012 |
| 1.1 | -0.309 | -0.2788 | 0.0576 | 0.0215 | -0.0074 | -0.0021 | 0.0024 | -0.003 | 0.006 | |
| 1.2 | -0.5878 | -0.2212 | 0.0791 | 0.0141 | -0.0095 | 0.0003 | -0.0006 | 0.003 | | |
| 1.3 | -0.809 | -0.1421 | 0.0932 | 0.0046 | -0.0092 | -0.0003 | 0.0024 | | | |
| 1.4 | -0.9511 | -0.0489 | 0.0978 | -0.0046 | -0.0095 | 0.0021 | | | | |
| 1.5 | -1 | 0.0489 | 0.0932 | -0.0141 | -0.0074 | | | | | |
| 1.6 | -0.9511 | 0.1421 | 0.0791 | -0.0215 | | | | | | |
| 1.7 | -0.809 | 0.2212 | 0.0576 | | | | | | | |
| 1.8 | -0.5878 | 0.2788 | | | | | | | | |
| 1.9 | -0.309 | | | | | | | | | |

Table 9: Forward Difference Table for a quarter of a cycle of a sinusoid

| Time (T) | Amplitude (x) |
|---|---|
| 1.05 | 1.000204 |
| 1.15 | -0.39804 |
| 1.25 | -0.3999 |
| 1.35 | -0.08203 |
| 1.45 | 0.497075 |
| 1.55 | 1.151958 |
| 1.65 | 1.761094 |
| 1.75 | 3.063 |
| 1.85 | 2.3698 |

Table 10: Calculated values of the sinusoid using equation (1)

Now from the graph it is seen that the variation between actual values and the obtained values are large.

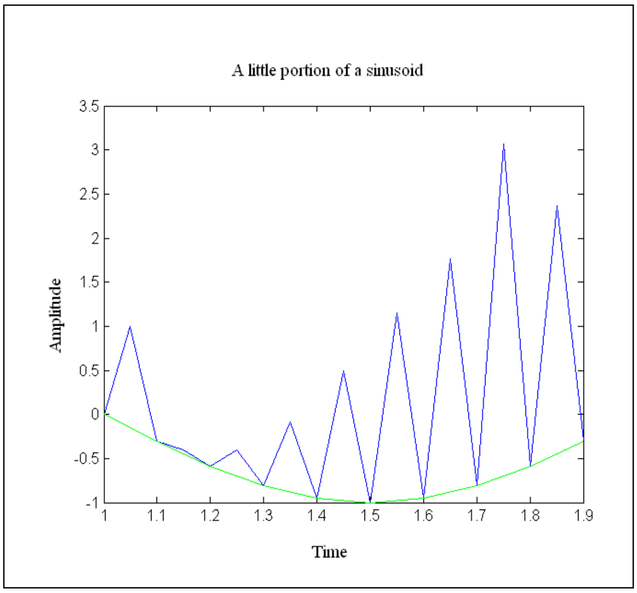


Figure 4: Comparison between the actual values and the calculated values of a sinusoid

From here we observe that to construct the signal in this case this method shows the worst performance.

## 4. Experimental results

From the analysis the limitation we've founded can expressed as follows:

i) For a step function it gives 100% accuracy

ii) For a linearly increased function like ramp function it gives 100% accuracy but if the function is decreasing (linearly) it can't give the right answers. A little bit modification is needed to get 100% accuracy.

iii) The worst case happens when the function can't be defined as a increasing or decreasing function but a combination.

Finally we can say that It gives the best result when the function is constant or increased linearly.

## 5. Conclusion

In this paper, according to the analysis the performance of Newton interpolation formula on different types of functions is presented. Experimental results show that for reconstructing a signal (e.g. image resizing) it works better for the area where signal values are relatively constant or increasing. In conclusion, it can be said that this formula is designed for a function whose value will increase or remain constant with the independent variable.

## 6. References

[1]. Meijering, Erik. "A Chronology of Interpolation: From Ancient Astronomy to Modern Signal and Image Processing." Proceedings of the IEEE. vol. 90, no. 3, pp. 319-42. March 2002

[2]. Jianping Xiao, Xuecheng Zou, Zhenglin Liu, Xu Guo, "Adaptive Interpolation Algorithm for Real-time Image Resizing," icicic, vol. 2, pp.221-224, First International Conference on Innovative Computing, Information and Control - Volume II (ICICIC'06), 2006

[3]. Numerical Methods for Engineers, by Steven C. Chapra and Raymond P. Canale, Forth Edition, Chapter 18, Page no: 474.

[4]. Digital Signal Processing, by John G. Proakis and Dimitris G. Manolakis, Forth Edition, Chapter 2, Page no: 43.

## Author Biography

**Name: Nasrin Akter Ripa**
Date of Birth: Date of birth: December 3, 1985
Place of Birth: Barisal, Bangladesh

**Educational Background:**

1. B. Sc. In Electronics and Telecommunication Engineering, Institute: Daffodil International University, City: Dhaka, Country: Bangladesh, Year of degree earned: 2008.
2. M. SC. In Electronics and Telecommunication Engineering (running), Institute: North South University, City: Dhaka, Country: Bangladesh.

**Working Experience:**

1. Designation: Lecturer, Dept. of Electronics and Telecommunication Engineering, Daffodil International University. Address: 102 Shukrabad, Mirpur Road, Dhanmondi, Dhaka-1207. Web address: www.daffodilvarsity.edu.bd. Working from 14th September, 2008 to now.

# Creating an Efficient Prefetching Mechanism by Leveraging Rule Based Agents

Jyoti[1], A. K. Sharma[2], Amit Goel[3]

[1]Sr. Lecturer, Dept of CE, YMCA Univ. of Science and Tech., Haryana, INDIA
[2]Prof. & Head, Dept of CE, YMCA Univ. of Science and Tech., Haryana, INDIA
[3]Manager, Evalueserve, Gurgaon, Haryana, INDIA
***justjyoti.verma@gmail.com, ashokkale2@rediffmail.com, goelamit1@yahoo.com***

***Abstract:*** Prefetching and caching are two well-known approaches for improving the performance of the Web and have become essential components of the Web infrastructure. But without their careful usage, both can result in the depletion of the performance which they could render complementing each other's drawbacks. In recent years, agents have become a very popular paradigm in computing because of their flexibility, modularity and general applicability to a wide range of problems. This paper provides a novel approach wherein agents have been introduced between client machines and proxy server to help the clients in getting the prefetched documents of their interests thereby balancing both the caching and prefetching.

***Keywords:*** Prefetching, agents, web mining, proxy server

## 1. Introduction

The Web has evolved rapidly from a simple information-sharing mechanism offering only static text and images to a rich assortment of dynamic and interactive services, such as video/audio conferencing, e-commerce, and distance learning. The explosive growth of the Web has imposed a heavy demand on networking resources and Web servers. Users often experience long and unpredictable delays when retrieving Web pages from remote sites [1]. Hence, an obvious solution in order to improve the quality of Web services would be the increase of bandwidth, but such a choice involves increasing economic cost. However, the higher bandwidth would solve temporarily the problems since it would ease the users to create more and more resource-hungry applications, bunching again the network. Therefore, the network limitations will remain or worsen unless effective software solutions are also provided. The authors of [2] have proposed a methodology to incorporate a Predictive Prefetching Engine (PPE) at the proxy level that helps in creating a database of rules that are extracted by applying the various data mining techniques at diverse levels on the proxy log. The current paper extends the work of [2] by introducing agents between clients and the proxy servers that will help in the triggering of rules for prefetching the web documents according to the client's requirements. The organization of the paper is as follows. The next three subsections give a brief outline of the web caching, web prefetching and the agents in general. Section II outlays the various factors that acted as motivation for this work. Section III discusses the proposed work followed by the conclusion and references.

### 1.1. The Web Caching Approach

Caching proved itself as an important technique to optimize the way the Web is used [3]. In particular, Web caching is implemented by proxy server applications developed to support many users. Proxy applications act as an intermediate between Web users and servers. Users make their connection to proxy applications running on their hosts. The proxy connects the server and relays data between the user and the server. At each request, the proxy server is contacted first to find whether it has a valid copy of the requested object. If the proxy has the requested object this is considered as a cache hit or otherwise a cache miss occurs and the proxy must forward the request on behalf of the user. Upon receiving a new object, the proxy services a copy to the end-user and keeps another copy to its local storage. From the above discussion follows that Web caching reduces bandwidth consumption, network congestion, and network traffic because it stores the frequently requested content closer to users. Also, because it delivers cached objects from proxy servers, it reduces external latency (the time it takes to transfer objects from the origin server to proxy servers). Finally, caching improves reliability because users can obtain a cached copy even if the remote server is unavailable. As far as concerned the performance of a Web proxy caching scheme, it is mainly dependent on the cache replacement algorithm [4] (identify the objects to be replaced in a cache upon a request arrival) which has been enhanced by the underlying proxy server. However, cache hit rates have not improved much with these schemes. Particularly, a Web caching scheme has three significant drawbacks:

- If the proxy is not properly updated, a user might receive stale data,
- as the number of users grows, origin servers typically become bottlenecks.

- Finally, several factors diminish the ideal effectiveness of Web caching. The obvious factors are the limited system resources of cache servers (i.e., memory space, disk storage, I/O bandwidth, processing power, and networking resources).

However, even if the cache space is unlimited, there are significant problems that cannot be avoided by such an approach. Specifically, large caches are not a solution because, the problem of updating such a huge collection of Web objects is unmanageable.

Therefore, we must resort to an approach, which will predict the future users' requests and retain in cache the most valuable objects.

### 1.2. The Web Prefetching Approach

Prefetching attempts to overcome these limitations by pro-actively fetching content before users actually request it [5]. Web prefetching is the process of deducing user's future requests for Web objects by locating popular requested objects into the cache prior to an explicit request for them. Unlike Web data caching, which exploits the temporal locality, the Web prefetching schemes are based on the spatial locality of Web objects. In particular, the temporal locality refers to repeated users' accesses to the same object within short time periods, whereas, the spatial one refers to users' requests where accesses to some objects frequently entail accesses to certain other objects. Typically, the main benefits of employing prefetching are that it prevents bandwidth underutilization and reduces the latency. Therefore, bottlenecks and traffic jams on the Web are bypassed and objects are transferred faster. Thus, the proxies may effectively serve more users' requests, reducing the workload from the origin servers. Consequently, the origin servers are protected from the flash crowd events as a significant part of the Web traffic is dispersed over the proxy servers. On the other hand, the main drawback of systems which have enhanced prefetching policies is that some prefetched objects may not be eventually requested by the users. In such a case, the prefetching scheme increases the network traffic as well as the Web servers' load. In order to overcome this limitation, high accuracy prediction models have been used [6]. From the above it occurs that caching and prefetching complement each other in order to reduce the noticeable response time perceived by users [7].

### 1.3. The Agents

The details of mobile agents and their employment in a distributed environment can be found elsewhere [8, 9]. Here, a brief introduction of mobile agents and their role in a network-based information system has been discussed. The term mobile agent is often context-dependent and has two separate and distinct concepts: *mobility* and *agency*. The term *agency* implies having the same characteristics as that of an agent. These are self contained and identifiable computer programs that can move within the network from node to node and act on behalf of a user or other entity. These can halt execution from a host without human interruption ([10]). The current network environment is based on the traditional client/server paradigm as shown in Fig.1.

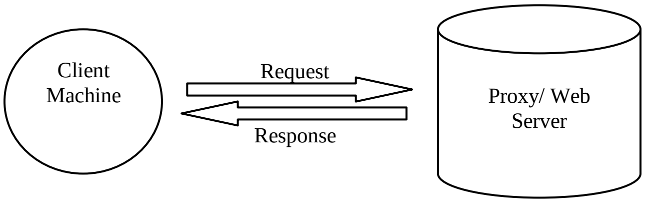


Fig. 1 Client/ Server Communication

However, in the case of mobile agents employed in a network, the service provision/utilization can be distributed in nature and is dynamically configured according to changing network performance metrics like congestion and user demand for service provision ([11]).

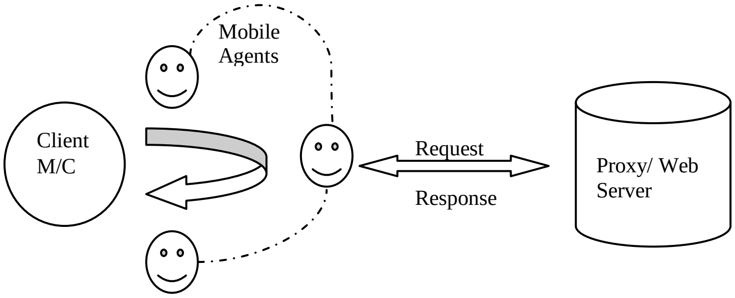


Fig. 2 Mobile Agent Communication

Mobile agents are typically suited to applications requiring structuring and coordinating wide area network and distributed services involving a large number of remote real time interactions. They can decide how best to handle a user's request based on past data or history of a similar request. These programs are therefore capable of learning from user behavior to some extent. Fig. 2 shows how agents can act as the intermediary between the clients and the proxy to meet the client's needs.

## 2. Motivation

Despite of the many efforts done by the industry and the research community to improve the World Wide Web performance, the web latency perceived by the users is still a perennial issue to reduce. The rising up of the Web architecture techniques such as web caching, prefetching and replication have became an important solution to reduce the user perceived latency. These techniques make use of the temporal, spatial and geographical locality properties of web objects to improve the web performance [12]. In the open literature, there are several proposals about the benefits of the above techniques applied at different elements of the generic Web architecture (i.e. clients, proxies, servers). In [13] authors suggests that the use of caching can reduce up to 26% the latency; also the use of prefetching can improve the web performance up to 57%. Furthermore, the combined use of caching and prefetching can reduce the latency perceived up to 60%. Nevertheless, authors in [14] take into account the current Web generation; point out a theoretical upper bound of 97% of latency reduction when prediction is done in a collaborative manner between proxies and servers. In [15], the authors present a non-interfering web prefetch system between clients and servers, unfortunately there is not a caching module in its architecture. An attempt to integrate web caching and prefetching was done by [16] in an interesting proposal where both techniques were applied only at the proxy server side and the used workload was generated by a synthetic workload generator. The authors in [17] present an extended study about a prefetching technique and its impact on the Proxy Cache Server in a real WAN environment (i.e. university campus). The later proposal contributes with many useful considerations (e.g. log analysis, session estimation, web object types) to take into account when prefetching is applied. In [2], authors have proposed a framework for extracting relevant web pages from WWW using data mining. They proposed a Predictive Prefetching Engine (PPE) that sits on the proxy server. Since a proxy server lies between a web browser and a web server, it is a potential tool that can be suitably employed to reduce the www latency i.e. it can intercepts all requests to the web server to see if it can fulfill the requests by itself. If not, then only it may forward the request to the web server. The job of PPE is to preprocess the proxy web log to perform the preprocessing tasks like reduction of search space, user and session identification and path completion. After preprocessing a cleaner version of log is formed called data mart. The next step applies data mining operations like rough set clustering so as to narrow down the look up into the log and thus reducing the complexity of the overall process. Following this is the rule generation phase which extracts the rules for prediction. The data mining operations like rough set clustering, markov and association rules if applied alone do not provide accuracy. Therefore, authors have carefully integrated the three operations to improve the prediction accuracy thereby making the repository of the rules that will help in prefetching of the desired documents [18]. The proposed work extends the effort of [2, 18] by introducing agents between clients and the proxy servers that will help in the triggering of rules for prefetching the web documents according to the client's requirements.

## 3. Proposed architecture

The proxy is chosen as the deployment location for PPE since the active involvement of the clients is not desirable as that would require the clients to involve actively which clients tend to avoid and the focus is to make the structure as transparent as possible. Our approach requires input from several clients and thus choosing a single client as a point of deployment would not have been beneficial. The web servers themselves are involved in several other tasks that require a lot of memory and processing time and including one more process would have affected the throughput of the web servers. The outcome of the PPE is the repository of rules of the form $D_{i=>} D_j$ or $D_{i=>} D_{j=>} D_k$

These rules will be of no use if they are not fired according to the client's demands. To effectively fire the desired rules, a layer of agents has been introduced between the client machines and the Proxy's PPE. The proposed framework is shown in fig.3. For every client machine, there will be a dedicated intelligent agent providing its services. The agents will work in the following manner:

### 3.1. Prefetching Scheme

Given that we have a set of rules in the repository created and maintained by the PPE, the prefetching scheme works as follows:

1. Let the request be for document A.
2. The agent scans the rule database[1] for the rules of the form A→X for some document X.
3. The agent then scans the database for every rule or part of the rule which has X in its sequence (e.g. A→ Y→X →Z). The only exception to this scan would be in the case of X being the last document in the sequence.

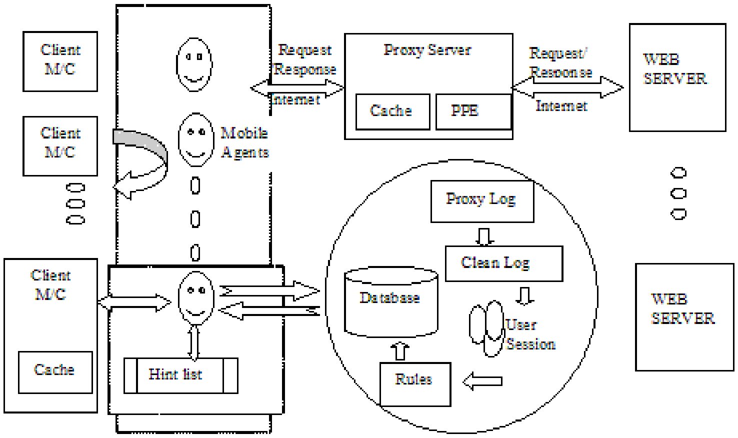


Fig.3 Proposed Architecture for prefetching the documents from the rule database of Proxy's PPE to the client's cache

4. As it scans, the agent brings all the documents that succeed X to the hint list maintained by the agent itself

[1] The rule database can be organized using some indexing scheme

and accordingly prefetch the corresponding web page from the proxy server to the client's cache.

5. The agent continues the scan and populates the hint list till such time the user requests for a web page which doesn't appear in the sequence. In such case, the agent cleans up the hint list and starts afresh (step 2).

## Conclusion

The paper discusses the new approach to prefetch the documents from the proxy server to the client's cache. For performing the above said task, agents have been introduced between the client machines and the proxy server. The Proxy server holds the PPE whose task is to make the rule database by applying the various data mining techniques on the proxy's web log that marks every entry which exists between the client and the web servers. The agents thus effectively trigger the desired rules according to the interests of the users. There is a dedicated agent rendering its services for each user so that at any instant when the user's line of interest changes, the agents change there line of action.

## References

[1] Pallis G, Vakali A., "Insight and perspectives for content delivery networks". Commun ACM (CACM) 2006; 49(1):101–6.

[2] Jyoti, A.K. Sharma, Amit Goel, "A Framework for Extracting Relevant Web Pages from WWW using web mining", In Proc of International journal of Computer Society and Network security, Seoul, Korea

[3] Rabinovich M, Spatsheck O., "Web caching and replication." Addison Wesley; 2002

[4] Podlipnig S, Boszormenyi L., "A survey of Web cache replacement strategies". ACM Comput Surveys 2003;35(4):374–98.

[5] Jiang Y, Wu M, Shu W., "Web prefetching: costs, benefits and performance". In: Proceedings of the 7th international workshop on web content caching and distribution (WCW2002). Boulder, Colorado; 2002.

[6] Yang Q, Zhang H., "Integrating Web prefetching and caching using prediction models." World Wide Web 2001;4(4):299–321.

[7] Kroeger TM, Long DDE, Mogul JM, "Exploring the bounds of web latency reduction from caching and prefetching." In Proceedings of the USENIX symposium on Internet technologies and systems. Monterey, California, USA; 1997

[8] Karmouch A. and V. A. Pham (1998), "Mobile Software agents: An Overview", IEEE Communications Magazine, 36(7), 26-37.

[9] M. K. Perdikeas et al (1999), "Mobile Agents Standards and Available Platforms", Comp.Net, 31(19), 1999-2016.

[10] J. Cao, G. H. Chan, W. Jia, and T. Dillon (2001), "Check-pointing and Rollback of Wide-Area Distributed Applications using Mobile Agents", Proceedings, International Parallel and Distributed Processing Symposium, San Francisco: IEEE Computer Society Press.

[11] Ahmad, H. F. and Helene Arfaoui, Mori, K (2001), "Autonomous Information Fading by Mobile Agents for Improving User's Access Time and Fault Tolerance", Proceedings 5th International Symposium on Autonomous Decentralized Systems, 279-283.

[12] M. Rabinovich and O. Spatscheck, "Web Caching and Replication". Addison Wesley, 2002.

[13] T. M. Kroeger, D. D. Long, and J. C. Mogul, "Exploring the bounds of web latency reduction from caching and prefetching," in Procc. of the 1st USENIX Symp. on Internet Technologies and Systems, Monterey, USA, 1997.

[14] J. Domenech, J. Sahuquillo, J. A. Gil, and A. Pont, "The impact of the web prefetching architecture on the limits of reducing user's perceived latency," in Procc. of the 2006 IEEE/WIC/ACM Inter. Conf. on Web Intelligence. IEEE, 2006.

[15] R. Kokku, P. Yalagandula, A. Venkataramani, and M. Dahlin, "NPS: A non-interfering deployable web prefetching system," in Procc. of the USENIX Symp. on Internet Technologies and Systems, Palo Alto, USA, 2003.

[16] W.-G. Teng, C.-Y. Chang, and M.-S. Chen, "Integrating web caching and web prefetching in client-side proxies," IEEE Transactions on Parallel and Distributed Systems, vol. 16, no. 5, pp. 444–455, 2005.

[17] C. Bouras, A. Konidaris, and D. Kostoulas, "Predictive prefetching on the web and its potential impact in the wide area." World Wide Web, vol. 7, no. 2, pp. 143–179, 2004.

[18] Jyoti, A.K. Sharma, Amit Goel, "A Novel Approach to Determine the Rules for Web Page Prediction using Dynamically Chosen K-Order Markov Models", In Proc of IEEE sponsored International Conference on Advances and Emerging trends in Computing Technologies, 2010

# The Dynamic Analysis of Investment System with Two Time Delays

H.X.Yao[1], K.Liu[2], R.Liu[3]

[1]Nonlinear Scientific Research Center, Xuefu Road 301
[2,3]Jiangsu University, Zhenjiang 212013, China
hxyao@ujs.edu.cn, l22k77@163.com,lr_smile@163.com

***Abstract***: In order to explore the intrinsic relationship between investment corporations and investment projects, according to enterprise ecological theory, the investment corporations in accordance with the scope of investment projects is divided into small and medium-sized investment corporation (SMIC) and large investment corporations. According to enterprise ecosystem, the article builds a dynamic system model to explain the complex relationship between investment business and investment projects and we verify the results of the analysis by numerical simulation.

***Keyword***: dynamic model; investment projects; investment corporations

## 1. Introduction

Recently, an increasing number of scholars applied ecological competition model theory to business and have made great achievements. On the basis of organizational ecology, Hannan[1] made a complete concept of organizational ecology and research framework and built a mathematical model which can weight the individual enterprise’s development, change and succession. As the increasing fusion of the world economy and the economic environment increasing deterioration, Moore[2] proposes enterprise ecosystem evolution theory and believes corporations should beyond the perspective of the enterprise and develop its own evolution strategy in a view of business ecosystem point. Analysis of economic phenomena by differential equations has been widely used by some scholars, and recognized by an increasing number of economists. H.X.Yao[3] described an economic system by discrete-time equations, while X.L.Ke[4] and X.Z.He[5] can describe an asset pricing model truly by continuous-time equations.

The article is organized as follows. In section 2, we build an ecological model of investment corporations. We discuss the stability of the model in section 3, and some simulation examples are given to illustrate the obtained results. The last part, we state our results.

## 2. The model

Gaining investment projects is the mainly aim for an investment company. The relationship between investment projects and investment corporations is similar to the relationship between predator and prey in predator-prey model. Meanwhile, there are competitions among the investment corporations, which are similar to internal competitions among predators in the predator-prey model. So we can change the biological model, and use it to analysis investment activities. In order to study the relationship between of investment corporations and investment projects, this is a new idea, and a new method.

### 2.1 The basic assumption of the model

(1) A regional business groups (Investment Corporations or other corporations) can be seen as a biosphere in a population, assuming that the number of investment corporations per unit area can be accurately represent by a variable.

(2) In the investment market, investment corporations’ scramble for resources is mainly reflected in the acquisition of the investment projects.

(3) We divide the investment corporations into small-medium investment corporations and large investment corporations according to the geographical scope of the projects which the investment corporations to bid for.

(4) Entrepreneurs are rational, before the investment company was established, and there is no competition in the same industry within a certain range.

(5) The entrepreneurs would not build investment corporations in a place until investment projects of the area have increased to a certain number.

### 2.2 Model and solution

We can get the investment corporation's ecological model by the stage-structure of predator-prey model proposed by Hasting[6]-[8].

1) investment projects

$$\frac{dx}{dt}=(a-\mu x)x-dx(z+y) \qquad (1)$$

$x(t)$, $y(t)$, $z(t)$ represent the density of investment projects, small-medium investment corporations, large investment corporations at time $t$ at the unit area respectively. We use the Logistic equation proposed by Verhulst to explain the change of investment projects according to time $t$ . The number of investment projects in a region not only has a natural expansion at the speed of $ax$, but also impacts by its own feedback at the rate of $-bx^2$. From the long term, it should meet the S-curve. $a$ is the natural growth rate of investment projects. $\mu$ is the inhibition coefficient of investment projects. $d$ is the effect coefficient between investment projects and investment corporations.

2) small and medium-sized investment corporation (SMIC)

$$\frac{dy}{dt} = cx\left(t-\tau_1\right)\left(z+y\right) - my^2 - eyz\left(t-\tau_2\right) \qquad (2)$$

$c$ is he growth rate of SMIC with the investment projects' increase. $m$ means the competition among SMIC. $e$ means the competition between SMIC and large investment corporations. $x\left(t-\tau_1\right)$ is the number of investment projects accumulated through time $\tau_1$, $z\left(t-\tau_2\right)$ is the number of SMIC exist at $t-\tau_2$,which become large investment corporation through time $\tau_2$.

3) large investment corporation

$$\frac{dz}{dt} = nyz\left(t-\tau_2\right) - bz^2 \qquad (3)$$

$n$ means the coefficient between SMIC and large investment corporations. $b$ is the coefficient among the large investment corporations .(All coefficients are positive.)

We can get the dynamic system of investment from (1)(2)(3),as followed:

$$\begin{cases} \dfrac{dx}{dt} = (a-\mu x)x - dx(z+y) \\ \dfrac{dy}{dt} = cx\left(t-\tau_1\right)\left(z+y\right) - my^2 - eyz\left(t-\tau_2\right) \\ \dfrac{dz}{dt} = nyz\left(t-\tau_2\right) - bz^2 \end{cases} \qquad (4)$$

System $(4)$ has four equilibriums：

$E_0 = \left(0,0,0\right)$,

$E_1 = \left(\dfrac{a}{\mu},0,0\right)$,

$E_2 = \left(\dfrac{ma}{cd+\mu},\dfrac{ca}{cd+\mu},0\right)$

$$E_3 = \left(\frac{ab\left(ne+bm\right)}{cd\left(n+b\right)^2 + b\mu(ne+bm)},\frac{abc\left(n+b\right)}{cd\left(n+b\right)^2 + b\mu(ne+bm)},\frac{acn(n+b)}{cd\left(n+b\right)^2 + b\mu(ne+bm)}\right)$$

## 3. Analysis the properties of the system's solutions

From $(4)$ we can get the jacobian determinant as followed

$$J = \begin{vmatrix} \lambda - a + 2\mu x + dy + dz & dx & dx \\ -c\left(y+z\right)e^{-\lambda\tau_1} & \lambda - cx + 2my + ez & -cx + eye^{-\lambda\tau_2} \\ 0 & -nz & \lambda + 2bz - nye^{-\lambda\tau_2} \end{vmatrix} \qquad (5)$$

The characteristic equation is as followed for the equilibrium $E_0$

$$\lambda^2\left(\lambda - a\right) = 0$$

We can get the root of it

$\lambda_1 = a$, $\lambda_2 = \lambda_3 = 0$,so $E_0$ is unstable point. In an economic view, the number of investment projects in one region is zero, the numbers of the two types of investment corporations are zero, too. This phenomenon could not exist in our real life, so we do not study.

The characteristic equation is as followed for the equilibrium $E_1$

$$\lambda\left(\lambda + a\right)\left(\lambda - \frac{ac}{\mu}\right) = 0$$

We can get the root of it

$\lambda_1 = 0$, $\lambda_2 = -a$, $\lambda_3 = \dfrac{ac}{\mu}$.So $E_1$ is unstable point, too. Namely there are investment projects, but the numbers of small-medium investment corporations and large-invested enterprises are zero. Because the number of investment projects is increasing with time $t$, so the numbers of investment corporations would increase. Namely $E_1$ is unstable point.

The same to $E_2$,we can get the system exist eigenvalue $\lambda > 0$,so $E_2$ is unstable point. Namely, there are investment projects and SMIC, but the number of large investment corporation is zero. As the development of the social, a certain number of SMIC can meet market demand, when there are few investment projects. The SMIC will eventually develop into large investment corporations. So $E_2$ is unstable point.

For the positive equilibrium $E_3$, we can get the characteristic equation of $E_3$

$$P(\lambda) + Q(\lambda)e^{-\lambda\tau_1} + R(\lambda)e^{-\lambda\tau_2} + S(\lambda)e^{-\lambda(\tau_1+\tau_2)} = 0 \qquad (6)$$

Let

$$\begin{aligned} P(\lambda) &= \lambda^3 + (2bz - cx + 2my + ez - a + 2\mu x + dy + dz)\lambda^2 \\ &\quad + \left[2bz(-cx + 2my + ez) + (2bz - cx + 2my + ez)(-a + 2\mu x + dy + dz) - cxnz\right]\lambda \\ &\quad + (-a + 2\mu x + dy + dz)\left[2bz(-cx + 2my + ez) - ncxz\right] \\ &= \lambda^3 + p_2\lambda^2 + p_1\lambda + p_0 \end{aligned}$$

$Q(\lambda) = \left(\lambda + 2bz + nz\right)cdx(y+z) = q_1\lambda + q_0$,

$q_1 > 0$, $q_0 > 0$

$$\begin{aligned} R(\lambda) &= -ny\lambda^2 - ny(-cx + 2my + ez - a + 2\mu x + dy + dz)\lambda \\ &\quad - ny(-a + 2\mu x + dy + dz)(-cx + 2my) \\ &= r_2\lambda^2 + r_1\lambda + r_0 \end{aligned}$$

$S(\lambda) = -cdnxy(y+z) = s_0$, $s_0 < 0$

### 3.1. $\tau_1 = \tau_2 = 0$

We can get for Eq.(6)

$$P(\lambda) + Q(\lambda) + R(\lambda) + S(\lambda) = 0 \quad (7)$$

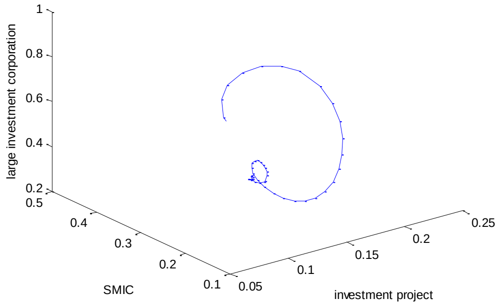
$$\lambda^3 + (p_2 + r_2)\lambda^2 + (p_1 + q_1 + r_1)\lambda + (p_0 + q_0 + r_0 + s_0) = 0$$

According to the Routh-Hurwitz criterion, the equilibrium point $E_3$ is stable if and only if

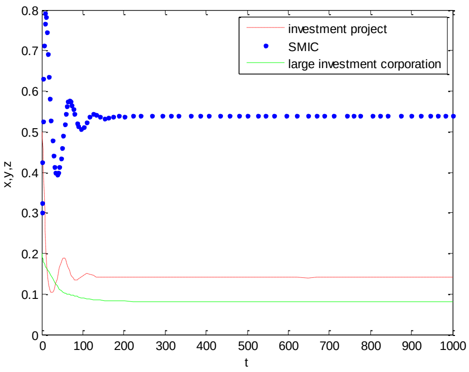
$(a)$ $\left(p_2+r_2\right)>0$

$(b)$ $\left(p_2+r_2\right)\left(p_1+q_1+r_1\right)>\left(p_0+q_0+r_0+s_0\right)$

We set a=0.2 , $\mu$=0.1 , d=0.3 , c=0.4 , m=0.06 , e=0.4 , $n=0.03$, b=0.2 ,and $x=0.5$ , $y=0.3$ , $z=0.2$ .the number of investment projects and investment corporations reach a equilibrium point $E_3\left(0.1406,0.5390,0.0808\right)$ ,and simulation diagram shown as fig1(a-b)



(a)



(b)

Fig1(a-b) $\tau_1=\tau_2=0$ , $E_3$ is asymptotically stable.

### 3.2. $\tau_1\neq0,\tau_2=0$

From Eq.(6), we can get

$$P(\lambda)+R(\lambda)+\left(S(\lambda)+Q(\lambda)\right)e^{-\lambda\tau_1}=0 \quad (8)$$

Namely

$$\lambda^3+(p_2+r_2)\lambda^2+(p_1+r_1)\lambda+(p_0+r_0)+(q_1\lambda+q_0+s_0)e^{-\lambda\tau_1}=0 \qquad (9)$$

Let $A_1=p_2+r_2$ , $B_1=p_1+r_1$ , $C_1=p_0+r_0$ , $D_1=q_1$ , $E_1=q_0+s_0$

We can get from (9)

$$\lambda^3+A_1\lambda^2+B_1\lambda+C_1+\left(D_1\lambda+E_1\right)e^{-\lambda\tau_1}=0 \qquad (10)$$

Lemma1.Eq. (8) has a unique pair of purely imaginary roots if $C_1^2<E_1^2$

**Proof.** If $\lambda=i\omega$ , $\omega>0$ is a root of (10), separating real and imaginary parts, we have the following:

$$\begin{cases}\omega^3-B_1\omega=D_1\omega\cos\omega\tau_1-E_1\sin\omega\tau_1\\ A_1\omega^2-C_1=D_1\omega\sin\omega\tau_1+E_1\cos\omega\tau_1\end{cases} \qquad (11)$$

Squaring and adding both equations we have

$$\omega^6+Q_1\omega^4+Q_2\omega^2+Q_3=0 \qquad (12)$$

Where

$Q_1=A_1^2-2B_1$ ,

$Q_2=B_1^2-2A_1C_1-D_1^2$ ,

$Q_3=C_1^2-E_1^2$

We know that

$$Q_1=\left(2bz-ny\right)^2+\left(-cx+2my+ez\right)^2+\left(-a+2\mu x+dy+dz\right)^2+2cnxz>0$$

$$Q_3=C_1^2-E_1^2<0$$

Then the condition of this lemma implies that there is a unique positive root $\omega_0$ satisfying (8). That is, (8) has a unique pair of purely imaginary roots $\pm i\omega_0$ .

From (11) $\tau_{1,n}$ can be obtained

$$\tau_{1,n}=\frac{1}{\omega_0}\cos^{-1}\frac{D_1\omega_0^4+\left(A_1E_1-B_1D_1\right)\omega_0^2-E_1C_1}{D_1^2\omega_0^2+E_1^2}+\frac{2n\pi}{\omega_0}$$

And , $n=0,1,2,\cdots$

Lamma2. If the following conditions

$C_1^2<E_1^2$, $A_1\left(B_1+D_1\right)>C_1+E_1$ ,

$\left(B_1^2-2A_1C_1\right)E_1^2>C_1^2D_1^2$ ,

hold, system (4) undergoes Hopf bifurcation at $E^*$ when $\tau_1=\tau_{1,\ n}$ , $n=0,1,2,\cdots$ ; furthermore $E^*$ is locally asymptotically stable if $\tau_1\in\left[0,\ \tau_{1,0}\right)$ ; and unstable if $\tau_1>\tau_{1,0}$ .

**Proof**. Differentiating (10) with respect $\tau_1$ , we get

$$\left[3\lambda^2+2A_1\lambda+B_1+D_1e^{-\lambda\tau_1}-\tau_1\left(D_1\lambda+E_1\right)e^{-\lambda\tau_1}\right]\frac{d\lambda}{d\tau_1}=\lambda\left(D_1\lambda+E_1\right)e^{-\lambda\tau_1}$$

That is

$$\left(\frac{d\lambda}{d\tau}\right)^{-1}=\frac{3\lambda^2+2A_1\lambda+B_1+D_1e^{-\lambda\tau_1}}{\lambda\left(D_1\lambda+E_1\right)e^{-\lambda\tau_1}}-\frac{\tau_1}{\lambda}$$

$$=-\frac{3\lambda^2+2A_1\lambda+B_1}{\lambda\left(\lambda^3+A_1\lambda^2+B_1\lambda+C_1\right)}+\frac{D_1}{\lambda\left(D_1\lambda+E_1\right)}-\frac{\tau_1}{\lambda}$$

Thus

$$\operatorname{Re}\left(\frac{d\lambda}{d\tau}\right)^{-1}\Bigg|_{\lambda=i\omega_0}=\operatorname{Re}\left(-\frac{B_1-3\omega_0^2+2A_1\omega_0 i}{i\omega_0\left[\left(C_1-A_1\omega_0^2\right)+i\omega_0\left(B_1-\omega_0^2\right)\right]}+\frac{D_1}{i\omega_0\left(D_1\omega_0 i+E_1\right)}\right)$$

$$=\frac{2D_1^2\omega_0^6+\left(3E_1^2+A_1^2D_1^2-2B_1D_1^2\right)\omega_0^4+\left(2A_1^2-4B_1\right)E_1^2\omega_0^2+\left(B_1^2E_1^2-C_1^2D_1^2-2A_1C_1E_1^2\right)}{\left[\omega_0^2\left(B_1-\omega_0^2\right)^2+\left(C_1-A_1\omega_0^2\right)^2\right]\left[\left(D_1\omega_0\right)^2+E_1^2\right]}$$

Let $M=\omega^2$

$$f\left(\omega\right)=2D_1^2\omega_0^6+\left(3E_1^2+A_1^2D_1^2-2B_1D_1^2\right)\omega_0^4$$
$$+\left(2A_1^2-4B_1\right)E_1^2\omega_0^2+\left(B_1^2E_1^2-C_1^2D_1^2-2A_1C_1E_1^2\right)$$

Then

$$G(M)=f(\omega)$$
$$=2D_1^2M^3+\left(3E_1^2+A_1^2D_1^2-2B_1D_1^2\right)M^2$$
$$+\left(2A_1^2-4B_1\right)E_1^2M+\left(B_1^2E_1^2-C_1^2D_1^2-2A_1C_1E_1^2\right)$$

And

$$G'=2\left[3D_1^2M^2+\left(3E_1^2+A_1^2D_1^2-2B_1D_1^2\right)M+\left(A_1^2-2B_1\right)E_1^2\right]$$
$$\Delta=\left(3E_1^2+A_1^2D_1^2-2B_1D_1^2\right)^2-12\left(A_1^2-2B_1\right)D_1^2E_1^2$$
$$=\left[3E_1^2-\left(A_1^2-2B_1\right)D_1^2\right]^2\geq 0$$

$G'=0$ has two real roots, which take the form

$$M_1=\frac{-\left(3E_1^2+\left(A_1^2-2B_1\right)D_1^2\right)+\sqrt{\Delta}}{6D_1^2}<0$$
$$M_2=\frac{-\left(3E_1^2+\left(A_1^2-2B_1\right)D_1^2\right)-\sqrt{\Delta}}{6D_0^2}<0$$

From above, we can get that $G(M)$ monotonously increases in $(M_1,+\infty)$, which means that $f(\omega)$ monotonously increases in $(0,+\infty)$.And as we know

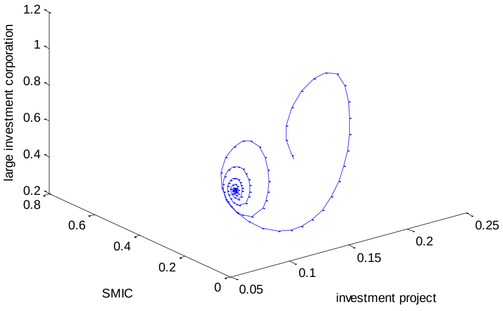
$$f(0)=B_1^2E_1^2-C_1^2D_1^2-2A_1C_1E_1^2>0$$

We have $f(\omega)>0$ for $\omega>0$. Then we obtain

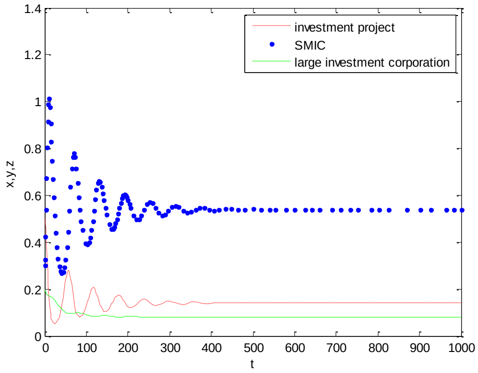
$$sign\frac{d\left(\operatorname{Re}\lambda(\tau)\right)}{d\tau}\bigg|_{\tau=\tau_n}=sign\operatorname{Re}\left(\frac{d\lambda}{d\tau}\right)^{-1}\bigg|_{\lambda=i\omega_0}>0$$

While $\tau_{1,0}$ is the minimum $\tau_n$ at which the real parts of these roots are zero. So $E_3$ is locally asymptotically stable if $\tau_1\in\left[0,\ \tau_{1,0}\right)$ and unstable if $\tau>\tau_0$.

From above analysis, the system weather stability is decided by delay coefficient. If $\tau_1\in\left[0,\ \tau_{1,0}\right)$ $E_3$ is locally asymptotically stable. Namely, with the increase of investment projects, to some extent, the number of SMIC and large investment corporations is stability. See as fig2(c-d).If $\tau>\tau_0$, $E_3$ is unstable. The number of investment corporations has periodic change as the change of interment projects.,see as fig3(e-f).

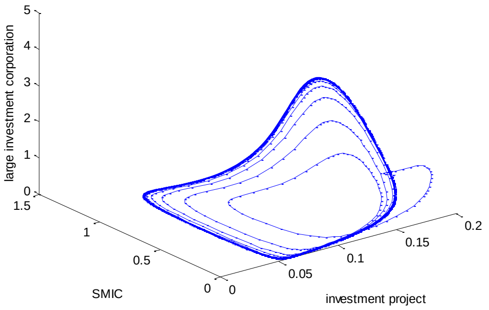


(c)

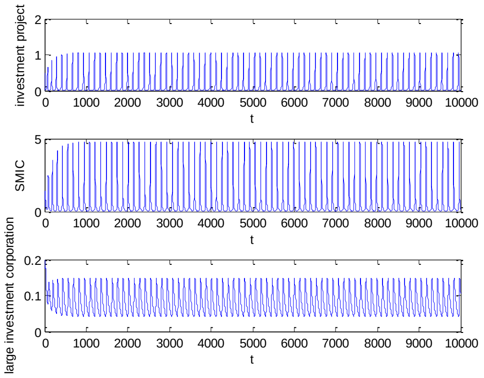


(d)

Fig2(c-d) $\tau_1<\tau_{1,0}$, $E_3$ is asymptotically stable.



(e)

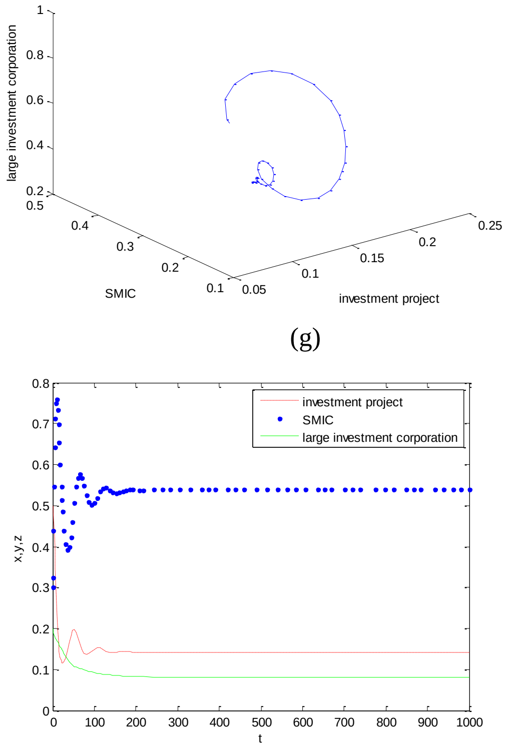


(f)

Fig3.(e-f) $\tau_1>\tau_{1,0}$ Bifurcation periodic solution form the equilibrium point $E_3$ occur.

### 3.3. $\tau_2\neq 0$

System simulation diagram is as follows



(g)

(h)

Fig4.(g-h）$\tau_1 = 0$，$\tau_2 \neq 0$，$E_3$ is asymptotically stable.

We found that time delay $\tau_2$ will not affect the stability of the system. The stability of the system has no matter with the accumulation of the time delay which the SIMC develop into the large investment corporations.

## 4. Summary

The article reveals the number of investment corporations with the change of the number of investment projects; through establish a model to analysis the relationship between investment corporations and investment projects. By adding time-delay to the investment model, we can find the market may rise to confusion for investment corporations entering market in different time. As time delay within the threshold, the system will reach the steady-state through a longer time. The system will have a period solution while time delay passes the threshold. What we study can provide a theoretical foundation for decision-makers choosing the time of entering the market, and provides some suggestions for managers to regulate the market better at macro point. The article doesn't take into account the complex relationship between investment and income, and we will consider it in the further research.

## References

[1] Hannan,M.T, Freeman,J.H, The population ecology of organization, The American Journal of Sociology (1977)929 - 964

[2] Moore, J. F, Predators and prey: a new ecology of competition, Harvard business review (1993)75-86

[3] Hong-Xing Yao, The Stability Analysis of Duopoly Investment Model with Bounded Rationality Based on China's Entry into the WTO, International Journal of Nonlinear Science Vol.3(2007) No.1,pp.44-51

[4] Ke, X.L., Shi, K.Stability and bifurcation in a simple heterogeneous asset pricing model, Economic Modeling 26 (2009) 680–688

[5] He,X.Z., Li,k., Market stability switches in a continuous-time financial market with heterogeneous beliefs, Economic Modeling 26 (2009) 1432–1442

[6] Yuan-Yuan Chen, Stability and Hopf bifurcation analysis in a prey–predator system with stage-structure for prey and time delay, Chaos, Solutions and Fractals 38 (2008) 1104–1114

[7] Kar,T.K, Stability and bifurcation of a prey–predator model with time delay, C. R. Biology 332 (2009) 642–651

[8]Hassard,B, Kazarinoff,D. Theory and application of Hopf bifurcation. Cambridge, Cambridge University Press; 1981

## Author Biographies

**H.X.Yao** male,Yangzhou Jiangsu Province of China, professor,,major research direction:the analysis of the complexity of economic systems modeling

**K.Liu** male, Yantai,Shandong Province of China,Master,research direction: the complexity of the financial system and Chaos Control.

**R.Liu** famle, Shangqiu,Henan Province of China ,Master,research direction: the complexity of the financial system and Chaos Control.

# Stability Analysis in a Pollution Defensive Model with Two Time Delays

H.X. Yao[1], Rui Liu[1,2], Kai Liu[1]

[1]Faculty of Science, Jiangsu University, Zhenjiang, Jiangsu Province, China, 212013
hxyao@ujs.edu.cn lr_smile@163.cn l22k77@163.com

***Abstract:*** In this paper a three-dimensional pollution defensive model with two time delays in the Chemical Industrial Area (CIA) is considered. The model is based on the interaction among chemical firms, pollution and capital stock in the CIA. The profit from chemical products is used for both defensive expenditure and an increase in capital stock. It shows that Hopf bifurcation occurs at the equilibrium point when the time delay reaches that point or the pollution defensive against the impact of chemical production is insufficient. In other words, if we do not guard against pollution sufficiently or control the production of chemical firms, it will lead to destabilization. Numerical simulations are given to illustrate the results.

***Keywords:*** Stability; Hopf bifurcation; Time delay; Pollution defensive

## 1. Introduction

Recent literature has demonstrated complicated relations between environment and capital. Becker [1] has examined the trade-off between capital accumulation and environmental quality through an analysis of regular maximum programs in the framework developed by Brock. The result is a constant utility path supported by government-imposed effluent charges and environmental rentals, sufficient conditions for a regular maximum path to satisfy the Hartwick Rule in calculating the combination of capital and environmental quality left for future generations. Similarly, a dynamic general equilibrium model based on environmental resources has been developed in a small open economy [2]. The result shows that the level of consumption and the fraction of income devoted to maximize the long-run welfare depend on both consumption level and environmental quality. Furthermore the possibility that both consumption and production affect the environment has been taken into account [3]. Consider the endogenous-growth model with physical and human capital accumulation [4]: the result shows that parameters on preferences, technologies and depreciation rates as well as fiscal policy are relevant to determine qualitatively the dynamic behavior of the economy. This paper considers a three-dimensional environmental defensive expenditures' model with time-delay bases demonstrating the interaction among visitors, quality of ecosystem goods and capital in protected areas [5].

This paper formulates a simple model with a unique positive equilibrium (when it exists) among the variable of state considered;moreover, this equilibrium is always stable. The aim of this work is to analyze how the stability of the equilibrium changes when two time-delays are considered because the dynamics of the pollution and capital stock at time $t$ depend on the profits of chemical production. In this model, we can see how such stability changes give rise to the Hopf bifurcation when time-delay passes through a sequence of critical values. The Hopf bifurcation allows us to find the

existence of a region of instability in the neighborhood of a fixed point where the manager of the CIA can stabilize the system if the time-delay is sufficiently short, but the model will become unstable when time delay is too long.

This paper is organized as follows: in Section 2 the model is presented; in Section 3 the fixed point, stability analysis and the existence of Hopf bifurcation are presented; and in Section 4, numerical simulations are presented.

## 2. The Model

The model, referring to the generic Chemical Industrial Area (CIA), has three variables: the production of chemical firm $V(t)$ in the CIA, the pollution $P(t)$ and the capital stock $K(t)$ intended as structures.

### 2.1 The production $V$ :

$$\dot{V}(t)=aK(t)-bP^2(t)-cV^2(t) \tag{1}$$

That parameter $a>0$ represents the production of unit capital stock, and $P(t)$ the current pollution. With the pollution growing, the production will decease; hence the coefficient is $b$. The production of the CIA never grows continuously, and then, $c$ represents recession coefficient.

### 2.2 The pollution $P$ :

Following Becker [1] and Cazzavillan and Musu [6], the pollution is defined as less than the maximum tolerable pollution $\bar{P}$ : that is, $0\le P(t)\le \bar{P}$

We assume that a constant proportion $0<r<1$ of the pollution is assimilated at each time $t$. Moreover, supposing that the pollution $P$ increases in proportion to the production $V$ of the CIA, the increase coefficient is $d$. When no resources are devoted to abatement expenditure, the CIA influences the pollution only by controlling production

$V(t)$.

$$\dot{P}(t)=dV(t)-rP(t) \qquad (2)$$

Chemical production makes a positive impact on this pollution. Therefore, chemical firms use a share $0<\rho<1$ of their profit to defend the environmental resources in the CIA in which the profit of one unit product is $n$ .This expenditure is directly proportional to the pollution. Therefore, the dynamics of the pollution is

$$\dot{P}(t)=dV(t)-rP(t)-mn\rho V(t) \qquad (3)$$

The parameter $m>0$ is a constant parameter determining how an additional unit of defensive expenditure decreases the pollution.

### 2.3 The capital stock $K$ :

The other share $(1-\rho)$ of the total profit of chemical product is used to increase capital stock

$$\dot{K}(t)=n(1-\rho)V(t)-\delta K(t) \qquad (4)$$

Capital stock is assumed to depreciate at the rate $\delta>0$. Considering two time delays, the model is formulated as follows:

$$\begin{cases}\dot{V}(t)=aK(t)-bP^2(t-\tau_2)-cV^2(t)\\ \dot{P}(t)=dV(t-\tau_1)-rP(t-\tau_2)-mn\rho V(t-\tau_1)\\ \dot{K}(t)=n(1-\rho)V(t-\tau_1)-\delta K(t)\end{cases} \qquad (5)$$

## 3. Qualitative Behavior of the Model

### 3.1 The equilibrium point:

It's easy to know that the equilibrium points are these:

$$F^0(0,0,0), \quad F^*(V^*,P^*,K^*)$$

where

$$V^*=\frac{anr^2(1-\rho)}{b\delta(d-mn\rho)^2+c\delta r}, P^*=\frac{d-mn\rho}{r}V^*, K^*=\frac{n(1-\rho)}{\delta}V^*$$

Obviously, $V^*, E^*$ and $K^*$ are only determined by the model (5), and they are always positive, if $(d-mn\rho)>0$.

### 3.2 Stability analysis:

The linearization of the model in the neighborhood of the positive equilibrium $F^*$ yields:

$$\begin{pmatrix}\dot{V}(t)\\ \dot{P}(t)\\ \dot{K}(t)\end{pmatrix}=\eta_1\begin{pmatrix}V(t)-V^*\\ P(t)-P^*\\ K(t)-K^*\end{pmatrix}+\eta_2\begin{pmatrix}V(t-\tau_1)-V^*\\ P(t-\tau_1)-P^*\\ K(t-\tau_1)-K^*\end{pmatrix}+\eta_3\begin{pmatrix}V(t-\tau_2)-V^*\\ P(t-\tau_2)-P^*\\ K(t-\tau_2)-K^*\end{pmatrix} \qquad (6)$$

where

$$\eta_1=\begin{pmatrix}0&0&a\\0&0&0\\0&0&-\delta\end{pmatrix}, \eta_2=\begin{pmatrix}-2cV^*&0&0\\(d-mn\rho)e^{-\lambda\tau_1}&0&0\\n(1-\rho)e^{-\lambda\tau_1}&0&0\end{pmatrix}$$

$$\eta_3=\begin{pmatrix}0&-2bP^*e^{-\lambda\tau_2}&0\\0&-re^{-\lambda\tau_2}&0\\0&0&0\end{pmatrix}$$

So the characteristic equation of model (5) is

$det(\lambda I-\eta_1-\eta_2e^{-\lambda\tau_1}-\eta_3e^{-\lambda\tau_2})=0$,

which leads to

$$D(\lambda,\tau_1,\tau_2)=S(\lambda)e^{-\lambda(\tau_1+\tau_2)}+R(\lambda)e^{-\lambda\tau_1}+Q(\lambda)e^{-\lambda\tau_2}+P(\lambda)=0 \qquad (7)$$

where

$$S(\lambda)=2bP^*(\lambda+\delta)(d-mn\rho)-anr(1-\rho);$$
$$R(\lambda)=-an\lambda(1-\rho);$$
$$Q(\lambda)=r(\lambda+\delta)(\lambda+2cV^*);$$
$$P(\lambda)=\lambda(\lambda+\delta)(\lambda+2cV^*);$$

### 3.3 The case $\tau_1=\tau_2=0$:

So, the characteristic polynomial is

$$D(\lambda,0,0)=S(\lambda)+R(\lambda)+Q(\lambda)+P(\lambda)=\lambda^3+A_0\lambda^2+A_1\lambda+A_2=0 \qquad (8)$$

where

$$A_0=2cV^*+\delta+r;$$
$$A_1=2cV^*(r+\delta)+2bP^*(d-mn\rho)+r\delta-an(1-\rho);$$
$$A_2=2b\delta P^*(d-mn\rho)-anr(1-\rho)+2cr\delta V^*;$$

According to the Routh-Hurwitz criterion [8], the equilibrium point is stable if, and only if

$$H_1: A_0>0 \quad A_0A_1-A_2>0;$$

### 3.4 The case $\tau_1\neq 0, \tau_2=0$ :

Let $\tau_2=0$ in Eq. (7), the characteristic polynomial becomes

$$\begin{aligned}D(\lambda,\tau_1,0)&=[S(\lambda)+R(\lambda)]e^{-\lambda\tau_1}+Q(\lambda)+P(\lambda)\\&=\lambda^3+B_0\lambda^2+B_1\lambda+B_2+[B_3\lambda+B_4]e^{-\lambda\tau_1}=0\end{aligned} \qquad (9)$$

where

$$B_0=r+\delta+2cV^*;$$
$$B_1=2cV^*(r+\delta)+r\delta;$$
$$B_2=2cr\delta V^*;$$
$$B_3=2bP^*(d-mn\rho)-an(1-\rho);$$
$$B_4=2b\delta P^*(d-mn\rho)-anr(1-\rho);$$

**Theorem 1.** Eq. (9) has a unique pair of purely imaginary roots if $(B_2+B_4)(B_2-B_4)<0$

**Proof:** If $\lambda=i\omega,\omega>0$ is a root of (9), separating real and imaginary parts, there will be the following equations:

$$\begin{cases}B_0\omega^2-B_2=B_4\cos\omega\tau+B_3\omega\sin\omega\tau\\ \omega^3-B_1\omega=B_3\omega\cos\omega\tau-B_4\sin\omega\tau\end{cases} \qquad (10)$$

Squaring and adding both equations above

$$\omega^6+Q_1\omega^4+Q_2\omega^2+Q_3=0 \qquad (11)$$

where

$$Q_1 = B_0^{\,2} - 2B_1,\ Q_2 = B_1^{\,2} - 2B\,B_0 - B_2^{\,2};\ Q_3 = B_3^{\,2} - B_2$$

leads to:

$$\begin{aligned} Q_1 &= B_0^{\,2} - 2B_1 \\ &= (\delta + r + 2cV^*)^2 - 2(\delta r + 2crV^* + 2c\delta V^* - an(1-\rho)) \\ &= 4c^2V^{*2} + \delta^2 + r^2 + 2an(1-\rho) > 0 \end{aligned} \tag{12}$$

$$Q_3 = B_2^{\,2} - B_4^{\,2} = (B_2 + B_4)(B_2 - B_4) < 0 \tag{13}$$

Then the conditions of this theorem imply that there is a unique positive root $\omega_0$ satisfying Eq. (9). That is, it has a unique pair of purely imaginary roots $\pm i\omega_0$.

From Eq. (10) $\tau_{1n}$ can be obtained

$$\tau_{1n} = \frac{1}{\omega_0} cos^{-1} \frac{B_3\omega_0^{\,4} + (B_0B_4 - B_1B_3)\omega_0^{\,2} - B_2B_4}{B_4^{\,2} + B_3^{\,2}\omega_0^{\,2}} + \frac{2k\pi}{\omega_0}$$
$$k = 0,1,2,... \tag{14}$$

**Theorem 2.** If the following conditions

$$H_2 : (B_2 + B_4)(B_2 - B_4) < 0\,;\ B_0(B_1 + B_3) > B_2 + B_4\,;$$
$$(B_1^{\,2} - 2B_0B_2)B_4^{\,2} > B_2^{\,2}B_3^{\,2}\,;$$

hold, model (5) undergoes Hopf bifurcation at $F^*(V^*, E^*, K^*)$ when $\tau = \tau_{10}$; furthermore, $F^*$ is locally asymptotically stable if $\tau_1 \in [0, \tau_{10})$, but unstable if $\tau_1 > \tau_{10}$.

**Proof.** It has been proved that, when $\tau_1 = 0$, all roots of Eq. (7) have negative real parts: that is to say, the equilibrium $F^*$ is locally stable for $\tau_1 = 0$. Subsequently, when $\tau_1 < \tau_{10}$, $F^*$ is still stable.

Then, if $Re\left(\frac{d\lambda}{d\tau_1}\right)\Big|_{\tau=\tau_{10}} > 0$, it indicates that when $\tau_1 > \tau_{10}$, at least a characteristic root with a positive real part will exist. According to the conditions of Hopf bifurcation theorem, the periodic solutions will occur when $\tau_1 > \tau_{10}$.

Differentiating Eq. (9) with $\tau_1$, it is as follows:

$$[3\lambda^2 + 2B_0\lambda + B_1 + B_3e^{-\lambda\tau_1} - \tau_1(B_3\lambda + B_4)e^{-\lambda\tau_1}]\frac{d\lambda}{d\tau_1} = \lambda(B_3\lambda + B_4)e^{-\lambda\tau_1} \tag{15}$$

that is,

$$\begin{aligned} \left(\frac{d\lambda}{d\tau_1}\right)^{-1} &= \frac{3\lambda^2 + 2B_0\lambda + B_1}{\lambda(B_3\lambda + B_4)e^{-\lambda\tau_1}} - \frac{\tau_1}{\lambda} \\ &= -\frac{3\lambda^2 + 2B_0\lambda + B_1}{\lambda(\lambda^3 + B_0\lambda^2 + B_1\lambda + B_2)} + \frac{B_3}{\lambda(B_3\lambda + B_4)} - \frac{\tau_1}{\lambda} \end{aligned} \tag{16}$$

Thus,

$$Re\left(\frac{d\lambda}{d\tau_1}\right)^{-1}\Bigg|_{\lambda=i\omega} = Re\left(-\frac{B_1 - 3\omega^2 + i2B_0\omega}{i\omega[(B_2 - B_0\omega^2) + i\omega(B_1 - \omega^2)]} + \frac{B_3}{i\omega(iB_3\omega + B_4)}\right)$$
$$= Re\left(\frac{2B_3^{\,2}\omega^6 + (3B_4^{\,2} + B_0^{\,2}B_3^{\,2} - 2B_1B_3^{\,2})\omega^4 + (2B_0^{\,2} - 4B_1)B_4^{\,2}\omega^2 + (B_1^{\,2}B_4^{\,2} - B_2^{\,2}B_3^{\,2} - 2B_0B_2B_4^{\,2})}{[\omega^2(B_1 - \omega^2)^2 + (B_2 - B_0\omega^2)^2][(B_3\omega)^2 + B_4^{\,2}]}\right)$$

We can rewrite the numerator as follows.

$$\begin{aligned} f(\omega) = {} & 2B_3^{\,2}\omega^6 + (3B_4^{\,2} + B_0^{\,2}B_3^{\,2} - 2B_1B_3^{\,2})\omega^4 \\ & + (2B_0^{\,2} - 4B_1)B_4^{\,2}\omega^2 + (B_1^{\,2}B_4^{\,2} - B_2^{\,2}B_3^{\,2} - 2B_0B_2B_4^{\,2}) \end{aligned} \tag{17}$$

Let $\eta = \omega^2$, then

$$\begin{aligned} G(\eta) = f(\omega) = {} & 2B_3^{\,2}\eta^3 + (3B_4^{\,2} + B_0^{\,2}B_3^{\,2} - 2B_1B_3^{\,2})\eta^2 \\ & + (2B_0^{\,2} - 4B_1)B_4^{\,2}\eta + (B_1^{\,2}B_4^{\,2} - B_2^{\,2}B_3^{\,2} - 2B_0B_2B_4^{\,2}) \end{aligned} \tag{18}$$

and

$$G'(\eta) = 2[3B_3^{\,2}\eta^2 + (3B_4^{\,2} + (B_0^{\,2} - 2B_1)B_3^{\,2})\eta + (B_0^{\,2} - 2B_1)B_4^{\,2}] \tag{19}$$

for $G'(\eta)$,

$$\begin{aligned} \Delta &= (3B_4^{\,2} + (B_0^{\,2} - 2B_1)B_3^{\,2})^2 - 12(B_0^{\,2} - 2B_1)B_3^{\,2}B_4^{\,2} \\ &= [3B_4^{\,2} - (B_0^{\,2} - 2B_1)B_3^{\,2}]^2 \geq 0 \end{aligned} \tag{20}$$

$G'$ has two real roots, which take the for

$$\eta_1 = \frac{-(3B_4^{\,2} + (B_0^{\,2} - 2B_1)B_3^{\,2}) + \sqrt{\Delta}}{6B_3^{\,2}} < 0\,,$$

$$\eta_1 = \frac{-(3B_4^{\,2} + (B_0^{\,2} - 2B_1)B_3^{\,2}) - \sqrt{\Delta}}{6B_3^{\,2}} < 0$$

From the above, clearly $G'(\eta)$ increases monotonously in $(\eta_1, +\infty)$. As far as concerned that $f(0) = B_1^{\,2}B_4^{\,2} - B_2^{\,2}B_3^{\,2} - 2B_0B_2B_4^{\,2} > 0$, there will be $f(\omega) > 0$, for $\omega > 0$, and it will have:

$$sign\frac{d(Re\,\lambda(\tau_1))}{d\tau_1}\Bigg|_{\tau_1=\tau_{1n}} = sign\,Re\left(\frac{d\lambda}{d\tau_1}\right)^{-1}\Bigg|_{\lambda=i\omega} > 0 \tag{21}$$

Theorem2 states that when these conditions obtain, the Hopf bifurcation will occur while $\tau_{10}$ is the minimum $\tau_{1n}$ at which the real parts of these roots are zero. That is to say, the model undergoes Hopf bifurcation at the equilibrium $F^*$ when $\tau_1 = \tau_{10}$, and, regarding the impact of production on pollution, the defensive expenditures is not elevated. In other words, if the pollution is not sufficiently decreased, it will reach one destabilization at the fixed points when $\tau_1 > \tau_{10}$.

**Theorem 3:** For Eq. (9), it can refer to:
If $H_1$ and $H_2$ hold, when $\tau_1 \in [0, \tau_{10})$ all roots of Eq. (9) have negative real parts, and when $\tau_1 > \tau_{10}$ Eq. (9) will have at least one root with positive real part.

**Proof:** As $H_1$ and $H_2$ hold, then the equilibrium of the Eq.(9) is stable and Eq.(9) has complex roots with negative real parts for $\tau_1 = 0$, and also for $\tau_1 = \tau_{10}$. Eq.(9) has purely imaginary roots, and the real parts of the root changes continuously with the increase of $\tau_1$ because of $sign\frac{d(Re\,\lambda(\tau_1))}{d\tau_1}\Big|_{\tau_1=\tau_{1n}} > 0$, so for $\tau_1 \in [0, \tau_{10})$ all roots of Eq.(9) have negative real parts and Eq.(9) has at least one root with positive real parts when $\tau_1 > \tau_{10}$.

### 3.5 The case $\tau_1 \neq 0, \tau_2 \neq 0$ :

Next, we return to the Eq. (7) with $\tau_2 > 0$ and $\tau_1$ in stable regions. Regard $\tau_2$ as a parameter following Ruan and Wei [10]

**Theorem 4:** If all roots of Eq. (9) have negative parts for $\tau_1 > 0$, then there will exist a $\tau_2^*(\tau_1) > 0$, subject to all roots of Eq. (7), and it will have negative real parts when $0 \le \tau_2 < \tau_2^*(\tau_1)$.

**Proof.** Following the theorem2.1 of Ruan and Wei, the left of Eq.(7)is analytic in $\lambda$ and $\tau_2$, and when $\tau_2$ varies, the sum of the multiplicities of zeros of the left of Eq.(7) in the open half-plane will change only if a zero is on, or crosses, the imaginary axis.

**Theorem 5:** Assume $H_1$ holds true; if $H_2$ holds, there exists $0 < \tau^*_{10} < \tau_{10}$ and $\tau_2 = \tau_2^*(\tau_1)$, then for any $\tau_1 \in [0, \tau^*_{10})$,the equilibrium of model (5) is locally asymptotically stable when $\tau_2 \in [0, \tau_2^*(\tau_1))$.

**Proof.** According to Theorem3 and Theorem4, there will be a result.

It's clear that the Hopf bifurcation occurs at $\tau_2^*(\tau_1)$ if it holds the conditions of Theorem 4 or Theorem 5 and, also, there may be a lot of stability-switches. If $\tau_1$ is in an unstable region, there may not exist $\tau_2^*(\tau_1)$ which makes the model (5) stable if $0 \le \tau_2 < \tau_2^*(\tau_1)$, but unstable if $\tau_2 > \tau_1^*(\tau_1)$.

## 4. Numerical Simulation

This section shows some numerical simulations at different value of $\tau_1$ and $\tau_2$.

Considering system (5) with the following parameters $a = 4$, $b = 0.5$, $c = 0.1$, $d = 0.9$, $r = 0.1$, $n = 1$, $\rho = 0.8$, $m = 1$, $\delta = 0.1$, initial values $V = 2, P = 1, K = 3$, the conditions of Theorem 1,2 hold. Supposing $\rho = 0.8$, the fixed point is $F^* = (13.75, 13.75, 27.50)$.

### 4.1 The case $\tau_1 \ne 0, \tau_2 = 0$

There is $\omega_0 = 0.1445$, $\tau_{10} = 12.6813$ period $T = 43.4782$.

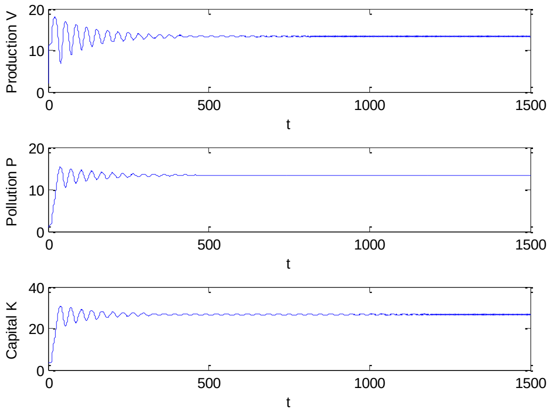


Fig 1when $\tau_2 = 0$ and $\tau_1 = 10 < \tau_{10}$

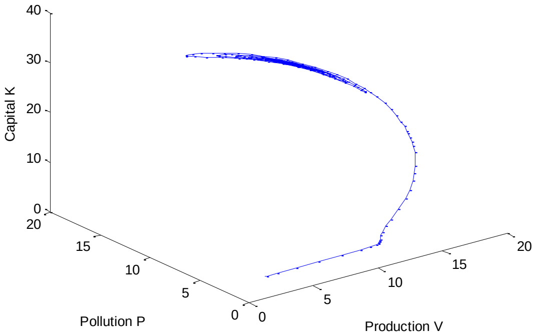


Fig2:when $\tau_2 = 0$ and $\tau_1 = 10 < \tau_{10}$

Figures1,2 show that when $\tau_2 = 0$ and $\tau_1 = 10 < \tau_{10}$, the chemical production, the pollution and the capital stock tend to be stable.

Figures3,4 show that when $\tau_2 = 0$ and $\tau_1 = 14 > \tau_{10}$, the chemical production, the pollution and the capital stock tend to be periodic solutions.

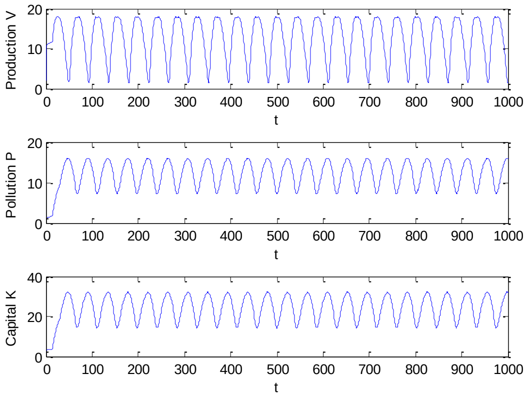


Fig3: when $\tau_2 = 0$ and $\tau_1 = 14 > \tau_{10}$

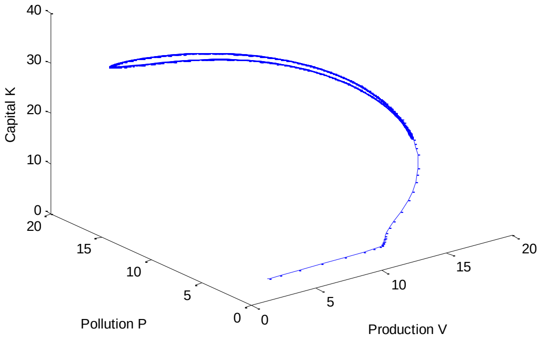


Fig4: when $\tau_2 = 0$ and $\tau_1 = 14 > \tau_{10}$

### 4.2 The case $\tau_1 \neq 0, \tau_2 \neq 0$

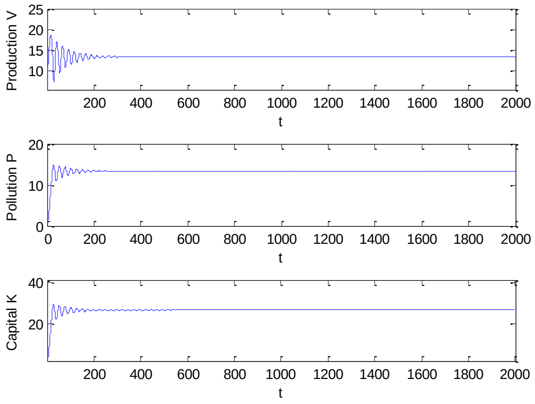


Fig5: when $\tau_1 = 5.7 < \tau^*_{10}$ and $\tau_2 = 0.7 < \tau_2^*(\tau_1)$

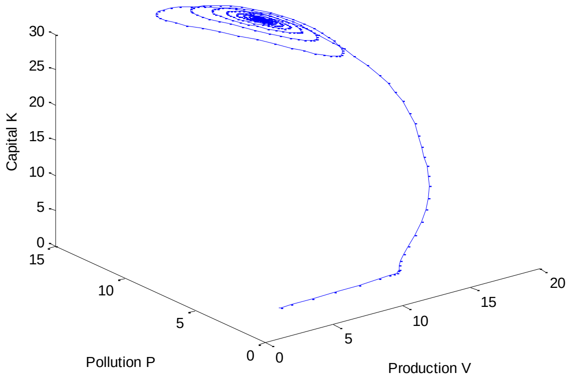


Fig6:when $\tau_1 = 5.7 < \tau^*_{10}$ and $\tau_2 = 0.7 < \tau_2^*(\tau_1)$

Figures5,6 show that when $\tau_1 = 5.7 < \tau^*_{10}$ and $\tau_2 = 0.7 < \tau_2^*(\tau_1)$, the chemical production, the pollution and the capital stock tend to be stable.

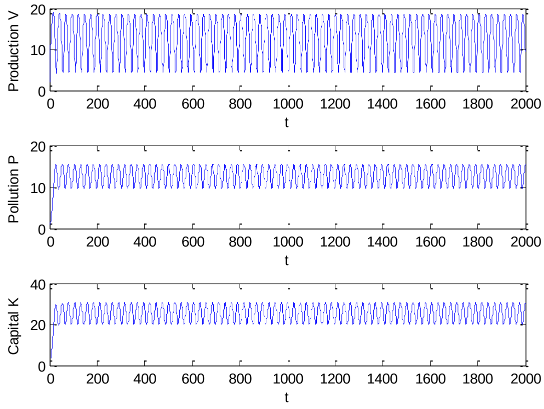


Fig7: when $\tau_1 = 5.7 < \tau^*_{10}$ and $\tau_2 = 1.0081 > \tau_2^*(\tau_1)$

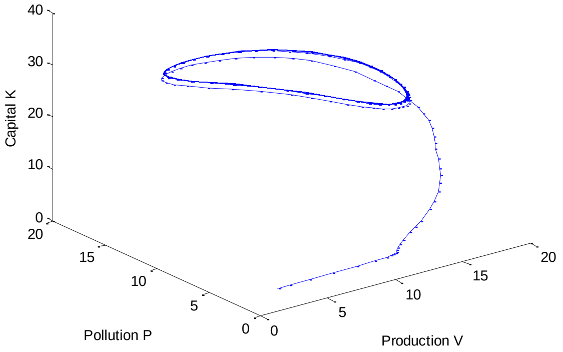


Fig8: when $\tau_1 = 5.7 < \tau^*_{10}$ and $\tau_2 = 1.0081 > \tau_2^*(\tau_1)$

## 5. Conclusion

The present work, starting from a simple model with a positive equilibrium, shows that a delay may generate instability and, as a consequence, problems in the sustainability of the CIA's decision if the condition $(d - mn\rho) < 0$ occurs: that is, the pollution defensive expenditure is not elevated sufficiently. Furthermore, if the conditions (1) $(C+E)(C-E) < 0$ ; (2) $A(B+D) > C+E$ ; (3) $(B^2 - 2AC)E^2 > C^2D^2$ hold, $\tau_1 > \tau_{10}$, the Hopf bifurcation occurs but, then, if $\tau_2 > 0$ and $\tau_1$ exist in stable regions, it's clear that Hopf bifurcation will occur at $\tau_2^*(\tau_1)$ if the conditions of Theorem 4 or Theorem 5 hold. Further developments can be identified or analyzed in a model with two variable delays.

## References

[1] Rober A. Becker. Intergenerational equity: The capital-environment trade-off [J]. J Environ Econ Manage. 1982; 9:165-85.

[2] Rinaldo Brau, Alessandro Lanza, Stefano Usai. Tourism and sustainable economic development: macroeconomic

model and empirical methods [M]. New York:William Partt House, 2008.

[3] Carlo Carraro, Domenico Siniscalco. New direction in the economic theory of the environment [M], London: Cambridge University Press, 1997.

[4] Salvador Ortigueira. Fiscal policy in an endogenous growth model with human capital accumulation [J]. Journal of Monetary Economics, 1998, 45(2): 323-335

[5] Paolo Russu. Hopf bifurcation in an environmental defensive expenditures model with time

(6): 991 - 1003.

[8] Hale J K. Theory of Functional Differential Equations [M]. New York: Spring2 Verlag, 1977.

[9] E.Beretta,Y.Kuang, Geomitric stability switch criteria in delay differential models with delay dependent parameters [J]. SIAM J.Math.Anal.33(2002) 1144-1165.

[10] S.Ruan,J.Wei, On the Zeros of transcendental functions with applications to stability of delay differential equations with two delays [J]. Dyn.Contin. Discrete Impuls. Syst. Ser.A Math. Anal. 10(2003) 863-874.

Delay [J]. Chao, Solitons & Fractals. 2009; 42:3147-59.

[6] Cazzavillan G, Musu I. Transitional dynamics and uniqueness of the balanced-growth path in
a simple model of endogenous growth with an environmental asset [J]. FEEM Working Paper 62.2001; 2001.

[7] Freedman H I, Sree Hari Rao V. The Trade 2 off Between Mutual Interference and Time Lags
in Predator 2 prey Models [J]. Bull Math Biol, 1983, 45

## Author Biographies

**H.X.Yao** male,Yangzhou Jiangsu Province of China, professor,,major research direction is the complexity of economic systems modeling and anslysis

**Liu Rui** famle,Shangqiu,Henan Province of China ,Master,research direction: the complexity of the financial system and Chaos Control.

**Liu Kai** male, Yantai,Shandong Province of China,Master,research direction: the complexity of the financial system and Chaos Contro

# Detection of Shapes of Objects Using Sophisticated Image Processing Techniques

Dr. T.C. Manjunath, *Ph.D.* (*IIT Bombay*), *Fellow IETE*,
Principal, Atria Institute of Technology, Bangalore-24, Karnataka, India.
Emails : tcmanjunath@gmail.com ; tcmanjunath@rediffmail.com

***Abstract* —** This paper features an efficient method of performing the shape analysis of objects in a binary image using a technique called as the moments. Using these moments, we can compute the centroid, moment of inertia, principal angle, orientation of the captured objects in the image. The simulation results show the effectiveness of the developed method.

***Index Terms* —** Moments, Invariance, Centroid, Moment of Inertia, Principal angle, Orientation, Binary image, Gray scale image.

## I. Introduction

SHAPE analysis is a method of finding the shape of irregular objects using two types of descriptors called as the line descriptors and the area descriptors and is used when the objects are not polyhedral objects. For example, circles, spheres, ellipses, boundaries, curves, arcs, objects of irregular shapes. There are two methods of performing the shape analysis of objects. viz., line descriptors & the area descriptors. Line descriptors are used to find out the length of the irregular boundary or curvature of an irregular object in a digital image in terms of pixels.

Area descriptors are used to find out the shape of the irregular object and its characteristic properties such as area, moments, central moments, centre of gravity (centroid), moment of inertia and the orientation of the object w.r.t. ( x , y ) axis of the image. The area descriptors are defined as the descriptors which are based on the analysis of the points enclosed by the boundary and are used to characterize the shape of the foreground region R in a image. In this paper, we discuss about the shape analysis concept of the captured objects in a 2 D image using moments [1].

The theory of moments provides an interesting method of describing the properties of an object in terms of its area, position and orientation parameters. The idea of moments was borrowed from the science of physics. In this paper, we discuss about a method of computing the features of objects in a 2D image.

The paper is organized as follows. A brief introduction about the shape analysis of objects was presented in the previous paragraphs. In section 2, a brief introduction about the foreground & background region in an image is dealt with. Shape analysis using moments is described briefly in section 3. Section 4 deals with the analytical / mathematical treatment of a simulation example. Section 5 shows the simulation results. Conclusions are presented in section 6 followed by the references.

## II. Foreground Region and Background Region in a Binary Image

Consider a binary image B ( k , j ) as shown in Fig. 1 which is obtained by the segmentation / thresholding / binarization of a digital image or a gray scale image I ( k , j ). Foreground is represented by 1's & background is represented by 0's. There was a circular object with a hole inside [2]. First, it was captured by a camera, digitized & then binarized. B ( k , j ) is the binary representation of the object with the hole inside & is as shown in the Fig. 1.

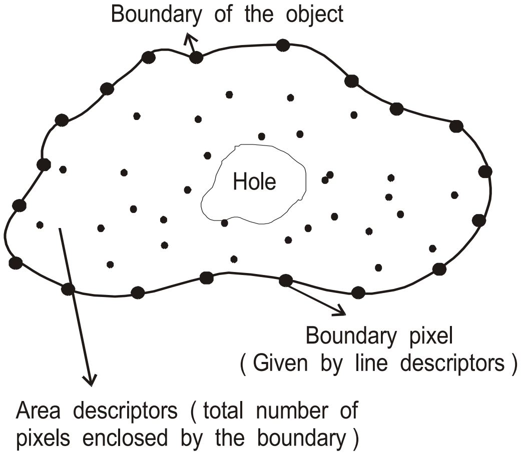
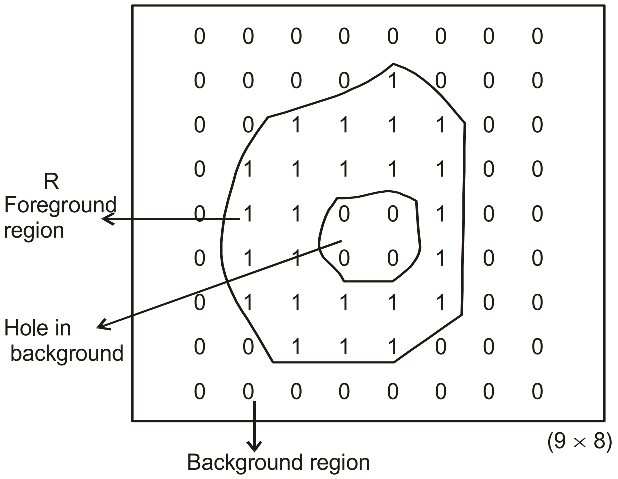

Boundary of the object
Hole
Boundary pixel
( Given by line descriptors )
Area descriptors ( total number of
pixels enclosed by the boundary )

Fig. 1 : 2D representation of a elliptical object in a image



| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 |
| 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 |
| 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 |
| 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Fig. 2 : A foreground region R in an binary image ;
White - 1 ( Foreground ), Black - 0 ( Background pixel )

A region R in a binary image is defined as a set of connected pixels as shown in Fig. 2, which are having the same gray level attribute. R is a connected set, i.e., for each pair of pixels in R, there is at least one path, which connects the pair. Foreground region is one connected set and background region forms another connected set. This makes sure that there are no breaks in the part. Since R is a connected set, R is a single part (with no breaks). Connected in the sense, there is a neighboring pixel, which is having the same gray scale value as the pixel considered. R can also have a hole in it. Now, we have to compute the shape of the alphabet O. In order to compute the shape of the alphabet O, we consider only the foreground region and the background region is neglected [3].

## III. Shape Analysis Using Moments

To analyze and characterize the shape of the given foreground region R, we compute some numbers. These numbers are called as the moments of the foreground region R. Moments gives the characteristic features of the objects such as the shape, area, centroid, moment of inertia and the orientation of the object in the image. There are four types of moments, viz., lower order moments, central moments, normalized central moments & the principal angle [4].

Moments are defined as the sequence of numbers which are used to characterize the shape of any object OR the sum of products of the row value **x** raised to the power k and column value **y** raised to the power j in R.

$$m_{kj} \overset{\Delta}{=} \sum_{\text{row , column} \in R} \text{row}^k \ \text{column}^j \qquad (1)$$

$$\overset{\Delta}{=} \sum_{x,\, y \in R} x^k y^j \ ; \ k \geq 0 ; \ j \geq 0 \qquad (2)$$

x and y : Row and column values of the pixel in the foreground region R.

Order of the moment = Sum of the powers = ( k + j ).

In calculation of moments, consider only the foreground pixels and ignore the background pixels.

### A. Lower Order Moments : Ordinary Moments, $m_{kj}$

Geometric meanings and physical significance of the LOM could be explained as follows. Let {$m_{kj}$} be the ordinary moments of the foreground region R of a binary image B ( k , j ), A being the area of the foreground region R and { $x_c$ , $y_c$} be the position of the centroid of the region R. Then [2], [5],

$$Area = A = m_{00}$$

$$x_c = \frac{m_{10}}{m_{00}} \qquad\qquad (3)$$

$$y_c = \frac{m_{01}}{m_{00}}$$

$$\left(x_c, y_c\right) = Centroid = \left\{\frac{m_{10}}{m_{00}}, \frac{m_{01}}{m_{00}}\right\}$$

- Zeroth order moment $m_{00}$ gives the area of the foreground region R or the count of the total number of pixels in the region R or it is the measure of the size of the region R.
- $m_{10}$ gives the first order moment ( lower order ) along x-axis.
- $m_{01}$ gives the first order moment ( lower order ) along y-axis.
- $m_{20}$ gives the second order moment ( lower order ) along x-axis.
- $m_{02}$ gives the second order moment ( lower order ) along y-axis.
- $m_{11}$ gives the product moment.
- $m_{22}$ gives the product moment.
- Normalize the first order moments with the zeroth order moments, i.e., divide the first order moments with the zeroth order moments $\frac{m_{10}}{m_{00}}, \frac{m_{01}}{m_{00}}$. This gives the centroid of the foreground region R.

The physical significance of lower order moments is, they just give the area and the centroid of the foreground region R.

### B. Central Moments of a Foreground Region R, $\mu_{kj}$

The geometric meanings and physical significance of central moments can be explained as follows. Central moments are called so because they are obtained using the centroid ( $x_c$ , $y_c$ ). The central moments of foreground region R are nothing but the ordinary moments $m_{kj}$ of the foreground region R, but, translated by an amount equal to the centroid so that the centroid ( $x_c$ , $y_c$ ) now coincides with the origin, as a result of which the first order central moments are equal to zero and are invariant to translations of the foreground region R ( i.e., $\mu_{10} = \mu_{01} = 0$ ). Let ( $x_{c,}$ $y_c$ ) be the centroid of a region R and let ( x, y ) be the row and column of a pixel p in R. The central moments $\mu_{kj}$ of R are given by

$$\mu_{kj} = \sum_{(x,y)\in R} (x - x_c)^k (y - y_c)^j \quad ; \ (k, j) \geq 0 \qquad (4)$$

The physical significance of the central moments are, they just give the area and the moment of inertia and they are invariant to translations, but are variant to scale changes and rotations [2], [6].

- Lower order central moments ( Zeroth ), $\mu_{00}$ is the same as $m_{00}$ ; i.e., gives the area of the region R or the size of the region, i.e., the number of pixel counts [16].
- Lower order central moments ( First ), $\mu_{10}$ : gives the first order central moment along x-axis = 0 ( since centroid coincides with origin ) & $\mu_{01}$ : gives the first order central moment along y-axis = 0 ( since centroid coincides with origin ).

During the calculation of the moments, each point ( x , y ) is shifted by an amount equal to the distance of the centroid from the origin ; so finally over the sum, the centroid coincides with the coordinate origin. Any translations of the region R are negated because of this shifting process. Since, central moments are invariant to translations, the first order central moments $\mu_{01}$ and $\mu_{10}$ are always = 0. Invariancy to translations means ; in the image, if the object is translated along x or y-axis the properties of the object remains the same. The second order central moments could be explained as follows :

- $\mu_{20}$ : Gives the second order central moment along x-axis. Gives the Moment of Inertia [ M I ] of the foreground region R about the x-axis.
- $\mu_{02}$ : Gives the second order central moment along y-axis. Gives the Moment of Inertia [ M I ] of the foreground region R about the y-axis.
- The x and y-axes pass through the centroid of the region R.
- $\mu_{11}$ is a product moment as it involves finding the product of ( x – $x_c$ ) & ( y – $y_c$ ) raised to a power after which, the products are summed to give $\mu_{11.}$
- $\mu_{22}$ is another product moment which is not of any physical significance [15].

### C. *Normalized Central Moments, $v_{kj}$*

The geometric meanings and physical significance can be explained as follows. Central moments are further normalized to produce another type of moments, called as the normalized central moments, which are invariant to scale changes of the foreground region R, in addition to translation invariance. If $\mu_{kj}$ are the central moments of region R and $v_{kj}$, then the normalized central moments $v_{kj}$ of R are given by [2]

$$v_{kj} = \frac{\mu_{kj}}{\mu_{00}^{(k+j+2)/2}} \quad ; (k, j) \geq 0 \qquad (5)$$

This property of invariancy to scale changes occurs because the area of the region R is scaled down by the factor of $\frac{(k+j+2)}{2}$. Invariancy to scale changes means ; in the image, if the object is zoomed in or zoomed out, the properties of the object remains the same.

- $v_{00}$ : Zeroth order NCM = 1, from this, we can come to a conclusion that the NCM are invariant to scale changes of R.
- $v_{10}$ and $v_{01}$ : First order NCM along x and y-axis = 0, since they are invariant to translations.
- $v_{20}$ and $v_{02}$ : Second order NCM along x and y-axis.
- $v_{11}$ and $v_{22}$ : Product moments, not of any physical significance.

The physical significance of NCM's are they are invariant to scale changes, in addition to translation invariance, but variant to rotations [7].

### D. *Principal Angle : Orientation or Inclination of the Region, $\phi$*

Lower order moments $m_{kj}$, central moments $\mu_{kj}$ and normalized central moments $v_{kj}$, characterize the region R and are invariant to translations and scaling of R, but are variant to rotations of the foreground region R.

Invariancy to rotations of R can also be obtained by finding the principal angle $\phi$ and is a measure of the orientation of the region R. The principal angle $\phi$ can be expressed in terms of the second order central moments and so does not involve additional calculations.

The physical significance of principal angle is ; it is invariant to rotations of the foreground region and gives the orientation of the foreground region R in the binary image. Invariancy to rotations means ; in the image, if the object is rotated by any amount, then, the properties of the object remains the same [8].

The physical interpretation of the principal angle $\phi$ is as follows. Draw a line L ($\beta$) through the centroid ( $x_{c,}$ $y_c$ ) of R at angle of $\beta$ with the x-axis. The moment of inertia of R about the line L($\beta$) depends on the angle $\beta$. Go on varying the angle $\beta$ as shown in the Fig. 3.

At one point, the MI of the object will be minimum. The angle at which the moment of inertia $I_R$ of region R is minimized is called the principal angle of R and is nothing but $\phi$ which is equal to $\beta$. It follows that the principal angle is well defined for an elongated object, but it becomes ambiguous when the object approaches a circular shape [9].

Principal angle $\phi$ of the foreground region R is defined as the angle at which the moment of inertia $I_R$ of region R is minimized. Mathematically, it is given by the formula [2]

$$\begin{aligned} \phi &= \frac{1}{2} a\tan 2\left(2\mu_{11}, \mu_{20} - \mu_{02}\right) \\ &= \frac{1}{2}\tan^{-1}\left(\frac{2\mu_{11}}{\mu_{20} - \mu_{02}}\right) \end{aligned} \qquad (6)$$

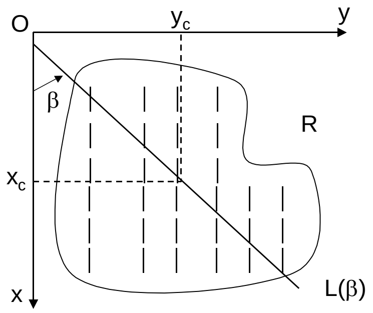


Fig. 3 : Principal angle of a region R

### E. *Invariant Moments*

Now, we know the centroid of R and the principal angle of R, i.e., ϕ. If we translate R by an amount = ( $-x_{c,}$ $-y_c$ ), then the centroid coincides with the origin. Rotate R by an angle of –ϕ ( clockwise ) ; then, the principal angle becomes zero. Now, if we take the moments of the region R, it will be seen that these moments are invariant to translations, rotations and scale changes [14]. The normalized moments of the resulting region will then be invariant to translations, rotations and scale changes of R. Such moments are called as invariant moments. Invariant moms are defined as the moments, which are invariant to translations, scale changes and rotations of the foreground region R [10].

### F. *Eccentricity*

It is the maximum chord length along the principal axes or major axis of object divided by the minimum chord length, which is perpendicular to chord length. The maximum chord length or major diameter D of an object O is defined as

$$\text{Eccentricity} = \frac{\text{Maxmimum chord length}}{\text{Minimum chord length}} \tag{7}$$

$$D = \max_{i,j} \sqrt{\left(x_j - x_i\right)^2 + \left(y_j - y_i\right)^2} \tag{8}$$

where $p_i = (x_i , y_i )$ and $p_j = (x_j , y_j )$ are the pixels in the boundary of the object O.

$$\text{Thinness} = \frac{(\text{Perimeter})^2}{2}$$

$$\text{Roundness} = \frac{\left(x_c{}^2 + y_c{}^2\right)^2}{A} \tag{9}$$

## IV. Simulation example (analytical calculation)

In this section, we consider a zig-zag object was captured by the camera, digitized, segmented & binarized and stored in the memory of the computer as shown in Fig. 4. It is necessary for us to compute the shape of the objects using the methods discussed in the previous section. A brief analytical / mathematical treatment to the problem considered is dealt with in this section. The following computations can be seen as below [11].

$$\begin{array}{cc|cccccccc|}
 & & \multicolumn{8}{c}{\text{Col} \quad j \quad \rightarrow} \\
 & & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
\hline
 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\text{Row} & 3 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\
k & 4 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\
\downarrow & 5 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\
 & 6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 & 7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 & 8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
\end{array}$$

Fig. 4 : An image of size ( 8 × 8 )

$$m_{00} = \sum_{x,\,y \in R} x^0\ y^0 = \text{Area}$$

$$= 10$$

$$m_{01} = \sum_{x,\,y \in R} x^0\ y^1$$

$$= 40$$

$$m_{10} = \sum_{x,\,y \in R} x^1\ y^0$$

$$= 40$$

$$m_{02} = \sum_{x,\,y \in R} x^0\ y^2 = \sum_{x,\,y \in R} y^2$$

$$= 174$$

$$m_{20} = \sum_{x,\,y \in R} x^2\, y^0 = \sum_{x,\,y \in R} x^2$$

$$= 166$$

$$m_{11} = \sum_{x,\,y \in R} x^1\ y^1$$

$$= 162$$

$$m_{22} = \sum_{x,\,y \in R} x^2\ y^2$$

$$= 2968$$

$$x_c = \frac{m_{10}}{m_{00}} = \frac{40}{10} = 4 \quad \text{and}$$

$$y_c = \frac{m_{01}}{m_{00}} = \frac{40}{10} = 4$$

Centroid , $(x_c\ y_c)$ = ( 4 , 4 )

$$\mu_{00} = \sum_{x,\,y \in R} (x-4)^0\ (y-4)^0 = 10$$

$$\mu_{02} = \sum_{x,\,y \in R} (x-4)^0\ (y-4)^2 = 14$$

$$\mu_{20} = \sum_{x,\,y\in R} (x-4)^2\ (y-4)^0$$

$$= 6$$

$$\mu_{11} = \sum_{x,\,y\in R} (x-4)^1\ (y-4)^1$$

$$= 2$$

$$\mu_{22} = \sum_{x,\,y\in R} (x-4)^2\ (y-4)^2$$

$$= 8$$

Calculation of normalized central moments as follows

$$v_{kj} \triangleq \frac{\mu_{kj}}{\mu_{00}^{(k+j+2)/2}}\ ;\ k \geq 0\ ,\ j \geq 0$$

$$v_{00} = \frac{\mu_{00}}{\mu_{00}^{(0+0+2)/2}} = \frac{\mu_{00}}{\mu_{00}^{1}} = 1$$

$$v_{10} = \frac{\mu_{10}}{\mu_{00}^{(1+0+2)/2}} = \frac{\mu_{10}}{\mu_{00}^{1.5}} = 0$$

$$v_{01} = \frac{\mu_{01}}{\mu_{00}^{(0+1+2)/2}} = \frac{\mu_{01}}{\mu_{00}^{1.5}} = 0$$

$$v_{02} = \frac{\mu_{02}}{\mu_{00}^{(0+2+2)/2}} = \frac{\mu_{02}}{\mu_{00}^{2}} = 0.140$$

$$v_{20} = \frac{\mu_{20}}{\mu_{00}^{(2+0+2)/2}} = \frac{\mu_{20}}{\mu_{00}^{2}} = 0.36$$

$$v_{11} = \frac{\mu_{11}}{\mu_{00}^{(1+1+2)/2}} = \frac{\mu_{11}}{\mu_{00}^{2}} = 0.02$$

$$v_{22} = \frac{\mu_{22}}{\mu_{00}^{(2+2+2)/2}} = \frac{\mu_{22}}{\mu_{00}^{3}} = 0.008$$

$$\text{Principal angle} = \phi = \frac{1}{2}\tan^{-1}\left(\frac{2\mu_{11}}{\mu_{20}-\mu_{02}}\right)$$

$$= \frac{1}{2}\tan^{-1}\left(\frac{2\times 2}{6-14}\right)$$

$$= 76.72°$$

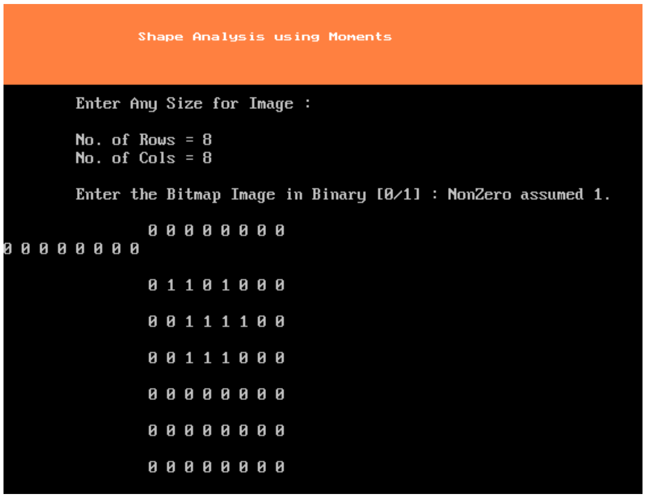
All the results are summarized as shown in the table 1. A graphical user interface program was developed in C / C++ language and the code was compiled and run [12].

On running the code, the following screens as shown below appeared, i.e., the inputs & the output screens, which are nothing but the simulation results as shown in the Figs. 5 - 8 respectively [13].

| Type | Lower Order Moms $m_{kj}$ | Central Moms $\mu_{kj}$ | Normalized Central Moms $v_{kj}$ |
|---|---|---|---|
| Zeroth ( 0 , 0 ) | 10 | 10 | 1 |
| First ( 1 , 0 ) | 40 | 0 | 0 |
| First ( 0 , 1 ) | 40 | 0 | 0 |
| Second ( 2 , 0 ) | 166 | 6 | 0.36 |
| Second ( 0 , 2 ) | 174 | 14 | 0.140 |
| Product ( 1 , 1 ) | 162 | 2 | 0.02 |
| Product ( 2 , 2 ) | 2968 | 8 | 0.008 |
| Centroid ( $x_c$ , $y_c$ ) | ( 4 , 4 ) | | |
| Principal angle $\phi$ | 76. 72° | | |

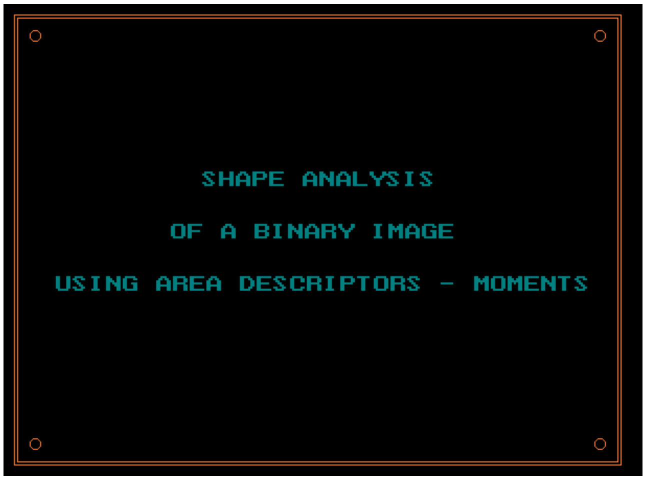
Table 1

## V. Simulation results



```
                 Shape Analysis using Moments

     Enter Any Size for Image :

     No. of Rows = 8
     No. of Cols = 8

     Enter the Bitmap Image in Binary [0/1] : NonZero assumed 1.
              0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0

              0 1 1 0 1 0 0 0

              0 0 1 1 1 1 0 0

              0 0 1 1 1 0 0 0

              0 0 0 0 0 0 0 0

              0 0 0 0 0 0 0 0

              0 0 0 0 0 0 0 0
```

Fig. 5 : Simulation result 1

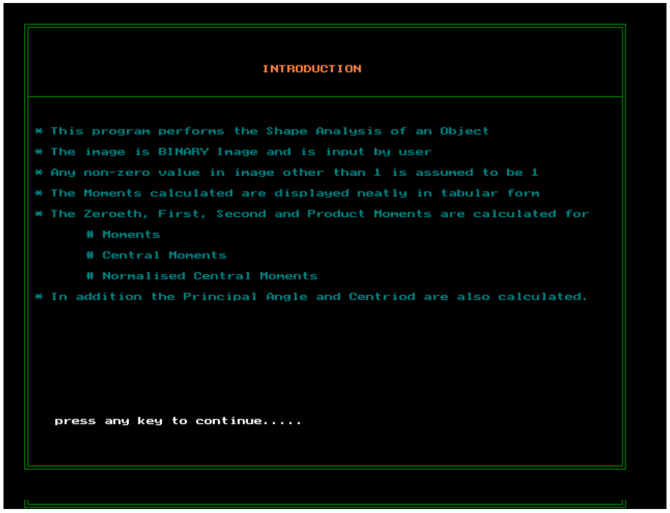


SHAPE ANALYSIS

OF A BINARY IMAGE

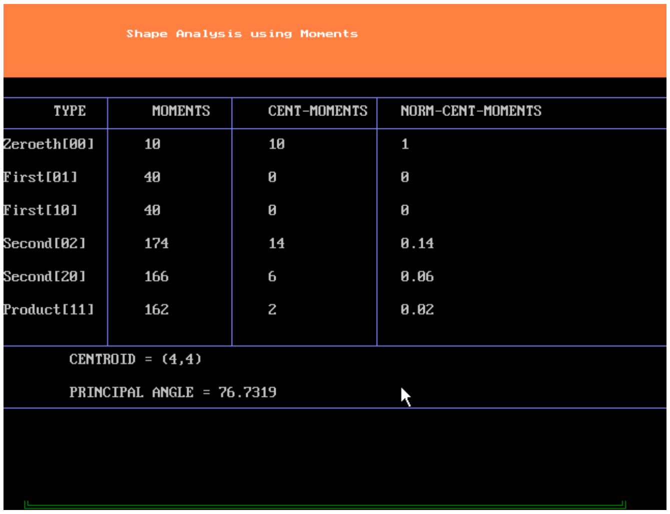
USING AREA DESCRIPTORS – MOMENTS

Fig. 6 : Simulation result 2



```
INTRODUCTION

* This program performs the Shape Analysis of an Object
* The image is BINARY Image and is input by user
* Any non-zero value in image other than 1 is assumed to be 1
* The Moments calculated are displayed neatly in tabular form
* The Zeroeth, First, Second and Product Moments are calculated for
      # Moments
      # Central Moments
      # Normalised Central Moments
* In addition the Principal Angle and Centriod are also calculated.


  press any key to continue.....
```

Fig. 7 : Simulation result 3



```
        Shape Analysis using Moments

   TYPE       MOMENTS     CENT-MOMENTS   NORM-CENT-MOMENTS
Zeroeth[00]   10          10             1
First[01]     40          0              0
First[10]     40          0              0
Second[02]    174         14             0.14
Second[20]    166         6              0.06
Product[11]   162         2              0.02

       CENTROID = (4,4)
       PRINCIPAL ANGLE = 76.7319
```

Fig. 8 : Simulation result 4

## VI. Conclusions

A method of computing the moments of a binary image was presented in this paper. The major disadvantage of moments in general is that they are global features rather than local. This makes them not suited for recognizing objects, which are partially obstructed. Moments are inherently location dependent, so some means must be adopted to insure location invariance (like the centroid). The mathematical results & the experimental results / simulated results shows the effectiveness of the developed method [2].

## References

[1] Craig J, *Introduction to Robotics* : *Mechanics, Dynamics & Control*, Addison Wessely, USA, 1986.

[2] Robert, J. Schilling, *Fundamentals of Robotics - Analysis and Control*, PHI, New Delhi.

[3] Klafter, Thomas and Negin, *Robotic Engineering*, PHI, New Delhi.

[4] Fu, Gonzalez and Lee, *Robotics: Control*, *Sensing*, *Vision and Intelligence*, McGraw Hill.

[5] Groover, Weiss, Nagel and Odrey, *Industrial Robotics*, McGraw Hill.

[6] Ranky, P. G., C. Y. Ho, *Robot Modeling, Control & Applications*, IFS Publishers, Springer, UK.

[7] Crane, Joseph Duffy, *Kinematic Analysis of Robotic Manipulators*, Cambridge Press, UK.

[8] Manjunath, T.C., (2005), *Fundamentals of Robotics*, Fourth edn., Nandu Publishers, Mumbai.

[9] Manjunath, T.C., (2005), *Fast Track to Robotics*, Second edn., Nandu Publishers, Mumbai.

[10] Dhananjay K Teckedath, *Image Processing*, Third edn., Nandu Publishers, Mumbai.

[11] Gonzalvez and Woods, *Digital Image Processing*, Addison Wesseley Publishers.

[12] Anil K Jain, *Digital Image Processing*, Prentice Hall, Englewood Cliffs, New Jersey, USA.

[13] http://www.wikipedia.org

[14] Michael Dipperstein**,** *Run Length Encoding* (*RLE*) *Discussion and Implementation*.

[15] Flusser, J.; Suk, T.; Saic, S., *Recognition of blurred images by the method of moments*, Image Processing, IEEE Transactions.

[16] Bob Bailey, *Moments in Image Processing*, Nov. 2002.

# Enhanced Method for Extracting Features of Respiratory Signals and Detection of Obstructive Sleep Apnea Using Threshold Based Automatic Classification Algorithm

A.Bhavani Sankar[1], D.Kumar[2], K.Seethalakshmi[3]
[1]Assistant Professor, Dept. of ECE, Anjalai Ammal - Mahalingam Engineering College, Kovilvenni
[2]Dean, Research,Periyar Maniyammai University,Vallam
[3]Senior Lecturer, Dept. of ECE, Anjalai Ammal - Mahalingam Engineering College, Kovilvenni.
E-Mail: absankar72@gmail.com, kumar_durai@yahoo.com, seetha.au@gmail.com

***Abstract*****—** Obstructive Sleep Apnea is a frequent disorder with detrimental health, performance and safety effects. The diagnosis of the disorder is cumbersome and expensive. New methods for screening and diagnosis are needed. The method we describe in this work is based on detection of four main features of respiratory signal. The automatic signal classification starts by extracting signal features from a 1 minute data segment through autoregressive modeling (AR) and other techniques. Four features are: signal energy, zero crossing frequency, dominant frequency estimated by AR and strength of dominant frequency based on AR. These features are then compared to threshold values and introduced to a series of conditions to determine the signal category for each specific epoch. The threshold values for the parameters were determined through experiment.

***Keywords***- Sleep Apnea, Motion Artifact, Energy Index, Respiration rate, Dominant frequency, Strength of Dominant frequency, Zero Crossing.

## 1. INTRODUCTION

Respiration monitors are of crucial importance in providing timely information regarding pulmonary function in adults and the incidence of Sudden Infant Death Syndrome (SIDS) in neonates. However, to accurately monitor respiration, the noise inherent in measuring devices, as well as artifacts introduced by body movements must be removed or discounted. With the recent success of media in creating awareness about the importance of sleep and effects of sleep apnea the classifying algorithm should be easy to use and provide a fair prediction that must contribute to public health. One can imagine a multitude of intelligent classification algorithms that could help to reach better identification mechanism. For example an algorithm should be capable of classifying different types of signal with different characteristics feature. Such an algorithm has the potential to become major classification tool. There have been enormous growth in developing efficient algorithm for classification of the respiratory signals, the reduced computational steps, reduced number of parameters used, increasing the capability to differentiate the signals and easy to implement in hardware setup to provide clinical support. An efficient algorithm should adopt itself to any kind of signals; it should not have any static rules for classifying the given input signal.

This work shows a simple method for respiratory signal classification using a MATLAB coding. It describes an automatic classification algorithm using features derived from the autoregressive modeling and threshold crossing schemes that was used to classify respiratory signals into the following categories: (1) normal respiration, (2) respiration with artifacts and (3) sleep apnea. This classification is capable of detecting fatigue of the human by identifying sleep apnea, early detection of sleep troubles and disorders in groups at risk, reduces the risks of being affected by serious heart diseases in future. The main contribution of this paper is the analysis of signals those are necessary for classification of the respiratory signals which yields not only the classification but also the analysis of various ailments.

Results in [1] indicate that respiratory signals alone are sufficient and perform even better than the combined respiratory and ECG signals. Respiratory signals are convenient to measure because they do not require electrodes on the skin, and people may wear the sensors for periods of several days and weeks. An apnea detection method based on spectral analysis was discussed in detail in [2]. In [3] the possibility of recognizing obstructive sleep apnea based on beat-by-beat features in ECG recordings was studied. It was also explored the application of time- varying autoregressive models and KNN linear classifier. A classification scheme of respiratory signal based on fuzzy logic was proposed in [4]. The paper [5] proposes an implementation of automatic classification of respiratory signals using a Field Programmable Gate Array (FPGA). The main novelty in [6] is that the phase difference between the two respiration signals is considered in order to determine the presence and grade of obstructive apnea. The work in [7] shows that the interval between zero crossings gives a good estimation of its frequency with reduced computational effort. The utilization of a second order autoregressive (AR) model to extract the dominant frequency and quantify its strength was discussed in [8].

## 2. SLEEP APNEA

Sleep apnea is a common sleeping disorder. When a person has sleep apnea, he or she stops breathing for short periods of time [3]. In most cases this lasts from 10 seconds to 1 minute or more while asleep. Then the person begins breathing again. A person may stop breathing only a few times or hundreds of times in the course of the night. If apnea is kept untreated it will lead to increase the risk for High blood pressure, Heart attack, Obesity and Diabetes, Increase the risk for worsen Heart failure, Make irregular heartbeats more likely, and Increase the chance of having work-related or driving accidents. Sleep apnea can be treated by focusing on reducing airway blockage and increasing the amount of oxygen in the body. The first step is often a serious attempt at losing weight. It is also crucial to avoid alcohol and sleeping pills.

If these measures do not help, the person may need a continuous positive airway pressure (CPAP). The individual wears a mask over the nostrils or mouth that pumps in pressurized air. This increases the amount of oxygen entering the lungs. It also relieves the symptoms of obstruction. The technique can be used with or without supplemental oxygen. Dental appliances may be used to reposition the tongue and lower jaw. Uvulopalatopharyngoplasty is a type of surgery that removes excess tissue at the back of the throat. If all other methods fail, a tracheostromy may be done. This involves cutting a small hole in the neck through which the person can breathe. Medicines may be needed to increase respiratory function while the person sleeps. Antidepressants may be prescribed. These reduce the amount of time a person spends in deep sleep.

### *Classification of Sleep Apnea*

There are three classifications of sleep apnea, including:

- OBSTRUCTIVE SLEEP APNEA, which means something, is blocking the airway or the airway does not open all the way during sleep.
- CENTRAL APNEA, in which the brain isn't signalling the muscles to breathe or the muscles don't receive or can't respond to the signal to breathe.
- MIXED APNEA, this is a combination of obstructive and central apnea.

The most common kind of sleep apnea is called Obstructive Sleep Apnea Syndrome [2]. Sleep apnea means "cessation of breath." It is characterized by repetitive episodes of upper airway obstruction that occur during sleep, usually associated with a reduction in blood oxygen saturation. In other words, the airway becomes obstructed at several possible sites. The upper airway can be obstructed by excess tissue in the airway, large tonsils, and a large tongue and usually includes the airway muscles relaxing and collapsing when asleep. Another site of obstruction can be the nasal passages. Sometimes the structure of the jaw and airway can be a factor in sleep apnea. A sleep test, called polysomnography is usually done to diagnose sleep apnea. There are two kinds of polysomnograms. An overnight polysomnography test involves monitoring brain waves, muscle tension, eye movement, respiration, oxygen level in the blood and audio monitoring. The second kind of polysomnography test is a home monitoring test. A Sleep Technologist hooks you up to all the electrodes and instructs you on how to record your sleep with a computerized polysomnograph that you take home and return in the morning. They are painless tests that are usually covered by insurance.

The positive effects of sleep deprivation on depressed people are used in psychiatry to treat a multitude of depression types without medication and are the most rapid antidepressant available today a lifestyle device. The developed algorithm could also be a primary sleep disorder prevention system that would be more powerful than only passive prevention methods and less expensive determination method.

## 3. NEED OF RESPIRATORY SIGNAL

The traditional methods for assessment of sleep related breathing disorders are sleep studies with the recordings of ECG, EEG, EMG and respiratory effort. Sleep apnea detection with ECG recordings requires more number of electrodes on the skin and people may wear it continuously for effective monitoring. EEG measurement can also be used for the detection of sleep apnea but the brain signals are always random in nature. For the complete detection, we need more number of samples for analysis. Also, the mathematical modeling of EMG signals is very complex for sleep apnea detection. From the results in [1], the respiratory signals alone are sufficient and perform even better than ECG, EEG and EMG. In our paper, we consider only the respiratory signal for the detection of sleep apnea since it is more convenient and do not require more number of electrodes on the skin.

The human respiratory signal as shown in Fig.1 is classified into three major classifications namely,

- Normal respirations.
- Motion artifacts.
- Sleep apnea.

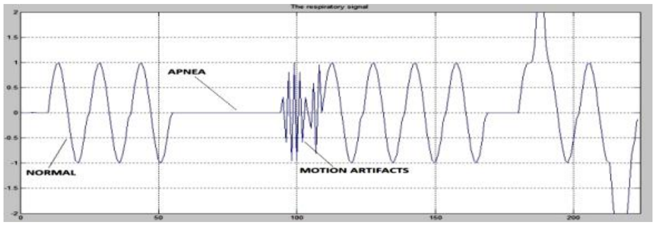


**Fig.1 Human Respiratory Signal**

### *Normal Respiration*

The normal respiration is characterized by the presence of a certain rhythm and the presence of some energy level in the signal.

### *Sleep apnea*

Apnea is easily classified as the absence of energy (ventilation activity) as well as a lack of rhythm. The respiration rate was below a critical level.

### *Motion artifact*

Motion Artifact is generally characterized by a sudden increase in the amplitude of the signal and by a sudden variation in the rhythm of the heart usually has the higher energy when compared to the normal respiration. Motion artifacts are transient baseline changes caused by changes in the skin impedance. This type of interference represents an abrupt shift in base line due to movement of the patient while the respiratory signal is being recorded

## FEATURE EXTRACTION

This classification algorithm extracts several features of respiratory signals and utilized for disease identification. The feature extraction plays a vital role since the classification is completely based on the values of the extracted features. The fundamental features of respiratory signal provide the numerical value which is compared with the threshold values and the classification results will be produced. The fundamental features of respiratory signals [8] are

- Energy Index (EI)
- Respiration frequency estimated by a modified Zero crossing scheme (FZX)
- Dominant frequency estimated by AR modeling (FAR)
- Strength of the dominant frequency estimated by AR modeling (STR)

### *Energy Index (EI)*

Given a continuous-time signal f (t), the energy contained over a finite time interval is defined as follows.

$$E(T_1, T_2) = \int_{T_1}^{T_2} |f(t)|^2 . dt, T_2 > T_1 \qquad (1)$$

$$E_f = \int_{-\infty}^{+\infty} |f(t)|^2 . dt \qquad (2)$$

Equation (1) defines the energy contained in the signal over time interval from $T_1$ till $T_2$. On the other hand, equation (2) defines the total energy contained in the signal. If the total energy of a signal is a finite non-zero value, then that signal is classified as an energy signal. **Typically the signals which are not periodic turn out to be energy signals.** The equation for computing Energy index is,

$$EI = \frac{1}{\mathrm{N}} \sum_{\mathrm{n=0}}^{\mathrm{N-1}} |x_n|^2 \qquad (3)$$

The average power of the signal x is defined as the energy per sample

$$p_x \triangleq \frac{\varepsilon_x}{N} = \frac{1}{\mathrm{N}} \sum_{\mathrm{n=0}}^{\mathrm{N-1}} |x_n|^2 \qquad (4)$$

### *Respiration Frequency (FZX)*

Zero-crossing is a commonly used term in electronics, mathematics, and image processing. In mathematical terms, a "zero-crossing" is a point where the sign of a function changes (e.g. from positive to negative), represented by a crossing of the axis (zero value) in the graph of the function as shown in Fig.2.

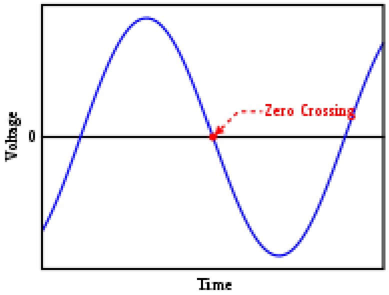


**Fig.2 Representation of Zero Crossing**

Counting zero-crossings is a method used in speech processing to estimate the fundamental frequency of speech. The interval between zero crossings gives a good estimation of its frequency [7]. Similarly, Respiration frequency (FZX) was determined by counting the number of times that x(n) crosses a baseline which is defined as the square root of EI.

$$FZX = \sqrt{EI} \qquad (5)$$

### *Dominant Frequency (FAR)*

In order to obtain the features FAR and STR, coefficients of a second order AR model have to be estimated. The respiration signal can be modeled as a second order autoregressive model [4] as the following,

$$x(n)=a_1x(n-1)+a_2x(n-2) + e(n) \qquad (6)$$

Where e(n) is the prediction error and {a1,a2} are AR model coefficients. Autoregressive **(AR)** spectral estimation techniques are known to provide better resolution than classical periodogram methods when short segments of data are selected for analysis. In our study, we adopted the Burg's method to compute AR coefficients. The major advantage of Burg method for estimating the parameters of the AR model are high frequency resolution, stable AR model and it is computationally efficient.

Using the second order autoregressive model coefficients, one can determine the dominant frequency and signal regularity strength as the following,

$$\mathrm{FAR} = \frac{F_s}{2\pi}\arctan\frac{a_1}{a_2} \qquad (7)$$

Where $f_s$ is the sampling frequency. A sampling frequency of 20Hz was used for analysis.

### *Strength of Dominant Frequency (STR)*

The AR coefficients were used to determine STR value as,

$$STR = \sqrt{a_1^2 + a_2^2} \qquad (8)$$

Basically, FAR and STR serve the same purpose as power spectrum usually does, indicating the dominant frequency and its corresponding power level. The classification of the signal is based on derived parameters shown above and thresholds would be properly initialized to allow accurate classification. The degree of reliability of the respiration rate estimate was determined by STR, which have a value between 0 and 1. For very regular rhythm, STR is very close to 1 (as in the case of normal respiration). If STR is too low, then the rate estimates FAR and FZX are deemed to be unreliable.

To make this algorithm more robust, the threshold values of classification parameters can be obtained as follows. Take the respiratory samples for one minute (two epochs) and the system performs the same analysis on these two epochs and extracts nominal values for each of the classification parameters. Then these nominal values are used to adjust the threshold values as follows [1]: low and high energy are 33% and 150% of average energy, low and high frequency are 50% and 150% of nominal frequency and low and high strength of dominant frequency are 75% and 95% of average strength. These values were determined experimentally.

The normal breathing frequency for a human being is usually between 0.2-0.3Hz and maximum frequency is unlikely to exceed 0.7- 0.8 Hz. Hence these values are used as the minimum and the maximum threshold for the respiration rate. Using the square root of the energy index as the appropriate baseline value for zero crossing, the number of times the signal crosses the baseline value was recorded and the respiration frequency was detected from it. A moving baseline was used to allow for changes in the mean respiration level. The calculated energy index, the respiration rate, the dominant frequency and the strength of the signal were compared with the set threshold values and were classified as normal respiration, apnea or respiration with artifact.

## 4. Simulation Results And Discussion

The input of 1200 samples of respiratory signal is as shown in the Fig.3.

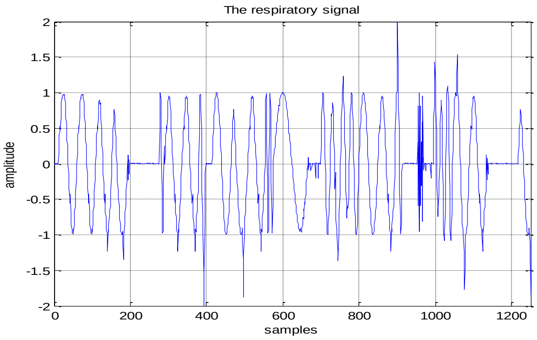


**Fig.3 Respiratory Input taken for Classification**

The input to our classification algorithm is human respiratory signal of 1 Hz which is given as the collection of data points which is obtained from the website www.physionet.com.The website hosted by a medical institute consist of various human bio-signals like ECG, EMG, EOG, RESPIRATORY SIGNALS., in the form of data points from which we can retrieve the required respiratory signal and can be given as the input to the classification algorithm developed.

The classification of the signal is performed by the sequential analysis of the signal in various phases of the program using MATLAB as,

- Form the mathematical model of the input respiratory signal using second order auto regressive modeling.
- Determine the various parameters of the derived mathematical model with the help of burgs algorithm.
- Extract the four fundamental features of human respiratory signals with the help of the provided mathematical equations.
- Compare the derived values with the optimum threshold values of the four fundamental features of the respiratory signal.

- Produce the classified results of the respiratory signals.

The signal was sampled at 20Hz and the total of 1200 data points is taken for analysis. The threshold values for the features to be extracted from the signal are provided as the optimum values in the program which is used for signal classification.

TABLE I. NOMINAL VALUES CALCULATED USING TWO EPOCHS

| Sample | E_Low | E_High | Str Low | Str High | FAR Min | FAR Max |
|---|---|---|---|---|---|---|
| Normal_1 | 0.43 | 1.97 | 0.20 | 0.26 | 0.57 | 1.72 |
| Normal_2 | 0.52 | 2.36 | 0.25 | 0.31 | 0.63 | 1.88 |
| Apnea_1 | 0.27 | 1.22 | 0.14 | 0.18 | 0.45 | 1.36 |
| Apnea_2 | 0.22 | 0.98 | 0.08 | 0.10 | 0.40 | 1.21 |
| Artifacts_1 | 0.74 | 3.36 | 0.45 | 0.57 | 0.75 | 2.25 |
| Artifacts_2 | 0.98 | 4.45 | 0.39 | 0.49 | 0.86 | 2.58 |

The nominal values calculated using two epochs to adjust the threshold values are given in Table I and the optimum values calculated as threshold is given in Table II.

TABLE II. CALCULATED OPTIMUM THRESHOLD VALUES

| Features | Normal | Apnea | Artifacts |
|---|---|---|---|
| E_Low | 0.5 | 0.25 | 0.8 |
| E_High | 2.2 | 1 | 4 |
| Str_Low | 0.2 | 0.05 | 0.4 |
| Str_High | 0.3 | 0.15 | 0.5 |
| FAR_Min | 0.6 | 0.4 | 0.8 |
| FAR_Max | 1.8 | 1.3 | 2.4 |

Respiration data was first divided into 20 second epochs and manually scored for comparison. The epochs were then processed with the automatic classification algorithm and compared to manual classification.

TABLE III. RESULT OF CLASSIFICATION ALGORITHM

| Episode | Manual | Simulation |
|---|---|---|
| Normal | 745 | 745 |
| Artifact | 148 | 148 |
| Sleep apnea | 292 | 292 |
| Unclassified | 15 | 15 |

The classified results provided in Table III shows that the proposed algorithm is capable of classifying with the accuracy of 100% in case of normal and sleep apnoic signals. Few disagreements were encountered with the detection of motion artifacts and hence 15 sections are termed as unclassified signals. The results obtained indicate that this algorithm can be an effective approach in respiration devices being developed to monitor infants at risk for SIDS and to accurately compute respiration rate on a regular base for selected patients.

## 5. CONCLUSION

This classifying algorithm with the help of MATLAB coding classifies the human respiratory signals into three major classifications such as normal respiration, motion artifacts and sleep apnea. The classification system is given with the human respiratory signal as the input, and the coding is developed in such way that it models the given signal as a mathematical equation using second order Auto Regressive model, then the parameters of the developed equation is determined with the help of Burgs algorithm. Then the fundamental features of the respiratory signal such as Energy index, Respiration frequency, Dominant frequency, Strength of the dominant frequency were calculated. The determined values then compared with the optimum predetermined values of the fundamental features and the results were developed with the help of the comparison. Hence the provided respiratory signal is classified successfully with the help of the formulated algorithm.

This work can be developed and implemented in real time application for detecting sleep apnea. To develop this project in real time, we have to design a processor and the algorithm should be improved by adding calibration procedures and is adjusted to run on FPGA [5]. The electrical signals which are analog in nature should be converted into digital by analog to digital converter (ADC) and is given to the FPGA kit. Then the processor will process and detect the appropriate signal. LCD or PC monitor can be used to display the name of the signal.

## References:

[1]. Walter Karlen, Claudio Mattiussi, and Dario Floreano, “Sleep and Wake Classification With ECG and Respiratory Effort Signals”, IEEE Transactions on Biomedical Circuits and Systems, Vol. 3, No. 2, April 2009.

[2]. Lorena S. Correa, Eric Laciar, Vicente Mut, Abel Torres,and Raimon Jané, “Sleep Apnea Detection based on Spectral Analysis of Three ECG - Derived Respiratory Signals” 31st Annual International Conference of the IEEE, EMBS, Minneapolis, Minnesota, USA, September 2009.

[3]. Martin O. Mendez, Davide D. Ruini, Omar P. Villantieri, Matteo Matteucci ,Thomas Penzel, Sergio Cerutti ,Anna M. Bianchi, “Detection of Sleep Apnea from surface ECG based on features extracted by an Autoregressive Model” Proceedings of the 29th Annual International Conference of the IEEE , EMBS, August 23-26, 2007.

[4]. Maria I. Restrepo, Susmita Bhandari, and Taikang Ning, "Classification of Respiration Episodes using Fuzzy Logic" IEEE 2006.

[5]. Bozidar Marinkovic, Matthew Gillette, and Taikang Ning, "FPGA Implementation of Respiration Signal Classification Using a Soft-Core Processor" IEEE 2005.

[6]. Peter Varady, Szabolcs Bongar, and Zoltan Benyo, "Detection of Airway Obstructions and Sleep Apnea by Analyzing the Phase Relation of Respiration Movement Signals", IEEE Transactions on Instrumentation and Measurement, Vol. 52, No. 1, February 2003.

[7]. Vladimir Friedman, "A Zero Crossing Algorithm for the Estimation of the Frequency of a Single Sinusoid in White Noise" IEEE Transactions on Signal Processing, Vol. 42, No. 6, June 1994.

[8]. Taikang Ning and Joseph D. Bronzino, "Automatic Classification of Respiratory Signals" IEEE Engineering in Medicine and Biology Society, 11th Annual International Conference, 1989.

## Author Biographies

**BHAVANI SANKAR.A.** He received his B.E in Electronics and communication Engineering from Institute of Road and Technology, Erode, Bharthiyar university, Coimbatore, Tamilnadu, India in 1994.He received his M.E in Power Electronics and drives fro Shunmugha Engineering College,Thanjavur,Tamilnadu, India in 2000. Currently, he is a pursuing his Ph.D., in the area of Bio-Signal Processing from Department of Electrical and Electronics Engineering from Anna university of Technology,Trichy, Tamilnadu,India. His field of research includes Bio-Medical, Signal Processing. Soft Computing. Currently, he is working as an Assistant Professor and Head, department of E.C.E in Anjalai Ammal Mahalingam Engineering College, Kovilvenni, Thiruvarur (Dt), Tamilnadu, India

**Dr.KUMAR.D.** He received his M.Tech and Ph.D., in Bio-Medical Instrumentation from IIT Madras, Chennai,Tamil nadu ,India.He is currently working as a Dean ,Research, Periyar Maniyammai University, Thanjavur, Tamil nadu, India. He has published many papers in both International and National journal and conferences. His research interests include Medical Electronics, Medical Imaging, Optical Imaging, Bio-Sensors.

SEETHA LAKSHMI.K. She received his B.E in Electronics and communication Engineering from A.C.Tech, Karaikudi, and Tamilnadu, India in 2002.She received his M.E in Communication Systems from Anna University, Chennai, Tamilnadu, India in 2010. Currently, she is working as a senior lecturer in E.C.E department in Anjalai Ammal Mahalingam Engineering College, Kovilvenni, Thiruvarur(Dt), Tamilnadu, India. Her field of interest in Optical Imaging, Soft computing, Medical informatics.

# Artificial Neural Network Modeling of Land Price at Sowcarpet in Chennai City

Sampathkumar.V[1], Helen Santhi. M[2]

[1] Research Scholar, Sathyabama University,Chennai, India
[2] Principal, Indira Gandhi College of Engg & Tech for Women, Chengalpet, India
Email: svsampathkumar@gmail.com : mhelensanthi@gmail.com

***Abstract:*** In this paper Artificial Neural Network (ANN) models are employed to forecast the unit land price of Sowcarpet, which is located in the core city as well as a part of Central Business District (CBD) of Chennai. The monthly average value of the selected factors such as National Gross Domestic Product, cost of crude oil, dollar equivalence to Indian currency, rate of inflation, gold and silver price, Mumbai and National share index, population in the study area, interest rate on home loan, unit cost of construction, guideline value and time factor from the year 1997 to 2008 are considered in the study and the models are validated with the land price of 2009 and 2010. The models are used to forecast the land price for the next five years and found that there will be a uniform annual increase of 17 % in the selected location.

***Keywords:*** Land Price, economic factors, neural network, land price model, future trend.

## 1. Introduction

Over the last two decades there have been a lot of research studies analysing land prices. Each study includes attributes of land price such as geographical location, the environment, size of plot, land use pattern, soil productivity, topography, drainage, population growth, economic development, infrastructure, agriculture, nearby developments, etc. [1]- [3]. Statistical models have been commonly used to estimate land prices [4] – [7].Recently, artificial neural network models have been applied in real estate prediction [8]-[11]. The studies show reasonable accuracy for complex problems using ANN models.

From the literature review, the following significant factors that influence the land price trend are selected for the study: Gross Domestic Product (%), Cost of crude oil ($),Dollar equivalence to Indian currency (Rs), Rate of inflation (%), Gold and Silver price per gram (Rs),Mumbai and National share index,Population in the study area, Interest rate on home loan (%),Unit cost of construction per Square foot, Guideline value per ground (Rs) and Time factor (Year and Month). This study focuses on artificial neural network to evaluate the influence of economic factors on land price and its future trend.

Land is an immovable, scarce resource, which helps to fulfill the basic need of a human. Owning a land and house is a prestigious issue in the society and because of potential it becomes an investment option now. Effective usage of land becomes an integral part of urban development. The urban-based economic activities account for more than 50 % of Gross Domestic Product (GDP) in all the countries. In India by 2011, urban area will contribute 65 % to GDP. The agricultural land was 84 % in the year 1980 and it will shrink to 35 % in 2020 due to rapid urbanization. Major urbanization pressure is to be addressed due to mass migration of people from small towns to urban centers. Presently 41 cities in India have more than one million population but before two decades, it was only 33 cities. In 2050, India needs to accommodate 900 million more people in cities, which requires 18500 square kilometers of land as per the conventional planning norms. According to World Development Indicator report, India’s urban population will increase to 75 % in 2050 from 38 % in 2009. Due to limited availability of land and to utilize it optimally, the development trend slowly shifted in vertical direction rather than radial and horizontal. This compact vertical development will make positive environmental impact and leads better accessibility and efficient transport.

Economic base of Chennai city has shifted from trade and commerce to administration and services. Buoyant Economy, increased employment rate, high disposable income, cosmopolitan atmosphere and improved life style are instrumental in driving the demand for high-rise apartments. The demand on residential property is consistent, price are also climbing up due to hike in the input cost. There is huge demand on developed plots and the supply level is virtually shrunk leads to hike in land price. In a market dominated by the end users, demand to supply mismatch continues. Invest on land for a secure future becomes a reality in India. Investment made on land yields better returns than apartments and other traditional investment options, at the range of 100 to 300 % over past few years in South and West suburbs of Chennai. Generally, the land price depends on economical, social and physical features. Compared to previous years the market has stabilized. Projects along the Old Mahabaliburam Road (OMR) have seen most new sales. The market is seeing a positive momentum with job security which increased number of end users. Most housing finance companies have kept their home loan rates stable without much increase helps to stabilize the market. Reasonable interest rate, increased supply and affordable prices are some reasons to predict the market will continue to do well. The revival of Information Technology (IT) sector leading to creation of new jobs and increased liquidity is also expected to give the land and housing market in Chennai a positive momentum. The study on land price trend becomes important to have a better idea on future land price which helps in planning issues.

## 3. About Chennai Metropolitan Area (CMA) and the Study Area

### 3.1 CMA

Vision 2026 is to make Chennai as a prime Metropolis, which will be livable, economically vibrant and environmentally sustainable. Chennai is the 4th largest metropolitan city in India. The City is at the core of CMA and is the centre for all commercial and social activities as well as a living area for majority of population. It is the place of focus on economic and cultural development. Chennai is situated on the Coromandel Coast in South India and the land is plane, which is located with latitude between $12^{0}50^{'}49^{''}$ and $13^{0}17^{'}24^{''}$ and longitude between $79^{0}59^{'}53^{''}$ and $80^{0}20^{'}12^{''}$. CMA comprises Chennai City Corporation, 16 municipalities, 20 town panchayats, 204 villages forming part of 10 panchayat unions in Thiruvallur and Kancheepuram districts. It extends over 1189 square kilometer area.

Chennai city and CMA have 55 and 70 lakhs of population respectively in 2009. It has a firm base of large industries and commerce including insurance, shipping and banking. The city has dramatically changed over a period and mushrooming of commercial building is an out-come of the changing spatial dynamics. From 1994 to 2007, Chennai city's GDP grew at 6.5 % while the states GDP grew at 6 %. The contribution of CMA to state GDP is 40 %. Chennai accounts for 30 % of national auto industry, 15 % of software exports and 50 % of leather exports. Land price scouring in city area and the development along IT corridor in South and electronic hardware corridor on West has given a virtual boost to land owners to increase the price. The exorbitant land value, which in turn upped the apartment prices to a new high in the city areas and in suburban properties, has a thrust on infrastructure development.

In real estate slump, the suburbs first bear the burnt, followed by City area and finally the CBD. Similarly, during an upturn, property revival first happens in muffussal area and the CBDs are the last to improve. Second master plan released by Chennai Metropolitan Development Authority (CMDA) [12] envisages a series of pragmatic measure for optimum utilization of available land. Higher FSI at 2.0 and redefining special buildings will move the belated developments. A survey by ICICI property services reveals that 30.70 million sq.ft. of residential space involving more than 21000 units by category A, B and C developers will enter in to the market by the end of 2009. Tidel park, existing industrial estates in Guindy and Ambattur, upcoming Sipcots on the fringes and IT units in West and Southern regions, proposed international airport, rapid transport and metro trains, over bridges, elevated and circular ring roads are the additional power of Chennai realty sector and the land price rise.

### 3.2 Study Area

Sowcarpet is one of the important parts of George Town (GT), which is the CBD of Chennai city. This area is abounding with jewelry shops, electrical shops, hardware shops, transport booking offices, vegetable and fruit markets, etc. The mixed land use pattern, narrow and winding unplanned road network which carries huge volume of mixed traffic more than its capacity have resulted a gradual degradation in the environmental quality of the area. It is filled with resettled population from North India, doing commercial, retail and wholesale business. Sowcarpet has 0.62 square kilometer area with latitude of $13^{0}05'29.15''$N and longitude $80^{0}16'52.39''$E and an elevation of thirteen meter above mean sea level and its location in Chennai city is shown in Figure 1.

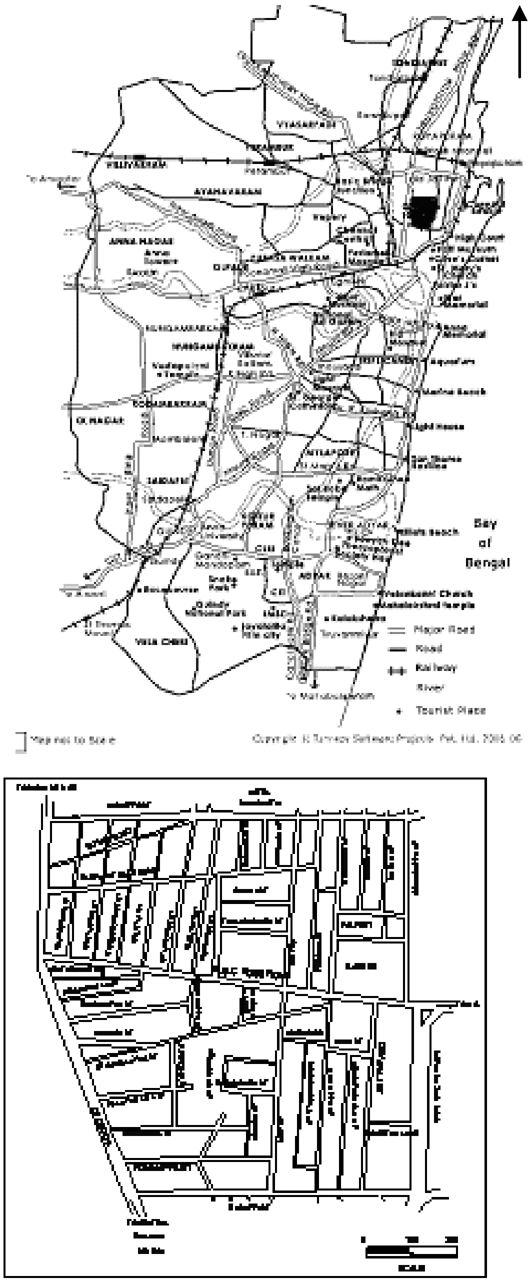


**Figure1.Location of Sowcarpet**

The road network of Chennai is dominated by a radial pattern converging at GT. Arterial roads leading to CBD carrying heavy traffic due to concentration of commercial, industrial and huge volume of employment related activities in the CBD are highly congested. Capacity of all the roads in CBD is reduced due to poor quality of riding surface and inadequate footpath facilities and unplanned intersections. Sowcarpet has 52 roads with total length of 14.6 kiolmeter and road to area ratio of just 0.14 which very much lesser than the urban planning guidelines. Even though the scope of widening of existing roads is negligible, many roads are converted into commercial zones. GT and Harbour becomes the commercial centre of the city. GT and its extension in South into Annasalai had the wholesale trade, specialized retail trade and banking and financial institutions and commercial activity was intense. More than two lakhs jobs constituting 48 % of the total work places in the CMA in 1971 were located here. General Hospital and Government Stanly Medical Hospital serves the CMA, is located next to sowcarpet. The growth rate of population in Chennai city and GT was in descending trend in the past two decade from 28 to 16 % and from 12 to 7.4 % respectively. The growth of area and population is shown in Table 1.

**Table 1. Population and its Growth in Chennai City (CC) and George Town (GT)**

| Year | Area (Sq. Km) | | Population (Lakhs) | |
|---|---|---|---|---|
| | CC | GT | CC | GT |
| 1961 | 128.83 | 5.34 | 17.5 | 2.67 |
| 1971 | 128.83 | 5.34 | 25.7 | 2.80 |
| 1981 | 170.0 | 5.34 | 32.75 | 3.13 |
| 1991 | 170.0 | 5.34 | 37.95 | 3.39 |
| 2001 | 172.0 | 5.34 | 44.0 | 3.64 |

Due to the influence of economic slowdown the annual number of land transactions registered in the registration office of Sowcarpet starts declined upto 80 % compared to the year 2007 after a sustained growth of past two decades. The results of questionnaire survey from the respondents of Sowcarpet yields the following information that it is located very close to Central railway station. Location of school, multi specialty hospital, vegetable markets and major bus terminus nearby High court are accessible from just 0.5 to 2 kilometer. Urban services like water supply, sewer drain, stome water drain and cleaning of garbage are satisfactory but the level of noise pollution is considerably high. Immigrated settlements from North India and from Andhra Pradesh, plenty of business and commercial activities like jewellery, cloth garments, hardware and electrical shops, are felt as the major factors of price rise of land and rental value of residence in Sowcarpet. Presently a two bedroom flat rents in five-digits and many house owners want to lease the house for huge sum of money.

The real estate growth in the city was dominated by the CBD and surrounding locations like Annasalai until 1990s. However, the scarce land availability and high real estate costs forced companies with large space requirements away from CBD. The current growth pattern of the city focused on the areas where government is planning roads and other infrastructure improvements. Decentralization of activities like vegetable market, flower bazaar and muffussal bus stand to peripheral areas like Koyambedu, Sathangadu and Madhavaram which are highly decongested the GT, will improve the living quality. Land price of CBD rise year by year. This is much above the affordability of the lower and middle in come group population. Triplicane, Mylapore, Purasiwalkam and the Northern part of GT like Royapuram are the old residential characterized by row housing with shopping along main roads. Shopping facilities of local significance had developed along almost all major roads. Experts say that for long term investors CBD area offer scope as even among commercial and residential sector one takes a beating, other sector will come to rescue so that there will be consistency in the flow of return on investment. The land price modeling study helps to ascertain the changes in growth occurred in the past and the reasons for the same. It will help to identify the potential for future growth and in planning aspects. To quantify the rise of land price over a period of time is essential for future policy implicating and to assess the compensation amount for the land which will be acquired for public purpose.

## 4. Theory of Artificial Neural Network

ANN is a computational technology from the artificial intelligence discipline whose architecture emulates the network of nerve cells in the human brain. An NN is a parallel distributed information-processing structure consisting of processing elements (PEs) which contains local memory. The PEs can also carry out localized information processing operations interconnected via unidirectional signal channels called connections. NN architecture such as a standard Back-propagation (BP) NN can be developed by using the various indicators as PEs to be investigated upon. The structure of back propagation is shown in Figure 2.

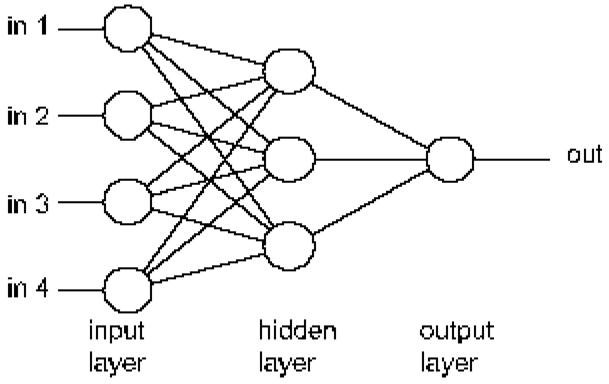


**Figure 2. Back propagation Network**

As in biological systems, the strength of these connections changes in response to the strength of each input and the use of transfer function by the PEs. All nodes (which are indicators) in the input-layer are fully connected to each of the hidden nodes in the hidden-layer and the process of learning involves all the input nodes and the hidden nodes. In other words, learning also involves all the other input nodes with each input node connected to every hidden node. The output value from each node of the hidden layer in turn becomes the excitatory input-value for a particular node in the output layer. In this study, there are 13 indicators, that is PEs, and one bias node in the input layer of the NN model is constructed. All the input values are normalized using the MinMax Table.

The principle behind this normalization process is:

Normalized value, N= [Original value - Minimum value] / [Maximum value - Minimum value]
Where, $0 \leq N \leq 1$

The module learns the underlying latent function through an error gradient-descent method and the training stops when the root-mean-Square-error for Output-target values falls below 0.0001 percent. It takes 88 Epochs to reach the desired target. More iteration in the training of data improves convergence. Each hidden node (that is $H_1$ to $H_3$) receives a set of feed–in signals (or values) from which an output value is generated. Finally, all nodes in the hidden-layer are fully connected to the output node.
Share of Influence Input Node, $I_i$, asserts on the subject Output Node = $S_i$%

$$S_i = \frac{\Sigma^{nj}_{j=1} (|w_{ij}||o_j|) / (\Sigma^{ni}_{i=1}|w_{ij}|)}{\Sigma^{ni}_{i=1} \Sigma^{nj}_{j=1}(|w_{ij}||o_j|) / (\Sigma^{ni}_{i=1}|w_{ij}|)} \text{ X 100} \quad (1)$$

Where $\mathbf{n_i}$ = number of input nodes
$\mathbf{n_j}$ = number of hidden nodes
$\mathbf{w_{ij}}$ = connection – weight from input node $I_i$ to hidden–node $H_j$
$\mathbf{o_j}$ = connection – weight from hidden node $H_j$ to subject output node $S_i$

An ANN learns to solve specific problems without the need for problem - specific algorithms. The learning strategy incorporates the minimization of mean square error across all training patterns. The user can set a desirable result and compare the network's performance with the target training set.

In the development of feasible neural network solution all 13 indicators are used to ascertain the effect and to predict the trend of land price, thereby to preserve reliability in subsequent comparison on the accuracy of neural network solution. BP NN is chosen as a basic since, it is widely accepted. The aim of the learning process is to minimize the global error E of the system by modifying the weights. A gradient descent rule is adopted in the learning across the training set. Suppose a vector i is presented as the input layer of the network and the desired output is D. Let O denote the actual output produced by the network with its current set of weights. Then the measure of the error in achieving that desired output is given by:

$$\mathbf{E = 0.5 \sum_{k} (D_k - O_k)^2} \quad \mathbf{(2)}$$

ANN is set to 10000 iterations. Training stops when convergence obtains at the required root-mean-square-error or when the error across the learning maxim generated by network has become consistently stable. Forecasts are being made over a period between 2011– 2015 and the Run dialog box in the neural programme will help to establish the actual output. The share of influence of selected indicators is established using Garson's method. In this regard, the neural programme has the NN tool dialog box that shows the change in output by weightages. This helps to know which of the indicators has the most effect on the output.

## 5. Results and Discussion

The forecast ability of the neural network solution is shown in Table 2. The actual and ANN model prices are compared and it can be seen that the difference is about 3 to 4 %. This indicates that the identified indicators may be used as reliable inputs for modeling of land price and also the model is validated with the prices in the years 2009 and 2010.

**Table 2. Prediction ability of ANN model**

| Year | Actual Market Price (Lakhs) | ANN Model Price (Lakhs) |
|---|---|---|
| 1997 | 36 | 35.97 |
| 1998 | 39 | 39.30 |
| 1999 | 42 | 42.60 |
| 2000 | 45 | 45.95 |
| 2001 | 48 | 49.23 |
| 2002 | 60 | 62.43 |
| 2003 | 72 | 75.60 |
| 2004 | 84 | 88.80 |
| 2005 | 96 | 99.97 |
| 2006 | 108 | 115.20 |
| 2007 | 144 | 150.80 |
| 2008 | 156 | 158.00 |
| 2009 | 144 | 149.31 |
| 2010 | 150 | 155.90 |

Table 3 shows the share of influence of identified factors towards the output node. ANN solution has ranked Construction Cost, Dollar equivalence and Crude oil price as major influencing factors towards Land price.

**Table3. Explanatory strength of indicators in ANN model**

| Indicators | $\Sigma_{j=1} [(\|w_{ij}\|\|o_j\|) / (\Sigma_{i=1}\|w_{ij}\|)]$ | Share of Influence | Strength of Indicators |
|---|---|---|---|
| Construction cost | 2.933248 | 9.865645 | 1 |
| Dollar equivalence | 2.637307 | 8.870282 | 2 |
| Crude oil price | 2.587336 | 8.702212 | 3 |
| Mumbai sensex | 2.560612 | 8.612329 | 4 |
| Inflation | 2.53264 | 8.518246 | 5 |
| Home loan interest | 2.371333 | 7.97571 | 6 |
| Time | 2.347322 | 7.894949 | 7 |
| Silver price | 2.060556 | 6.930445 | 8 |
| Population | 2.02966 | 6.82653 | 9 |
| National sensex | 2.031778 | 6.833656 | 10 |
| Guideline value | 1.964782 | 6.60832 | 11 |
| Gold price | 1.849004 | 6.218914 | 12 |
| GDP | 1.826364 | 6.142767 | 13 |
| Sum of signal transfer | 29.73194 | 100 | - |

The validated land price model and the future model are shown in Figures 3 and 4, respectively. From the results, the predictions offered an average STDEV of 3.75 which is marginal and the annual rise of land price in Sowcarpet will be 17 %.

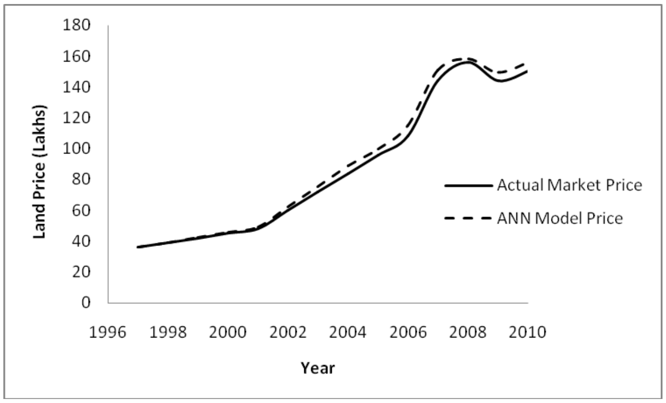


**Figure 3. Validated Land Price Model**

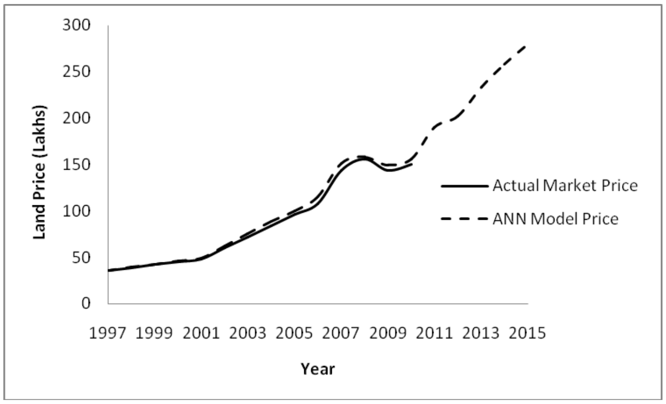


**Figure 4. Forecasted Land Price Model**

## 6. Conclusion

In this paper, the land price at Sowcarpet in Chennai City is evaluated by Artificial Neural Network. Artificial neural network model prices are tested for their predictive power using economic factors collected from the year 1997 to 2008. The forecasted results show that the annual rise in land price at Sowcarpet is about 17 % in the next 5 years.

## REFERENCES

[1]] BillMundy, John.A.Kilpatrik,“Factors Inflincing CBD land prices”, Real Estate Issues,Vol.25, No.3,pp 39,2000.

[2] Malphillamy.C.H, “Factors Affecting Rural Land Price in N.S.W. and the construction of indexes of Rural Land Values”, Australian Jl. of Agricultural Economics, Vol.7, No.4, pp 145-160, 1964.

[3] Goh, B.H., “Residential Construction Demand Forecasting Using Economic Indicators: A Comparative Study of Artificial Neural Networks and Multiple Regression”, Construction Management and Economics, Vol.4, No.1, pp 25-34,1996.

[4] Bruce,R., P.Sundell, “Multiple Regression Analysis: History and Applications in the Appraisal Profession”, Real Estate Appraisal and Analyst Vol.43, pp 37-44, 1997.

[5] Ramsland, M.O, D.E.Markham, “ Market- Supported Adjustments Using Multiple Regression Analysis”, The Appraisal Journal, 1998.

[6] Hannonen, “ Predicting urban land prices : a comparison of four approaches”, International Journal of Strategic Property Management, Vol.2, pp 20-25, 2008.

[7] Billie Ann Brotman, “Linear and Nonlinear Appraisal Models”, The Appraisal Journal, Vol.58, No.2,1990.

[8] A Quang Do, Grudnitski, Gary, “ A neural network approach to residential property appraisal”, Real Estate Appraiser, Vol.58, No.3, 1992.

[9] Rossini, P.A., "Artificial Neural Networks versus Multiple Regression in the Valuation of Residential Property", Australian Land Economics Review, Vol.3, No 1, 1997.

[10] Do, A. Q. and G. Grudnitski, “A Neural Network Approach to Residential Property Appraisal”, The Real Estate Appraiser, Vol.58, pp 38 – 45,1992.

[11] Samad, T, “Backpropgation is Significantly Faster if the Expected Value of the Source Unit is Used of Update”, International Neural Network Society Conference Abstracts, 1998**.**

[12] Chennai Metropolitan Development Authority, “Master plan for Chennai Metropolitan Area- 2011, Draft 2”, 2007.

# A Survey on Applications of Wireless Sensor Network Using Cloud Computing

Sanjit Kumar Dash[1], Subasish Mohapatra[2] and Prasant Kumar Pattnaik[3]

[1]Department of Information Technology, College of Engineering and Technology, India
sanjitkumar303@gmail.com
[2]Department of Computer Science and Engineering, Institute of Technical Education and Research, India
subasish.mohapatra@gmail.com
[3] Department of Computer Science and Engineering, Konark Institute of Science and Technology, India
patnaikprasant@gmail.com

***Abstract*****—**Popularity of cloud computing is increasing day by day in distributed computing environment. There is a growing trend of using cloud environments for storage and data processing needs. Cloud computing provides applications, platforms and infrastructure over the internet. It is a new era of referring to access shared computing resources. On the other hand, wireless sensor networks have been seen as one of the most essential technologies for the 21st century where distributed spatially connected sensor node automatically forms a network for data transmission and receive among themselves is popularly known as Sensor Network. For security and easy access of data, cloud computing is widely used in distributed/mobile computing environment. This is possible due to miniaturization of communication technology. Many researchers have cited different types of technology in this context. But the application scenario are of important consideration while designing a specific protocol for Sensor network with reference to Cloud Computing. In this paper, we surveyed some typical applications of Sensor Network using Cloud computing as backbone. Since Cloud computing provides plenty of application, platforms and infrastructure over the Internet; it may combined with Sensor network in the application areas such as environmental monitoring, weather forecasting, transportation business, healthcare, military application etc. Bringing various WSNs deployed for different applications under one roof and looking it as a single virtual WSN entity through cloud computing infrastructure is novel.

***Index Terms*****—**Cloud Computing, Distributed Computing, Internet, Sensor Network, WSN

## 1. INTRODUCTION

The communication among sensor nodes using Internet is often a challenging issue. It makes a lot of sense to integrate sensor networks with Internet [1]. At the same time the data of sensor network should be available at any time, at any place. It is possibly a difficult issue to assign address to the sensor nodes of large numbers; so sensor node may not establish connection with internet exclusively. Cloud computing strategy can help business organizations to conduct their core business activities with less hassle and greater efficiency. Companies can maximize the use of their existing hardware to plan for and serve specific peaks in usage. Thousands of virtual machines and applications can be managed more easily using a cloud-like environment. Businesses can also save on power costs as they reduce the number of servers required.

Fig.1 consists of WSNs (i.e. WSN1, WSN2, and WSN3), cloud infrastructure and the clients. Clients seek services from the system. WSN consists of physical wireless sensor nodes to sense different applications like Transport Monitoring, Weather Forecasting, and Military Application etc. Each sensor node is programmed with the required application. Sensor node also consists of operating system components and network management components. On each sensor node, application program senses the application and sends back to gateway in the cloud directly through base station or in multi-hop through other nodes. Routing protocol plays a vital role in managing the network topology and to accommodate the network dynamics. Cloud provides on-demand service and storage resources to the clients. It provides access to these resources through internet and comes in handy when there is a sudden requirement of resources.

The organization of our work is as follows. In Section 2 & Section 3 we have presented an overview of Clouds and Sensor Network. In section 4 we have discussed various application scenarios of Sensor Network using Cloud Computing. Lastly, Section 5 concludes our work.

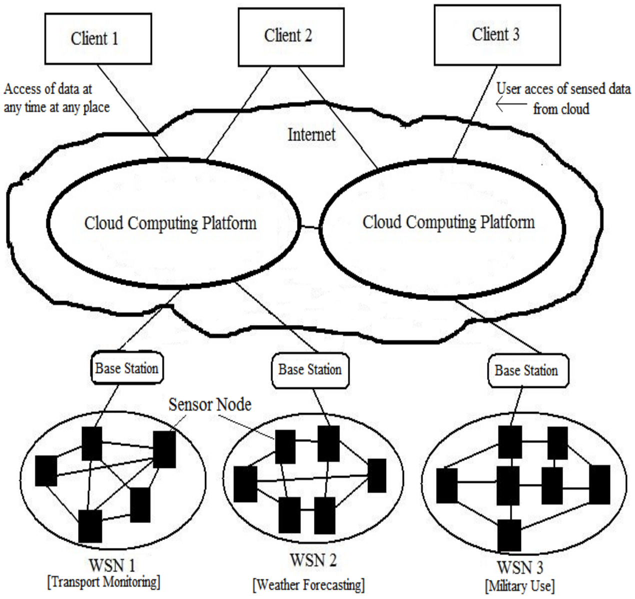


Fig. 1 WSN- Cloud Computing Platform

## 2. CLOUD: OVERVIEW

Cloud computing is a term used to describe both a platform and type of application. A cloud computing platform dynamically provisions, configures, reconfigures servers as needed. Servers in the cloud can be physical machines or virtual machines. It is an alternative to having local servers handle applications. The end users of a cloud computing network usually have no idea where the servers are physically located—they just spin up their application and start working. Advanced clouds typically include other computing resources such as storage area networks (SANs), network equipment, firewall and other security devices. Cloud computing also describes applications that are extended to be accessible through the Internet. These cloud applications use large data centers and powerful servers that host Web applications and Web services. Anyone with a suitable Internet connection and a standard browser can access a cloud application.

Many formal definitions have been proposed in both academia and industry, the one provided by U.S. NIST (National Institute of Standards and Technology) [2] appears to include key common elements widely used in the Cloud Computing community:

*Cloud computing is a model for enabling convenient, on demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction* [2].

### *A. Features*

The following are the essential features of cloud computing:

*1) Service on demand*: The request of the clients to avail resources can be fulfilled automatically without human interaction.

*2) Elasticity of demand:* There is no formal agreement or contract on the time period for using the resources. Clients can use the resources whenever they want and can release when they finish.

*3) Abstraction:* Resources are hidden to clients. Clients can only use the resources without having knowledge regarding location of the resource from where data will be retrieved and where data will be stored.

*4) Network access*: The client application can perform in various platform with the help of mobile phone, laptop and PDA using a secure internet connection.

*5) Service measurement:* Although computing resources are pooled and shared by multiple clients (i.e. multi-tenancy), the Cloud infrastructure can measure the usage of resources for each individual consumer through its metering capabilities.

*6) Resource pooling:* The resources are dynamically assigned as per clients' demand from a pool of resources [2].

### *B. Services*

The cloud provides following three services:

*1) SaaS(Software as a Service):* This model provides services to clients on demand basis. A single instance of the service runs on the cloud can serve multiple end user. No investment is required on the client side for servers and software licenses. Google is one of the service providers of SaaS.

*2) PaaS(Platform as a Service):* This model provides software or development environment, which is encapsulated & offered as a service and other higher level applications can work upon it. The client has the freedom to create his own applications, which run on the provider's infrastructure. PaaS providers offer a predefined combination of OS and application servers. Google's App Engine is a popular PaaS example.

*3) IaaS(Infrastructure as a Service):* This model provides basic storage and computing capabilities as standardized services over the network. Servers, storage systems, networking equipment, data centre space etc. are pooled and made available to handle workloads. The customer would typically deploy his own software on the infrastructure. The common example of IaaS is Amazon.

### *C. Cloud Computing Models*

The following models are presented by considering the deployment scenario:

*1) Private Cloud:* This cloud infrastructure is operated within a single organization, and managed by the organization or a third party irrespective of its location. The objective of setting up a private cloud in an organization is to maximize and optimize the utilization of existing in-house resources, providing security and privacy to data and lower data transfer cost [3] from local IT infrastructure to a Public Cloud.

*2) Public Cloud:* Public clouds are owned and operated by third parties. All customers share the same infrastructure pool with limited configuration, security protections, and availability variances. These are managed and supported by the cloud provider.

*3) Community Cloud:* This cloud infrastructure is constructed by number of organization jointly by making a common policy for sharing resources. The cloud infrastructure can be hosted by a third-party vendor or within one of the organizations in the community.

*4) Hybrid Cloud:* The combination of public and private cloud is known as hybrid cloud. In this model, service providers can utilize 3rd party Cloud Providers in a full or partial manner so that the flexibility for using the resources are increased.

## 3. Sensor Network: overview

A wireless sensor network (WSN) consists of spatially distributed autonomous sensors to cooperatively monitor physical or environmental conditions, such as temperature, sound, vibration, pressure, motion or pollutants.[4,5] The development of wireless sensor networks was motivated by military applications such as battlefield surveillance. They are now used in many industrial and civilian application areas, including industrial process monitoring and control, machine health monitoring [6], environment and habitat monitoring, healthcare applications, home automation, and traffic control [4, 7].Each node in a sensor network is typically equipped with a radio transceiver or other wireless communications device, a small microcontroller, and an energy source, usually a battery. The size of sensor node may vary from shoebox down to a grain of dust. The cost of sensor nodes is also varies from hundreds of dollars to a few pennies, depending on the size of the sensor network and the complexity required of individual sensor nodes [4]. Size and cost constraints on sensor nodes result in corresponding constraints on resources such as energy, memory, computational speed and bandwidth [4].

A sensor network is a computer network Composed of a large number of sensor nodes. [8] The sensor nodes are densely deployed inside the phenomenon, they deploy random and have cooperative capabilities. Usually these devices are small and inexpensive, so that they can be produced and deployed in large numbers, and so their resources in terms of energy, memory, computational speed and bandwidth are severely constrained. There are different Sensors such as pressure, accelerometer, camera, thermal, microphone, etc. They monitor conditions at different locations, such as temperature, humidity, vehicular movement, lightning condition, pressure, soil makeup, noise levels, the presence or absence of certain kinds of objects, mechanical stress levels on attached objects, the current characteristics such as speed, direction and size of an object. Normally these Sensor nodes consist there components: sensing, processing and communicating [9]. The development of sensor networks requires technologies from three different research areas: sensing, communication, and computing (including hardware, software, and algorithms). Thus, combined and separate advancements in each of these areas have driven research in sensor networks. Examples of early sensor networks include the radar networks used in air traffic control. The national power grid, with its many sensors, can be viewed as one large sensor network. These systems were developed with specialized computers and communication capabilities, and before the term “sensor networks” came into vogue.

### *A. Terminology*

Following are the important terms which are used widely in sensor network:

*1) Sensor:* A transducer that converts a physical phenomenon such as heat, light, sound or motion into electrical or other signal that may be further manipulated by other apparatus.

*2) Sensor node:* A basic unit in a sensor network, with processor, memory, wireless modem and power supply.

*3) Network Topology:* A connectivity graph where nodes are sensor nodes and edges are communication links.

*4) Routing:* The process of determining a network path from a source node to its destination.

*5) Resource:* Resource includes sensors, communication links, processors and memory and node energy.

*6) Data Storage:* The run-time system support for sensor network application. Storage may be local to the node where the data is generated, load balanced across a network, or anchored at a few points.

### *B. Routing Protocols in WSNs*

Routing protocols in WSNs are broadly divided into two categories: Network Structure based and Protocol Operation based. Network Structure based routing protocols are again divided into flat-based routing, hierarchical-based routing, and location-based routing. Protocol Operation based are again divided into Multipath based, Query based, QoS based, Coherent based and Negotiation based.

In flat-based routing, all nodes are typically assigned equal roles or functionality sensor nodes collaborate together to perform the sensing task. Due to the large number of such nodes, it is not feasible to assign a global identifier to each node. The examples of flat-based routing protocols are –SPIN [10,11], Directed Diffusion [12], Rumor Routing [13], MCFA [14], GBR[15], IDSQ & CADR [16], COUGAR [17], ACQUIRE [18], Energy Aware Routing [19] etc.

In hierarchical-based or cluster based routing, nodes will play different roles in the network. In a hierarchical architecture, higher energy nodes can be used to process and send the information while low energy nodes can be used to perform the sensing in the proximity of the target. This means that creation of clusters and assigning special tasks to cluster heads can greatly contribute to overall system scalability, lifetime, and energy efficiency. Hierarchical routing is an efficient way to lower energy consumption within a cluster and by performing data aggregation and fusion in order to decrease the number of transmitted messages to the BS. Hierarchical routing is mainly two-layer routing where one layer is used to select cluster heads and the other layer is used for routing. The examples of hierarchical-based routing protocols are – LEACH [20], PEGASIS [21], TEEN[22], APTEEN [23], MECN [24], SMECN [25], SOP[26], Sensor Aggregate routing [27], VGA[28], HPAR [29], TTDD [30] etc.

In location-based routing, sensor nodes' positions are exploited to route data in the network. In this kind of routing, sensor nodes are addressed by means of their locations. The distance between neighboring nodes can be estimated on the basis of incoming signal strengths. Relative coordinates of neighboring nodes can be obtained by exchanging such information between neighbors [37, 38, 39]. Alternatively, the location of nodes may be available directly by communicating with a satellite, using GPS (Global Positioning System), if nodes are equipped with a small low power GPS receiver [40]. The examples of location-based routing protocols are – GAF [31], GEAR [32], GPSR [33], MFR, DIR, GEDIR [34], GOAFR [35], SPAN [36] etc.

In multipath routing, communication among nodes uses multiple paths to enhance the network performance instead of single path. In Query based routing, the destination nodes propagate a query for data from a node through the network and a node having this data sends the data which matches the query back to the node, which initiates the query. Usually these queries are described in natural language, or in high-level query languages. In QoS-based routing protocols, the network has to balance between energy consumption and data quality. The network has to satisfy certain QoS metrics, e.g., delay, energy, bandwidth, etc for delivering data to the BS. In coherent routing, the data is forwarded to aggregators after minimum processing. The minimum processing typically includes tasks like time stamping, duplicate suppression, etc. In Negotiation based routing, protocols use high level data descriptors in order to eliminate redundant data transmissions through negotiation. Communication decisions are also taken based on the resources that are available to them.

## 4. APPLICATION SCENARIOS

Combining WSNs with cloud makes it easy to share and analyze real time sensor data on-the-fly. It also gives an advantage of providing sensor data or sensor event as a service over the internet. The terms *Sensing as a Service* (SaaS) and *Sensor Event as a Service* (SEaaS) are coined to describe the process of making the sensor data and event of interests available to the clients respectively over the cloud infrastructure.

Merging of two technologies makes sense for large number of application. Some applications of sensor network using cloud computing are explained below:

### *D. Transport Monitoring*

Transport monitoring system includes basic management systems like traffic signal control, navigation, automatic number plate recognition, toll collection, emergency vehicle notification, dynamic traffic light etc. [42].

In transport monitoring system, sensors are used to detect vehicles and control traffic lights. Video cameras are also used to monitor road segments with heavy traffic and the videos are sent to human operators at central locations. Sensors with embedded networking capability can be deployed at every road intersection to detect and count vehicle traffic and estimate its speed. The sensors will communicate with neighboring nodes to eventually develop a “global traffic picture” which can be queried by users to generate control signals. Data available from sensors is acquired and transmitted for central fusion and processing. This data can be used in a wide variety of applications. Some of the applications are - vehicle classification, parking guidance and information system, collision avoidance systems, electronic toll gates and automatic road enforcement.

In the above scenarios, both the applications require storage of data and huge computational cycles. They also require analysis and prediction of data to generate events. Access to this data is limited in both the cases. Integrating these WSN applications with the cloud computing infrastructure will ease the management of storage and computational resources. It also provides an improvement on the application data over the internet through web.

### *A. Military Use*

Sensor networks are used in the military for Monitoring friendly forces, equipment and ammunition, Battlefield surveillance, Reconnaissance of opposing forces, Targeting, Battle damage assessment and Nuclear, biological and chemical attack detection reconnaissance etc [43].

The data collected from these applications are of greatest importance and needs top level security which may not be provided using normal internet connectivity for security reason. Cloud computing may be one of the solution for this problem by providing a secure infrastructure exclusively for military application which will be used by only Defense Purpose.

### *B. Weather Forecasting*

Weather forecasting is the application to predict the state of the atmosphere for a future time and a given location. Weather monitoring and forecasting system typically includes- Data collection, Data assimilation, Numerical weather prediction and Forecast presentation [41].

Each weather station is equipped with sensors to sense the following parameters—wind speed/direction, relative humidity, temperature (air, water and soil),barometric pressure, precipitation, soil moisture, ambient light (visibility), sky cover and solar radiation. The data collected from these sensors is huge in size and is difficult to maintain using the traditional database approaches. After collecting the data, assimilation process is done. The complicated equations that govern how the state of the atmosphere changes (weather forecast) with time require supercomputers to solve them.

### *C. Health Care*

Sensor networks are also widely used in health care area. In some modern hospital sensor networks are constructed to monitor patient physiological data, to control the drug administration track and monitor patients and doctors and inside a hospital.

In the above scenario, the data collected from the patients are very sensitive and should be maintained properly as collected data are required by the doctors for their future diagnosis. In traditional approach the patient’s history database is maintained in the local nursing home. So reputed doctors who are specially invited from abroad to handle critical cases cannot analyze the patient’s disease frequently. They will only make diagnosis when they will visit the particular nursing home. This problem may be solved by forming a cloud where the critical data of the patients can be maintained and authorized doctors sitting in abroad can analyze the data and give proper treatment.

## 5. CONCLUSION

The communication among sensor nodes using Internet is a challenging task since sensor nodes contain limited band width, memory and small size batteries. The issues of storage capacity may be overcome by widely used cloud computing technique. In this paper, we have discussed some issues of cloud computing & sensor network. To develop a new protocol in sensor network, the specific application oriented scenarios are of important consideration. Keeping this in mind we have discussed some application of Sensor Network using Cloud Computing.

## REFERENCES

[1] C. Ulmer, L. Alkalai and S. Yalamanchili, Wireless distributed sensor networks for in-situ exploration of mars, Work in progress for NASA Technical Report. Available in: http://users.ece.gatech.edu/

[2] P. Mell and T. Grance, "Draft nist working definition of cloud computing - v15," *21. Aug 2005,* 2009.

[3] M. Armbrust, A. Fox, R. Griffith, A. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, and I. Stoica, "Above the clouds: A Berkeley view of cloud computing," *EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2009-28,* 2009.

[4] Römer, Kay; Friedemann Mattern (December 2004), "The Design Space of Wireless Sensor Networks", *IEEE Wireless Communications* 11 (6): 54–61, doi:10.1109/MWC.2004.1368897, http://www.vs.inf.ethz.ch/publ/papers/wsn-designspace.pdf

[5] Thomas Haenselmann (2006-04-05). *Sensornetworks*. GFDL Wireless Sensor Network textbook. http://pi4.informatik.uni-mannheim.de/~haensel/sn_book. Retrieved 2006-08-29.

[6] Tiwari, Ankit et. al, *Energy-efficient wireless sensor network design and implementation for condition-based maintenance*, ACM Transactions on Sensor Networks (TOSN), http:// portal.acm.org/citation.cfm?id=1210670

[7] Hadim, Salem; Nader Mohamed (2006), "Middleware Challenges and Approaches for Wireless Sensor Networks", *IEEE Distributed Systems Online* **7** (3): 1, doi:10.1109/MDSO.2006.19, http://doi.ieeecomputersociety.org/10.1109/MDSO.2006.19 art. no. 0603-o3001.

[8] http://en.wikipedia.org/wiki/Sensor_Networks

[9] Akyildiz, I.F., W. Su, Y. Sankarasubramaniam, E. Cayirci, "A Survey on Sensor Networks", IEEE Communications Magazine, August, 102-114(2002).

[10] W. Heinzelman, J. Kulik, and H. Balakrishnan, "Adaptive Protocols for Information Dissemination in Wireless Sensor Networks," Proc. 5th ACM/IEEE Mobicom Conference (MobiCom '99), Seattle, WA, August, 1999. pp. 174-85.

[11] J. Kulik, W. R. Heinzelman, and H. Balakrishnan, "Negotiation-based protocols for disseminating information in wireless sensor networks," Wireless Networks, Volume: 8, pp. 169-185, 2002.

[12] C. Intanagonwiwat, R. Govindan, and D. Estrin, "Directed diffusion: a scalable and robust communication paradigm for sensor networks," Proceedings of ACM MobiCom '00, Boston, MA, 2000, pp. 56-67.

[13] D. Braginsky and D. Estrin, \Rumor Routing Algorithm for Sensor Networks," in the Proceedings of the First Workshop on Sensor Networks and Applications (WSNA), Atlanta, GA, October 2002.

[14] F. Ye, A. Chen, S. Liu, L. Zhang, \A scalable solution to minimum cost forwarding in large sensor networks", Proceedings of the tenth International Conference on Computer Communications and Networks (ICCCN), pp. 304-309, 2001.

[15] C. Schurgers and M.B. Srivastava, \Energy efficient routing in wireless sensor networks", in the MILCOM Proceedings on Communications for Network-Centric Operations: Creating the Information Force, McLean, VA, 2001.

[16] M. Chu, H. Haussecker, and F. Zhao, \Scalable Information-Driven Sensor Querying and Routing for ad hoc Heterogeneous Sensor Networks," The International Journal of High Performance Computing Applications, Vol. 16, No. 3, August 2002.

[17] Y. Yao and J. Gehrke, \The cougar approach to in-network query processing in sensor networks", in SIGMOD Record, September 2002.

[18] N. Sadagopan et al., The ACQUIRE mechanism for efficient querying in sensor networks, in the Proceedings of the First International Workshop on Sensor Network Protocol and Applications, Anchorage, Alaska, May 2003.

[19] R. C. Shah and J. Rabaey, \Energy Aware Routing for Low Energy Ad Hoc Sensor Networks", IEEE Wireless Communications and Networking Conference (WCNC), March 17-21, 2002, Orlando, FL.

[20] W. Heinzelman, A. Chandrakasan and H. Balakrishnan, "Energy-Efficient Communication Protocol for Wireless Microsensor Networks," Proceedings of the 33rd Hawaii International Conference on System Sciences (HICSS '00), January 2000.

[21] S. Lindsey, C. Raghavendra, \PEGASIS: Power-Efficient Gathering in Sensor Information Systems", IEEE Aerospace Conference Proceedings, 2002, Vol. 3, 9-16 pp. 1125-1130.

[22] A. Manjeshwar and D. P. Agarwal, "TEEN: a routing protocol for enhanced efficiency in wireless sensor networks," In 1st International Workshop on Parallel and Distributed Computing Issues in Wireless Networks and Mobile Computing, April 2001.

[23] A. Manjeshwar and D. P. Agarwal, "APTEEN: A hybrid protocol for efficient routing and comprehensive information retrieval in wireless sensor networks," Parallel and Distributed Processing Symposium., Proceedings International, IPDPS 2002, pp. 195-202.

[24] V. Rodoplu and T. H. Meng,\Minimum Energy Mobile Wireless Networks", IEEE Journal Selected Areas in Communications, vol. 17, no. 8, Aug. 1999, pp. 133344.

[25] L. Li, and J. Y. Halpern,\Minimum-Energy Mobile Wireless Networks Revisited," IEEE International Conference on Communications (ICC) 2001. Vol. 1, pp. 278-283.

[26] L. Subramanian and R. H. Katz, \An Architecture for Building Self Configurable Systems", in the Proceedings of IEEE/ACM Workshop on Mobile Ad Hoc Networking and Computing, Boston, MA, August 2000.

[27] Q. Fang, F. Zhao, and L. Guibas, \Lightweight Sensing and Communication Protocols for Target Enumeration and Aggregation", Proceedings of the 4th ACM international symposium on Mobile ad hoc networking and computing (MOBIHOC),
2003, pp. 165-176.

[28] Jamal N. Al-Karaki, Raza Ul-Mustafa, Ahmed E. Kamal, "Data Aggregation in Wireless Sensor Networks - Exact and Approximate Algorithms'", Proceedings of IEEE Workshop on High Performance Switching and Routing (HPSR) 2004, April 18-21, 2004, Phoenix, Arizona, USA.

[29] Q. Li and J. Aslam and D. Rus,\Hierarchical Power-aware Routing in Sensor Networks", In Proceedings of the DIMACS Workshop on Pervasive Networking, May, 2001.

[30] F. Ye, H. Luo, J. Cheng, S. Lu, L. Zhang, \A Two-tier data dissemination model for large-scale wireless sensor networks", proceedings of ACM/IEEE MOBICOM, 2002.

[31] Y. Xu, J. Heidemann, D. Estrin,\Geography-informed Energy Conservation for Ad-hoc Routing," In Proceedings of the Seventh Annual ACM/IEEE International Conference on Mobile Computing and Networking 2001, pp. 70-84.

[32] Y. Yu, D. Estrin, and R. Govindan, \Geographical and Energy-Aware Routing: A Recursive Data Dissemination Protocol for Wireless Sensor Networks", UCLA Computer Science Department Technical Report, UCLA-CSD TR-01-0023, May 2001.

[33] B. Karp and H. T. Kung, \GPSR: Greedy perimeter stateless routing for wireless sensor networks", in the Proceedings of the 6th Annual ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom '00), Boston, MA, August 2000.

[34] I. Stojmenovic and X. Lin. \GEDIR: Loop-Free Location Based Routing in Wireless Networks", In International Conference on Parallel and Distributed Computing and Systems, Boston, MA, USA, Nov. 3-6, 1999.

[35] F. Kuhn, R. Wattenhofer, A. Zollinger,\Worst-Case optimal and average-case efficient geometric ad-hoc routing", Proceedings of the 4th ACM International Conference on Mobile Computing and Networking, Pages: 267-278, 2003.

[36] B. Chen, K. Jamieson, H. Balakrishnan, R. Morris, \SPAN: an energy-efficient coordination algorithm for topology maintenance in ad hoc wireless networks", Wireless Networks, Vol. 8, No. 5, Page(s): 481-494, September 2002.

[37] N. Bulusu, J. Heidemann, D. Estrin,\GPS-less low cost outdoor localization for very small devices", Technical report 00-729, Computer science department, University of Southern California, Apr. 2000.

[38] A. Savvides, C-C Han, aind M. Srivastava,\Dynamic ¯ne-grained localization in Ad-Hoc networks of sensors," Proceedings of the Seventh ACM Annual International Conference on Mobile Computing and Networking (MobiCom), July 2001. pp. 166-179.

[39] S. Capkun, M. Hamdi, J. Hubaux,"GPS-free positioning in mobile ad-hoc networks", Proceedings of the 34th Annual Hawaii International Conference on System Sciences, 2001 pp. 3481-3490.

[40] Y. Xu, J. Heidemann, D. Estrin,\Geography-informed Energy Conservation for Ad-hoc Routing," In Proceedings of the Seventh Annual ACM/IEEE International Conference on Mobile Computing and Networking 2001, pp. 70-84.

[41] http://en.wikipedia.org/wiki/Weather_forecasting

[42] http://en.wikipedia.org/wiki/Intelligent_transportation_system
[43] Chee-Yee Chong; Kumar, S.P., "Sensor networks: Evolution, opportunities, and challenges, "Proc IEEE, August 2003

**Sanjit Kumar Dash** received the B.Tech. degree in Information Technology from Biju Patnaik University of Technology, Orissa, India, in 2004 and pursuing M.Tech. degree in Computer Science and Engineering at Institute of Technical Education and Research, Bhubaneswar, India. He is also working as a faculty member at the Information Technology Department, College of Engineering and Technology, Bhubaneswar, India. His research interests include Cloud Computing, Sensor Network, and Mobile Computing.

**Subasish Mohapatra** received the B.Tech. degree in computer science and engineering from Biju Patnaik University of Technology, Orissa, India, in 2003 and the M.Tech. degree in computer science and information technology from College of Engineering and Technology, Bhubaneswar, India, in 2007. He was a faculty member at the Computer Science and Engineering Department, College of Engineering and Technology, Bhubaneswar, India, during 2004–2009. Currently, he is an Assistant Professor at the Department of Computer Science and Engineering at the Institute of Technical Education and Research, Bhubaneswar, India. His research interests include Virtualization, Ad-hoc Network, and Sensor Network.

**Prasant Kumar Pattnaik** , M.Tech, PhD. At present working as a Professor and Head of Department of Computer Science and Engineering at Konark Institute of Science and Technology, Bhubaneswar, India. He is a senior member of International Association of Computer Science and Information Technology, Singapore. His research interests include Ad-hoc and Sensor Network.

# An Intelligent Routing Strategy for Ad hoc Networks

Shailender Gupta[1] and C. K. Nagpal[2]

[1]YMCA University of Science and Technology,
[2]YMCA University of Science and Technology,
Corresponding Adresses
{Shailender81@gmail.com, nagpalckumar@rediffmail.com}

***Abstract***: The advancement in computing technology has led to the design, development and deployment of high end portable computing devices. These devices are free to move along with their users. Due to their mobility these devices are designed to share the information through temporary networks which don't require much of infrastructure. The ad hoc network is one such network, which can even be deployed in war hit or disaster hit areas where the basic infrastructure facilities such as power network or communication network are destroyed. In the ad hoc network, the data transmission is through intermediate nodes to save the power and every participating node has to work as the router as well. This feature of ad hoc networks has led to the enormous possible routing algorithm design. The basic criteria for efficiency of a routing algorithm include high throughput, minimum routing overhead, path optimality, minimum average delay and minimum packet lost. The literature of ad hoc network contains many routing algorithms, some of these are proactive and others are reactive with each having its own merits and demerits. In such a scenario, hybrid routing can take advantages of both. This paper presents an efficient hybrid routing algorithm.

***Keywords***: Ad hoc Networks, Routing Protocols.

## 1. Introduction

W The origination of Mobile Ad hoc Networks (MANET) from packet radio network (PRNET) [1] and SURAN project [2] has not only rendered the mobile and wireless networks important but the design of an efficient and portable routing protocol a challenging task for the researchers. While designing a routing protocol a designer has to keep in mind the following performance metrics [5]:

Throughput: Defined as total number of packets received by the destination.

Routing overhead: The ratio between the total numbers of routing packets transmitted to the total number of data packets.

Path optimality: The difference between the number of hops a packet takes to reach its destination and the length of the shortest path that physically existed.

Packets lost: Measure of the number of packets dropped by the routers due to various reasons.

Average Delay: Average amount of time taken by a packet to go from source to destination.

A critical aspect involved in the design of ad hoc routing protocol is to ensure the use of minimum power as the devices involved are battery driven and power constrained especially when deployed in the disaster hit area or in war situation where [3, 4, 5] there are no infrastructural facilities. The metrics such as packets lost, high average delay and path optimality are directly related to the power consumption. These aspects of the network can be improved by minimizing the number of collisions [6], reducing congestion and by optimizing the hop count [7, 8]. Thus the basic goal of the routing protocol is to increase the throughput of the network with minimum usage of power.

The paper has been divided in five sections. The Section 2 constitutes literature survey. Section 3 constitutes the details of the proposed protocol. Section 4 provides the simulation results. The paper completes with conclusion and future scope.

## 2. LITERATURE SURVEY

The routing protocols can be classified into three categories proactive, reactive and hybrid routing. A proactive routing is also called as table driven routing because in this each node has to maintain routing tables. The tables are updated periodically or on demand basis so as to have complete topological structure of the network. Therefore, a source node can get routing path immediately when it requires. Most proactive routing protocols have inherited properties of wired routing protocols so there are certain limitations such as high control overhead, the scalability is poor but there lies the advantage of using these routing schemes that the time complexity involved in searching the route to the destination is O(1). Reactive routing is also called as on demand routing as the routes are formed whenever there is a requirement. A route discovery operation invokes a route-determination procedure. The discovery procedure terminates either when a route has been found or no route is available after examination for all route permutations. This involves certain delay in routing the data packets. Though the scheme has delay while routing but the control overhead is quite low and scalability is quite good when compared with proactive routing schemes. Hybrid routing protocols can be designed to derive the merits of both proactive and reactive routing protocols and avoid their shortcomings. For example, a hybrid routing protocol such as ZRP assumes hierarchical network architectures with proactive routing approach for intra domain routing and reactive routing approach for inter domain routing. The merit of ZRP hybrid protocol is the reduced control overhead of as compared to proactive routing approaches and the reduced latency as compared to route search operations in reactive routing approaches.

### 2.1 Table driven routing protocol (proactive routing)

T In table driven routing protocols the nodes maintains routing table to store paths for each possible destination. The tables are continuously updated by message passing techniques. The updates can be periodic or on demand that is when a change occurs in the network. The popular routing protocols in this category are Wireless Routing Protocol (WRP) [10], Destination Sequence Distance Vector (DSDV) [11], and Fisheye State Routing (FSR) [12] etc. The proactive routing is being described with the help of DSDV. The DSDV protocol [11] is similar to WRP but its routing mechanism is quite different. In routing tables of DSDV, an entry stores the next hop towards a destination, the cost metric for the routing path to the destination and a destination sequence number that is created by the destination. Sequence numbers are used in DSDV to distinguish stale routes from fresh ones and avoid formation of route loops. To reduce the overhead of control packet incremental updates along with full dump updates are used. The incremental updates are used when there are minor changes in the topology or when the network is less mobile and full dump updates are sent when ever there are major changes in the topology or the network is mobile.

Disadvantages

- Memory requirement: The need to maintain routing tables results in a large memory requirement
- Bandwidth requirement: Though it uses incremental updates to reduce the routing messages but still the number of routing messages exchanged is quite large resulting in high bandwidth requirement.
- Scalability: The table driven routing protocols are not easy to scale. The scalability is quite poor.

### 2.2 On Demand routing protocol (Reactive routing)

Reactive routing protocols for mobile ad hoc networks are also called "on-demand" routing protocols. In a reactive routing protocol, routing paths are searched only when needed. A route discovery operation invokes a route-determination procedure. The discovery procedure terminates either when a route has been found or no route available after examination for all route permutations. The popular routing protocols in this category are Ad hoc On Demand Distance Vector Routing (AODV), Dynamic Source Routing (DSR) [14], and Temporally Ordered Routing Algorithm (TORA) [18, 19]. The basic operation of on demand routing is explained with the help of AODV. In AODV [13], routing information is maintained in routing tables at nodes. Every mobile node keeps a next-hop routing table, which contains the destinations to which it currently has a route. A routing table entry expires if it has not been used or reactivated for a pre-specified expiration time. In AODV if a source node wants to send a packet to the destination node then firstly it starts route discovery process in which a route request packet is broadcasted by the source node. The route request packet contains the following information

- Source node id
- Destination node id
- Sequence number

When this packet reaches to a node it records originator information as well as from whom the packet is sent to it in its route cache. The node process the route request only if it has not seen this RREQ previously and this is how the loop formation is avoided. The process of broadcasting continues till the RREQ reaches to the destination node or to a node which has information about the destination node in its route cache. After the route request phase the route reply packet is generated using the same path as selected in case of route request path selected.

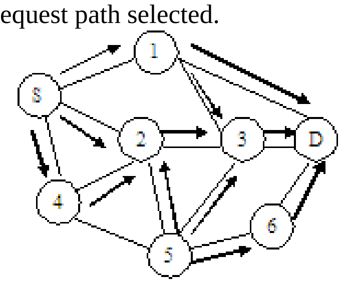


Figure.1a The Route Request packets flooding in AODV

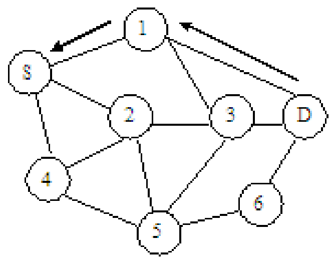


Figure. 1b The Route Reply packet in AODV

As shown in the above Fig. 1a the source node (S) broadcast route request to all the neighboring nodes 1, 2, and 4 which in turn broadcast the route request to all their neighboring nodes. The process goes until the packet reaches to the destination node (D). The route request packet reaches the destination from different paths but the path from which the packet reaches first is selected for route acknowledgement as shown in the Fig. 1b.

Disadvantage: AODV does not support unidirectional and multiple routing paths as supported by DSR.

### 2.3 Hybrid Routing Protocols

Hybrid routing protocols are proposed to combine the merits of both proactive and reactive routing protocols and overcome their shortcomings. Normally, hybrid routing protocols for mobile ad hoc networks exploit hierarchical network architectures. Proper proactive routing approach and reactive routing approach are exploited in different hierarchical levels, respectively. The protocols that fall in this category are Zone Routing Protocol (ZRP) [15], The Zone-based Hierarchical Link State routing (ZHLS) [20], and the Hybrid Ad Hoc Routing Protocol (HARP) [21]. The hybrid protocols are explained with the help of ZRP. In ZRP, the network is divided into routing zones according to distances between mobile nodes. Given a hop distance d and a node N, all nodes within hop distance at most d from N belong to the routing zone of N. Peripheral nodes of N are N’s neighboring nodes in its routing zone which are exactly d hops away from N. In ZRP, different routing approaches are exploited for inter-zone and intra-zone packets. The proactive routing approach, i.e., the Intra-zone Routing protocol (IARP), is used inside routing zones and the reactive Inter-zone Routing Protocol (IERP) is used between routing zones, respectively. The IARP maintains link state information for nodes within specified distance d. Therefore, if the source and destination nodes are in the same routing zone, a route can be available immediately. Most of the existing proactive routing schemes can be used as the IARP for ZRP. The IERP reactively initiates a route discovery when the source node and the destination are residing in different zones. The route discovery in IERP is similar to DSR with the exception that route requests are propagated via peripheral nodes.

- The hybrid protocols are proposed to reduce the control overhead of proactive routing approaches
- Decrease in the latency caused by route search operations in comparison with reactive routing approach..

## 3. THE PROPOSED PROTOCOL

To understand the working of the proposed protocol, familiarity with the basic packets and tables involved is necessary. Therefore, we start this section with the brief introduction about these aspects.

### 3.1 Data Packet Format

This packet is used for exchange of data between the mobile nodes. The basic format of data packet header is given in Fig. 2 Each data packet header will have several fields like packet type, source address of the node that initiates the packet, destination address of the node to which the packet must finally be handed over, a list of addresses of previously visited nodes and hop count indicating the no. of intermediate nodes in the path.

| Packet Type | Source Address | Destination Address | Visited Nodes | Hop count | Data |
|---|---|---|---|---|---|

Figure 2 Format of data packet header

*Packet Type: identifies the type of packet*
*Source Address: denotes the address of source that initiates the packet*
*Destination Address: denotes the address of node that finally receives the packet*
*Visited nodes: List of intermediate nodes*
*Hop count: total number of intermediate nodes*
*Data: data to be sent to the destination (payload)*

### 3.2 Nodes Gateway Table (NGT)

Each node in the network maintains information about its neighboring nodes by broadcasting a hello request and receiving a reply packet in turn. Fig. 3 shows a representative ad hoc network in which each node has been shown connected to its neighbors with the help of dotted lines. The Table 1 shows the different nodes in the network with the list of their neighbors.

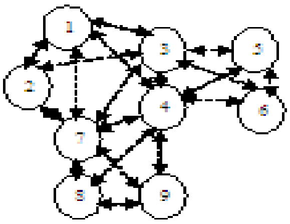


Figure. 3 Ad Hoc Network

**Table 1** Neighboring Node Information

| Node | Neighboring Nodes | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | 3 | 4 | 7 | 2 | | | |
| 2 | 1 | 7 | 3 | | | | |
| 3 | 4 | 5 | 6 | 1 | 2 | 7 | |
| 4 | 1 | 3 | 5 | 6 | 7 | 8 | 9 |
| 5 | 3 | 4 | 6 | | | | |
| 6 | 3 | 4 | 5 | | | | |
| 7 | 1 | 2 | 3 | 4 | 8 | 9 | |
| 8 | 7 | 9 | 4 | | | | |
| 9 | 8 | 7 | 4 | | | | |

Thereafter, every node multicasts a Gateway request packet to its neighbors listed in the neighbor table to determine the best first intermediate node its communicate with the rest of the nodes in the network except for immediate neighbors. We name such a node as the gateway node as it acts as the gateway for the node (under consideration) to communicate with rest of the network. For example, according to the Table 1, Node 1 will multicast gateway request packets to nodes 3,4,7,2 (see Fig. 5) to find its gateway for rest of the network barring the immediate neighbors 3, 4, 7 and 2. However, there is no need to search gateway if the destination node is next neighbor i.e 3, 4, 7, or 2. Similar process is repeated by the other nodes to search for their gateway. Fig. 4 shows the basic design of Gateway Request Packet.

| Packet Type | Source Address | Destination Address |
|---|---|---|

Figure.4 Gateway request packet

*Packet type: identifies the type of packet*
*Source Address: denotes the address of source that initiates the packet*
*Destination Address: denotes the address of node that finally receives the packet*

Where packet type identifies the packet as a gateway request packet, source address contains the address of the sender node and destination address contains the broadcast address.

| | | |
|---|---|---|
| Gateway Request packet | Node 1 | Node 3 |

Figure 5 Gateway request packet of node 1 for node 3

When a gateway request packet is received by a node, it sends a gateway reply packet to the sender node in the format as shown in Fig. 6a.

Figure 6a Gateway Reply packet

The Fig. 6b shows the Node 3 gateway reply packet when it receives gateway request packet from its neighbors.

| | | | |
|---|---|---|---|
| Gateway Reply packet | Node 3 | Node 1 | 4,5,6,1,2, 7 |

Figure 6b Gateway Reply packet of node 3

After getting complete neighbor list from the neighboring nodes, the node compares this list with its neighboring node list and creates gateway priority table according to the rule that a node with highest new nodes information in comparison to the source node neighbor table is selected as the best gateway. The process of searching the new node information with all the neighboring nodes is continued till a sorted gateway list is created. The process can be made clearer with the help of Table 2.

- The node 4 has four new nodes information in comparison to node 1 so it is given the first gateway priority.
- The node 3 and 7 has only two new nodes information in comparison to 1 so both of them should have same priority but the node 3 has been given higher priority on the basis of serial number.
- The node 1 has four neighbors 3, 4, 7, 2 while the node 2 has three neighbors 1, 7, 3 there is no new node information for node 1 so it is given the lowest priority.

Table 2 Gateway priority table of node 1

| Node | Neighboring Nodes | | | | | | | Gateway priority |
|---|---|---|---|---|---|---|---|---|
| 1 | 3 | 4 | 7 | 2 | | | | |
| 2 | 1 | 7 | 3 | | | | | 4 |
| 3 | 4 | 5 | 6 | 1 | 2 | 7 | | 2 |
| 4 | 1 | 3 | 5 | 6 | 7 | 8 | 9 | 1 |
| 7 | 1 | 2 | 3 | 4 | 8 | 9 | | 3 |

- It may be noted that every node creates a table called Nodes Gateway Table. The entries of neighboring nodes in NGT are stored in accordance with decreasing value of gateway priority as shown in Table 3 for node 1.

Table 3 NGT of node 1

| Nodes | Nodes gateway table | | | |
|---|---|---|---|---|
| 1 | 4 | 7 | 3 | 2 |

The nodes in the ad-hoc network repeat the process of gathering the NGT (Nodes Gateway Table) entries in the table until every node has details about their neighboring nodes.

### 3.3 Algorithmic Details

To initiate or forward data packets the nodes use information contained in NGT table. Whenever due to mobility of nodes there is a packet loss, the NGT entries are updated. However, after a regular interval of time, every node within a cell retransmits a gateway Request packet in order to update the network information. This process helps nodes in acting as per the latest network topology.

| | | | |
|---|---|---|---|
| Packet Type | Source Address | Destination Address | List of Neighbors |

**The Proposed Routing Algorithm**

```
Dp.hop_count=16;
If (DP.hop_count <1)
Drop the packet  /* DP has to be retransmitted from the source of DP
else
{
   if( DP.Dest_Addr.== Node Address )
   {
   Consume the data packet;
   }
   else
   {
    DP.hop_count -- ;
    Best Gateway selection ();
    }
  }
```

*The algorithm says that, initially the hop count is set to 16 and when the hop count becomes less then 1 the data packet has again to be retransmitted and if the count is greater then 1 the data packet will call the gateway selection algorithm, which is discussed, in the next section.*

```
 Best Gateway Selection Algorithm
Best Gateway selection ( )
{
i=1;
                while(sort_gate_list!=null)
{
                if (sort_gate_list[I]==visited node)
                  /*If the Ith entry of the sorted gateway list
                    Matches with the visited node if yes then
                    there is no need to visit it again*/
                {
                 i++;  /* check the next entry of the
                                  Visited node*/
                 }
                 else
                {
                 Unicast the packet to that sort_gate_list[i];
                  /* deliver the packet to the best gateway*/
                exit( )_;
                }
                }
                     if all the nodes are visited
                         {
```

```
    Set i=1;
    Select sort_gate_list [i];
                          }
  }
```

## 4. Simulation Results

A simulator was designed in C++ in which an area of 35*35 sq. unit's size was chosen (see Fig. 7). The nodes were distributed randomly in the given area and following performance metrics [16] results were recorded in an output file

1. Average Power left per node
2. Average Throughput
3. Average no. of hop count for successful transmission
4. Average no. of retransmission required

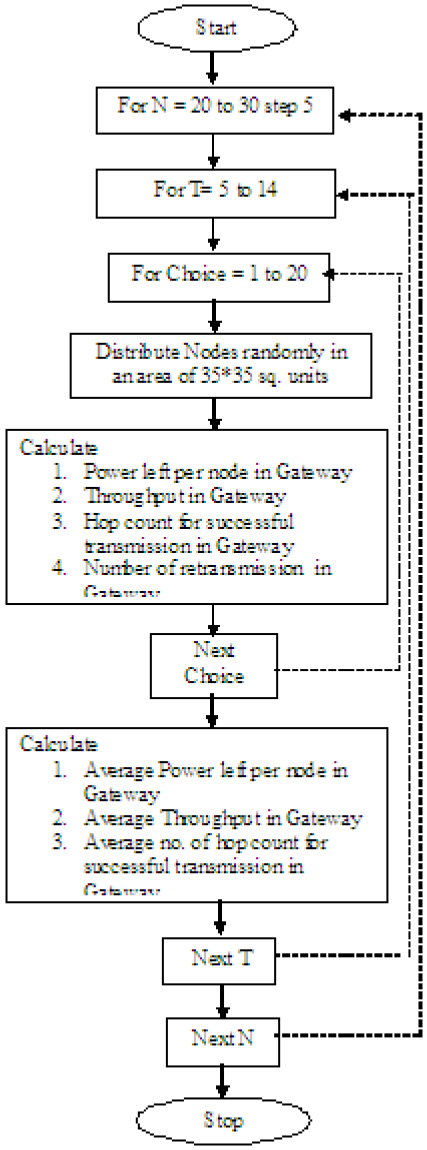

Start
For N = 20 to 30 step 5
For T= 5 to 14
For Choice = 1 to 20
Distribute Nodes randomly in an area of 35*35 sq. units
Calculate
1. Power left per node in Gateway
2. Throughput in Gateway
3. Hop count for successful transmission in Gateway
4. Number of retransmission in Gateway
Next Choice
Calculate
1. Average Power left per node in Gateway
2. Average Throughput in Gateway
3. Average no. of hop count for successful transmission in Gateway
Next T
Next N
Stop

Figure 7 Flow chart of simulator

The following assumptions were made while measuring the above mentioned parameters

1. Antennas are Omni directional
2. Sleep mode power dissipation has been considered as null
3. Mobility has not been taken into consideration

The following assumptions were made in measuring average power:

1. The node looses 2 units battery in transmitting a packet
2. The node looses 1.5 units of its battery power while receiving

Initially each node was given 100 units of power. The simulator designed selects random source and destination every run. Each execution of program involved 20 runs. The average power left per node after all the 20 execution was recorded in an output file as shown in Fig 8.

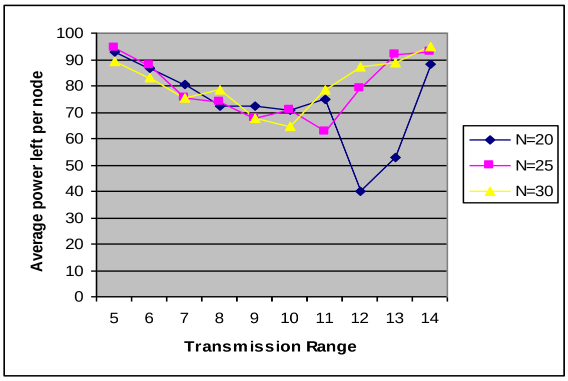


Figure 8 Battery status after 20 transmissions for different transmission radius

The Average throughput increases as the transmission range increases due to the fact that with increase in transmission radius neighboring nodes get increased (see Fig. 9) resulting in higher probability for a data packet to reach to its destination within permissible hop count.

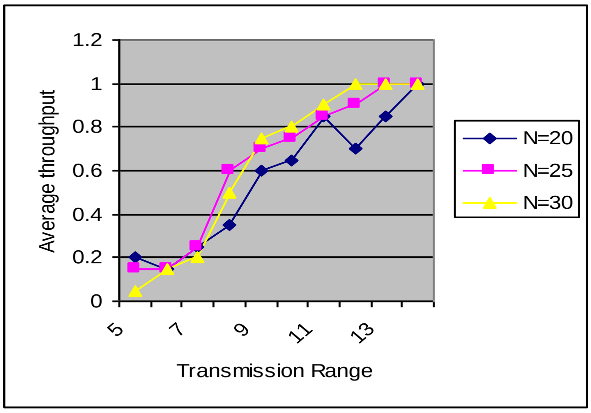


Figure 9 Average Throughput for different transmission radius

The average number of intermediate nodes that a packet takes to reach to its destination for successful transmission is shown in Fig. 10. As seen in the graph the average number of hops are fewer at lower transmission range since the number of neighboring nodes is quite less as well as the average throughput is also very low. As the transmission range is increased the average numbers of hops for successful transmission increases since the number of neighboring nodes gets increased as well as the throughput also get increased as shown in Fig. 10.

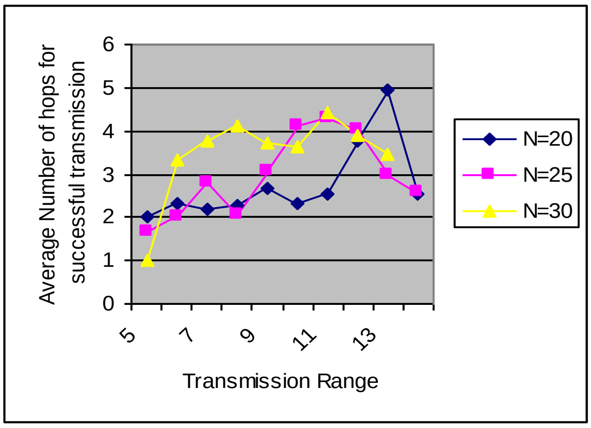


Figure 10 Average numbers of intermediate hops for different transmission radius

The average number of retransmission decreases as the transmission range increases as shown in the graph since as the transmission range increases the number of neighbor gets increased and the probability to reach to the destination also gets increased (see Fig. 11).

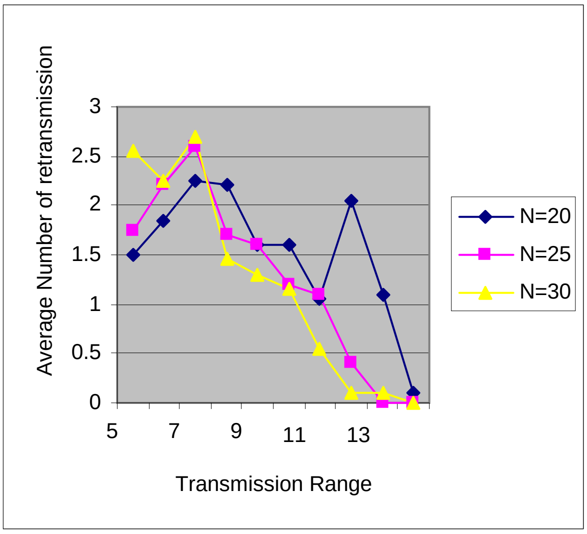


Figure 11 Average Retransmission for different transmission radius

## Conclusion and Future Scope

The proposed routing protocol tries to select the node with highest available gateway priority. The advantage of this protocol is its heuristic approach to find the route. The proposed approach uses both proactive and reactive counterparts for route discovery thereby utilizing the advantages of both. Most of the transmissions used for finding the route involve multicasting which is in contrast to the AODV, which use blind search and major portion of the signals in route finding process involves broadcasting. This leads to major saving of the energy.

As described earlier, the proposed protocol has been implemented in isolation by designing a simulator in C++. Depending upon the requirement the proposed protocol can be implemented on simulators like NS2 or Qualnet and be compared with other existing protocols in the similar conditions as per the requirement. The readers can ask for the C++ code by contacting the authors.

## References

[1] John Jubin and Janet ,D.Tornow, “The DARPA packet radio network protocols”. proceeding of IEEE 75(1):21-32-jan1987.

[2] Gregray S.Lawer, “Packet Radio Routing in Communication Network” edited by Martha.E.Steen Strup,prentice hall, Engle wood cliffs, New Jersey .1995

[3] C. S. R. Murthy and B. S. Manoj, Ad hoc Wireless Networks: Architectures and Protocols, Prentice Hall PTR, 2004.

[4] C. –K. Toh, Ad hoc Mobile Wireless Networks: Protocols and Systems, Prentice Hall PTR, 2002.

[5] M. S. Curson and J. Macker, “Mobile Ad hoc Networking (MANET): Routing Protocol Performance Issues and Evaluation Considerations,” http://www.faqs.org/rfcs/rfc2501.html, Network Working Group Request for Comments: 2501, January 1999.

[6] Chun-Cheng Chen, Hwangnam Kim, and Haiyun Luo SELECT: “Self-Learning Collision Avoidance for Wireless Networks” IEEE TRANSACTIONS ON MOBILE COMPUTING, VOL. 7, NO. 3, MARCH 2008

[7] C. R. Lin and J.-S. Liu, “QoS Routing in Ad Hoc Wireless Networks,” IEEE JSAC, vol. 17, no. 8, Aug. 1999, pp. 1426–38.

[8] T. Shivaprakash, G. S. Badrinath, N. Chandrakanth, K. R. Venugopal, L. M. Patnaik, “Energy Efficient Routing in Adhoc Networks” IEEE 2006

[9] Changling liu, jorg Kaiser “A survey of mobile ad hoc network routing protocol” IEEE 2003.

[10] Murthy, S. and J.J. Garcia-Luna-Aceves, An Efficient Routing Protocol for Wireless Networks, ACM Mobile Networks and App. J., Special Issue on Routing in Mobile Communication Networks, Oct. 1996, pp. 183-97.

[11] C. E. Perkins and P. Bhagwat. Highly dynamic Destination-Sequenced Distance-Vector Routing (DSDV) for mobile computers, ACM Computer Communication Review, Vol. 24, No.4, (ACM SIGCOMM’94) Oct. 1994, pp.234-244.

[12] G. Pei, M. Gerla and T.-W. Chen, Fisheye State Routing in Mobile Ad Hoc Networks. In Proceedings of the 2000 ICDCS Workshops, Taipei, Taiwan, Apr. 2000, pp. D71-D78.

[13] C.E. Perkins and E.M. Royer. Ad hoc on demand Distance Vector routing, mobile computing systems and applications, 1999. Proceedings. WMCSA ’99. Second IEEE Workshop on, 1999, p90 - p100.

[14] D. Johnson, D. A. Maltz, Dynamic source routing in ad hoc wireless networks, in Mobile Computing (T. Imielinski and H. Korth, eds.), Kluwer Acad. Publ., 1996.

[15] Z. J. Haas. The Zone Routing Protocol (ZRP) for ad hoc networks, Internet Draft, Nov. 1997.

[16] S. Corson, and J. Macker, Mobile Ad hoc Networking: “Routing Protocol Performance Issues and Evaluation Consideration” rfc 2501, Jan

[17] L. R. Ford Jr. and D. R. Fulkerson, Flows in Networks, Princeton Univ.Press 1962

[18] V. D. Park and M. S. Corson. A highly adaptive distributed routing algorithm for mobile wireless networks, INFOCOM '97, Sixteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Driving the Information Revolution, Proceedings IEEE, Volume: 3, 1997 Page(s): 1405 - 1413 vol.3.

[19] V. Park, and S. Corson, Temporally-Ordered Routing Algorithm (TORA) Version 1 Functional Specification. IETF Internet draft, 1997.

[20] M. Joa-Ng and I-Tai Lu, A peer-to-peer zone-based two-level link state routing for mobile ad hoc net-works, IEEE on Selected Areas in Communications, vol. 17, no. 8, pp. 1415 1425, 1999.

[21] Navid Nikaein, Christian Bonnet and Neda Nikaein, HARP - Hybrid Ad Hoc Routing Protocol, in proceeding of IST 2001: International Symposium on Telecommunications, Iran/Tehran 2001.

## Author Biographies

**First Author** Shailender Gupta is B. Tech (Electronics) and M. Tech (Computer Engg.) and pursuing his Ph. D in the area of ad hoc mobile network security. His academic interest includes network security, automata theory and fuzzy logic.

Shailender Gupta
Lecturer (Electronics Engg.)
YMCA instt. of Engg. Faridabad
E-mail: shailender81@gmail.com

**Second Author** T Chander Kumar is M. Tech (Computer Engg.) and Ph. D (Computer Science). His academic interest includes network security, software reliability and artificial intelligence.

Dr. C.K.Nagpal
Asstt. Prof.(Computer Engg.)
YMCA instt. of Engg. Faridabad
E-mail: nagpalckumar@rediffmail.com

# A Comparative Study and Analysis of Power Factor Control Techniques

Sanjay L. Kurkute, Pradeep M. Patil, Vinod H. Patil

**Abstract: -** Power factor correction (PFC) is the capacity of absorbing the reactive power produced by a load. The major industrial loads have an inductive power factor. The current tends to go beyond the power is usually used for the power conversion. This paper presents an active input power factor correction with single phase boost converter topology using various control techniques. A comparative study of several analog and digital power factor control techniques is studied. This investigation is to identify a low cost, efficient and reliable PF control technique. Digital implementation by using Microcontroller and DSP achieves more reliability. For Digital Signal Processor based PFC technique power factor is above 0.99 and very close to unity.

***Keywords:*** Power Factor Correction, Digital Signal Processor, PF Control Technique,

## 1. Introduction

In recent years, the power quality of the AC system has become a great concern due to the rapidly increased numbers of electronic equipment, power electronics and high voltage power system. Now passive PFC schemes are implemented & used by customer for commercial & industrial applications.

Drawbacks of Passive PF Method:

3 Large size of reactive elements.
3 Power factor improvement for a narrow operating region.
3 Large output dc voltage ripple.

Active high frequency power factor correction makes the load behave like a resistor, leading to near unity load power factor and the load generating negligible harmonics for variable loads.

To compensate for the higher reactive power demand by the converters at high power transfer levels, power factor correction becomes mandatory. This is also consistent with the goals of switch mode conversion. A variety of topologies can be used including the boost converter and the buck converter. For reasons of relative simplicity and popularity, the boost converter is described here.

A diode rectifier effects the ac/dc conversion, while the controller operates the switch in such a way to properly shape the input current ig according to its reference shown in Fig.1. The output capacitor absorbs the input power pulsation, allowing a small ripple of the output voltage VL. The boost topology is very simple and allows low-distorted input currents and almost unity power factor with different control techniques. Moreover, the output capacitor is an efficient energy storage element due to the high output voltage value and the ground-connected switch simplifies the drive circuit [1-2].

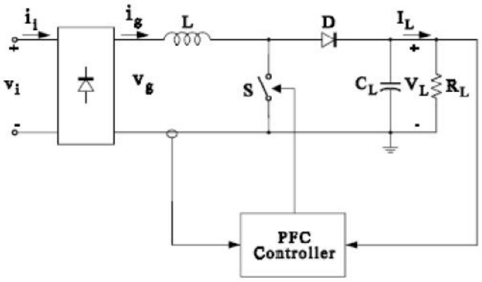


*Fig.1 – Boost Converter with active PFC*

The paper is organized as follows: Section II provides seven different PF control techniques. Section III describes the analysis of all the control techniques. Section IV presents some conclusion along with future issues that need to be addressed.

## 2. PF Control Techniques

### A) Continuous Current Mode Control:

### 1. INDUCTOR CURRENT CONTROL

The switching pre-regulator circuit of Fig. 2 is a high frequency boost converter. The output voltage of the pre-regulator can be transformed via conventional switched-mode methods to generate low voltage dc outputs. There are twocontrollers in the pre-regulator circuit. These are the voltage and current-loop

---------------------

The research is supported in part by Electronics Engineering -ACTEL BVDU Lab, Bharati Vidyapeeth Deemed University College of Engineering, Pune.

Sanjay L. Kurkute, Professor and Head, Electronics Engineering Department, Bharati Vidyapeeth Deemed University College of Engineering, Pune-43 (India),
e-mail: slkurkute@bvucoep.edu.in, kurkutesanjay@yahoo.co.in

Pradeep M. Patil, Professor and Head, Electronics Engineering Department, Vishwakarma Institute of Technology,Pune-37(India).e-mail: patil_pm@rediffmail.com

Vinod H. Patil, Lecturer Electronics Engineering Department, Bharati Vidyapeeth Deemed University College of Engineering, Pune-43(India), e-mail:vinodpatil93@gmail.com

controllers. The voltage-loop controller regulates the output voltage around the desired nominal value while the current-loop controller shapes the inductor current into a rectified sinusoid in phase with the input ac voltage. Thus, near unity power factor can be achieved.

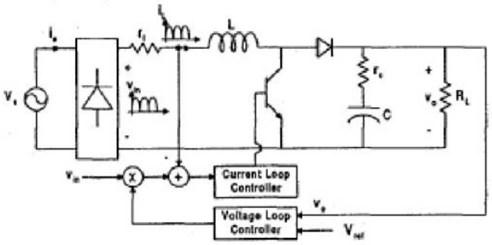


*Fig.2 - Boost Converter with switching pre-regulator under current mode control technique.*

The controller specifies a peak switch current in each cycle, or a peak inductor current, rather than the duty cycle. The switch is turned on at the beginning of the switching cycle, and is turned off when its current reaches a specified upper threshold value, im. This threshold value is the primary control variable and the duty ratio becomes an indirectly determined auxiliary variable. This method has lead to an inductor current that approximates a rectified sinusoid in phase with the input voltage. It can yield power factors in the range of 0.95 to 0.99, which reduces the total harmonic distortion of the source voltage amplitude, permits the use of a smaller capacitor.

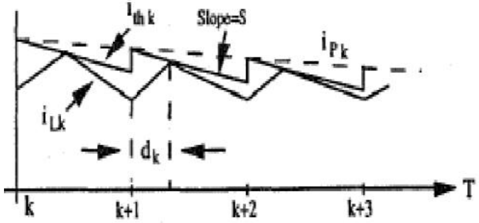


*Fig. 3: Inductor current under CMC*

Fig. 3 shows the relationship between the threshold current and the inductor current. The threshold current, ith, is determined from the sum of two signals: a slowly varying signal, ip, determined by the voltage controller on the basis of the discrepancy between the reference and output voltages, and a regular saw-tooth ramp of slope-S at the switching frequency.

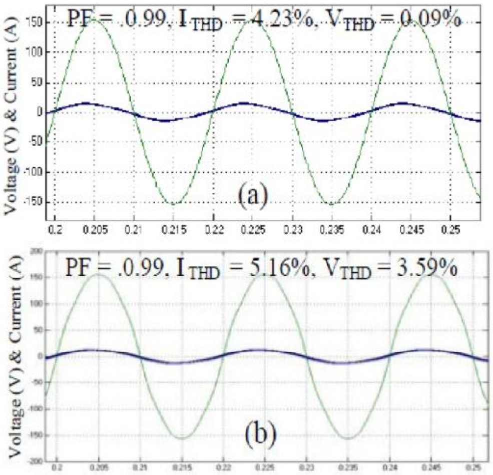


*Fig.4 - Simulation Results; Performance of Boost converter using CMC*

## 2. PEAK CURRENT CONTROL

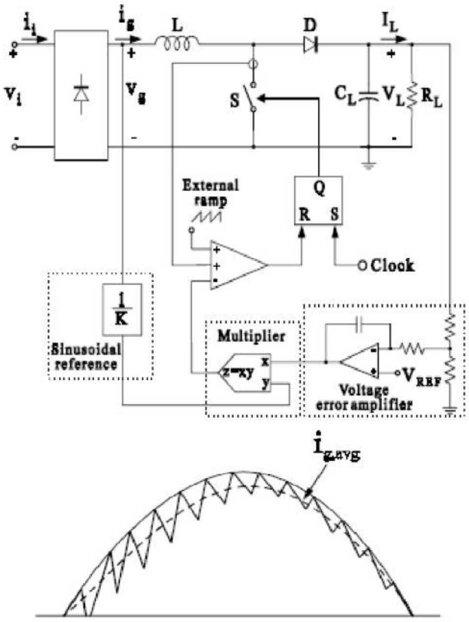


*Fig.5 - Peak current control technique*

The switch is turned on at constant frequency by a clock signal, and is turned off when the sum of the positive ramp of the inductor current or switch current and a compensating ramp reaches the sinusoidal current reference. This reference is usually obtained by multiplying a scaled replica of the rectified line voltage vg times the output of the voltage error amplifier, which sets the current reference amplitude. In this way, the reference signal is naturally synchronized and always proportional to the line voltage, which is the condition to obtain unity power factor. As Fig.5 reveals, the converter operates in Continuous

Inductor Current Mode (CICM); this means that devices current stress as well as input filter requirements are reduced. Moreover, with continuous input current, the diodes of the bridge can be slow devices (they operate at line frequency). On the other hand, the hard turn-off of the freewheeling diode increases losses and switching noise, calling for a fast device [3].

***Advantages*:**

- Constant switching frequency;
- Only the switch current must be sensed and this can be accomplished by a current transformer, thus avoiding the losses due to the sensing resistor;
- No need of current error amplifier and its compensation network;
- Possibility of a true switch current limiting.

***Disadvantages*:**

- Presence of sub harmonic oscillations at duty cycles greater than 50%, so a compensation ramp is needed;
- input current distortion which increases at high line voltages and light load and is worsened by the presence of the compensation ramp;
- Control more sensitive to commutation noises.

## 3. AVERAGE CURRENT CONTROL

It allows a better input current waveform, is the average current control represented in Fig.6. Here the inductor current is sensed and filtered by a current error amplifier whose output drives a PWM modulator. In this way the inner current loop tends to minimize the error between the average input current ig and its reference. This latter is obtained in the same way as in the peak current control. The converter works in CICM, so the same considerations done with regard to the peak current control can be applied [4].

***Advantages*:**

- Constant switching frequency;
- No need of compensation ramp;
- Control is less sensitive to commutation noises, due to current filtering;
- Better input current waveforms than for the peak current control since, near the zero crossing of the line voltage, the duty cycle is close to one, so reducing the dead angle in the input current.

***Disadvantages*:**

- Inductor current must be sensed;
- A current error amplifier is needed and its compensation network design must take into account the different converter operating points during the line cycle.

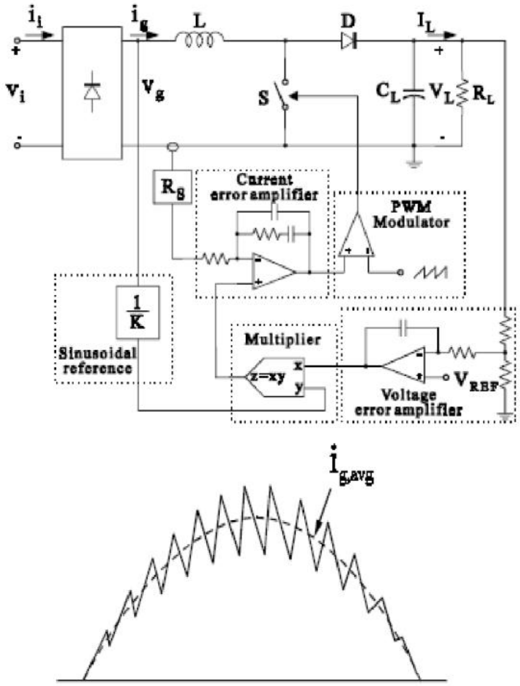


*Fig.6 - Average current control technique*

### B) Discontinuous current PWM control:

With this approach, the internal current loop is completely eliminated, so that the switch is operated at constant on-time and frequency .As shown in Fig.7, with the converter working in discontinuous conduction mode (DCM), this control technique allows unity power factor when used with converter topologies like flyback, Cuk and Sepic. Instead, with the boost PFC this technique causes some harmonic distortion in the line current.

***Advantages*:**

- Constant switching frequency;
- No need of current sensing;
- Simple PWM control;

***Disadvantages*:**

- Higher devices current stress than for borderline control;
- Input current distortion with boost topology.

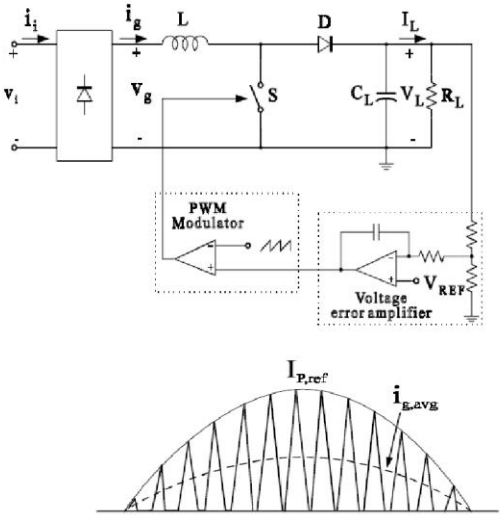


*Fig.7 - Discontinuous current PWM control technique*

### C) Digital Signal Processor Control

DSP controllers provide many distinctive advantages over Traditional analog control, viz:

1. Standard control hardware design for multiple platforms.
2. Less susceptibility to aging and environmental variations.
3. Better noise immunity.
4. Ease of implementation of sophisticated control algorithms.
5. Flexible design modifications to meet a specific customer need.

Fig. 8 shows circuit diagram of PFC using digital technique a typical power-factor-corrected rectifier based on a boost converter. Current loop is designed so that the converter input current follows the waveform of the input voltage. In the ideal case these two waveforms have the same wave shape and are in phase, thus the rectifier presents a resistive load to the system. The outer loop regulates the voltage across the energy-storage capacitor. This voltage always has ripple at twice the line frequency 2 L. To maintain low input current harmonics output of the voltage regulator u(t) must not have significant line frequency harmonics. Consequently, to avoid distortion of the ac line current through feedback, the capacitor voltage ripple in conventional designs the bandwidth of the voltage loop is limited to frequencies significantly lower than the line frequency (typically 10- 20Hz).

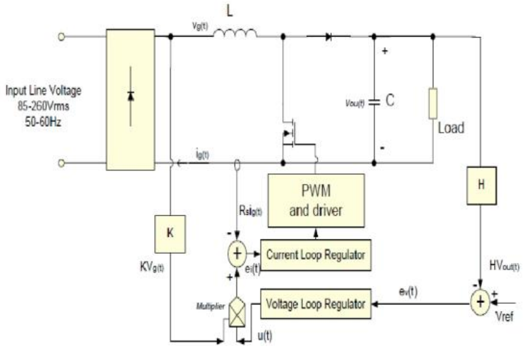


*Fig.8 - Digital Signal Processor Control technique*

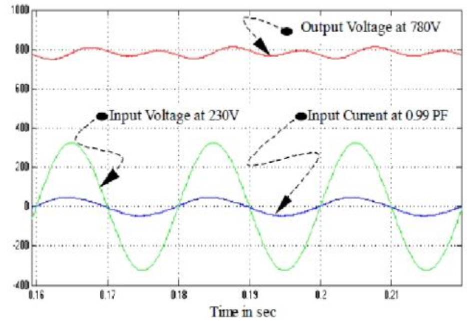


*Fig.9 - Simulation Results; Performance of Boost converter using DSP Control*

To increase systems reliability, it is proposed to implement converter controls in digital domain. Digital implementation also enhances systems programmability and reliability by removing few drawbacks of analog implementation such as parts count, ageing and environment effects and limited flexibility [5-6].The recently available high-speed digital signal processor executes controller algorithm faster and enhances converter-switching frequency to 20 kHz and higher. The control algorithm written in high-level language provides ease and flexibility. The digital implementation reduces number of components, increases reliability and hence attractive for UPS application.

## 3. Comparative Analysis

Table: 1 shows comparison of analog and digital control techniques. Simulation results of current mode and DSP control techniques are shown in Fig. 4 and Fig. 9.

*Table: 1 Comparison of analog and digital control techniques for PFC.*

| Control Techniques | Features | Power Factor |
|---|---|---|
| **Continuous current** | **Inductor/Peak current control used** | **0.95-0.99** |
| **Discontinuous current** | **Simple PWM control** | **PF close to unity** |
| **Digital Signal Processor** | **Flexible design modification and easy implementation** | **above 0.99 to very close to unity** |

## 4. Result and Conclusion

In this paper a comprehensive summary of several control techniques are studied and analyzed for PFC boost converter. The analog and digital control techniques are compared for variable load. Current mode controls have presence of sub harmonic oscillations and more sensitive to commutation noises. A DSP based control technique also enhances systems programmability and reliability and it is most suitable latest technique for achieve very close unity power factor than others.

Three phase boost converter PFC is one of the important future issue for various load demand for industrial applications.

## IV. References

[1] Sangsun Kim and Dr. P. Enjeti," Digital Control of Switching Power Supply - Power Factor Correction Stage" International Conference on Smart Manufacturing Application April. 9-11, 2008 in KINTEX, Gyeonggi-do, Korea.

[2] A. Muthuramalingam and S. Himavathi, "Evaluation of Power Factor Corrected AC – DC Converters and Controllers to meet UPS Performance Index" International Journal of Electronics, Circuits and Systems 3:1 2009.

[3] Gui-Jia Su, IEEE Senior Member, Donald J. Adams, "Comparative Study of Power Factor Correction Converters For Single Phase Half-Bridge Inverters" PESC 2001.

[4] David M. Van de Sype, Koen De Gussemé, Alex P. Van den Bossche, IEEE, and Jan A. A. Melkebeek, "A Sampling Algorithm for Digitally Controlled Boost PFC Converters" IEEE TRANSACTIONS ON POWER ELECTRONICS, VOL.19,NO.3,MAY2004.

[5] Feel-Soon Kang, Sung-Jun Park, Member, IEEE, and Cheul-U Kim, Member, IEEE, "ZVZCS Single-Stage PFC AC-to-DC Half-Bridge Converter" IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS, VOL. 49, NO. 1, FEBRUARY 2002.

[6] Supratim Basu and Math H. J. Bollen, Fellow, IEEE," A Novel Common Power Factor Correction Scheme for Homes And Offices" IEEE TRANSACTIONS ON POWER DELIVERY, VOL. 20, NO. 3, JULY 2005.

[7] P. M. Patil and S. L. Kurkute, "Speed control of Three
Phase Induction Motor using Single Phase Supply along with Active power Factr Correction," in ICGST International Journal on Automatic Control and Systems Engineering (ACSE), vol. VI, Issue: III, , October 2006,
pp. 23–32.

[8] P.M. Patil, D.N. Kyatanavar, R.G. Zope and D.V. Jadhav, "Three-phase ac drive using single phase supply", Journal of The Institution of Engineers (India), Vol. 82, June2001, pp 43-47.

[9] P.M. Patil, J.V. Kulkarni and D. B. Kshirsagar, "A Noble firing scheme for three-phase controllers", In Proceedings of International Conference on Computer Applications in Electrical Engineering RecentvAdvances (CERA01) held at IIT Roorkee, February 2002, pp 412-417.

[10] P.M. Patil and S.L. Kurkute,"Programmable active power factor improvement technique for single phase switch mode boost rectifiers", in proceedings of IEEE International Conference on Industrial Technology ICIT 2006 held at IIT Bombay, December 2006, pp 2119-2124.

[11] S.L. Kurkute and P.M. Patil ,"Single phase power factor correction for high frequency buck-boost switching converters" in proceedings of Pravara International Conference on Emerging Trends in Engineering (PICETE 2008) held at PREC, Loni. pp 112-113 on 20-22 Dec. 2008.

[12] Sanjay L Kurkute, Pradeep M Patil and K.C. Mohite, "A Digital Power Factor Correction using Floating Point Processor for Pulse Width Modulation Control in Boost Converters" International Journal of Electronic Engineering Research Volume 1 Number 2 (2009) pp. 135–146.

**AUTHORS BIBLOGRAPHY**

### Sanjay L. Kurkute

**He is born in Loni, Dist Ahmednagar, Maharashtra, India on December 10, 1973. He received B.E.(Electronics) degree in 1997 under Pune University and M.Tech in Power Electronics degree under Visveshwaraiah Technological University, Belgaum in Jan 2001.**

**Presently he is doing the Research/ PhD in Power Electronics. He is working as Professor and Head in the Department of Electronics Engineering at Bharati Vidyapeeth University College of Engineering, Pune-43. Also he was Principal Investigator for Research Project (2008, 2009) under BCUD, University of Pune. His work has been published in an International Journal ICGST on Automatic Control and System Engineering (ACSE), International Journal of Electronic Engineering Research (IJEER), International IEEE papers and various National Conferences.**

**Dr. Pradeep M. Patil**

He is born in Bhusawal, District Jalgaon, Maharashtra, India on December 13,1966. He received B. E. (Electronics) degree in 1988 from Amravati University, Amravati, India and M. E. (Electronics) degree in 1992 from Marathwada University, Aurangabad, India. From 1988 to 2002. He received the Ph.D. degree in Electronics and Computer Engineering in 2004 at Swami Ramanand Teerth Marathwada University. He is member of various professional bodies like IE, ISTE, IEEE and Fellow of IETE. Presently he is working as Professor and Head of Electronics Department at Vishwakarma Institute of Technology, Pune, India. His research areas include pattern recognition, fuzzy neural networks and power electronics. His work has been published in various international and national journals and conferences including IEEE.

**Vinod H. Patil**

He was born in Tambave; district Sangli, Maharashtra, India on June 05, 1987.

He received his BE degree in Electronics and telecommunication with distinction from Shivaji University, Kolhapur, India in the year 2008. He is appearing for M.Tech (Electronics- VLSI) in Bharati Vidyapeeth Deemed University, Pune, India. He is currently working as an lecturer in Bharati Vidyapeeth deemed University College of Engineering, Pune, India. His research interests are Power Electronics, VLSI and Wireless communication.

# Impulsive stabilization of two kinds of third-order delay differential equations

Hongxing Yao[1], Lili Qi[2] and Hong Pan[3]

[123]Nonlinear Scientific Research Center, Jiangsu University, Zhenjiang 212013, China
[1]Department of Applied Mathematics, University of Waterloo, Waterloo,Canada N2L3G1
hxyao@ujs. edu.cn, square04@126.com

***Abstract*:** These instructions provide you guidelines for preparing papers for International Journal. Use this document as a template and as an instruction set. This paper is concerned with the impulsive stabilization problems for two kinds of 3th-order delay differential equations. By the method of Lyapunov function, we prove that the non-impulsive equations can be stabilized by the proper impulse control. Our results has improved and extended some results. We also give examples to illustrate the efficiency of our results.

***Keywords*:**Third-order differential equation; Impulsive stabilization; Delay; Lyapunov function

## 1. Introduction

When you submit your paper print it in two-column format, including figures and tables. In addition, designate one author as the "corresponding author". This is the author to whom proofs of the paper will be sent. Proofs are sent to the corresponding author only. Third-order differential equation; Impulsive stabilization; Delay; Lyapunov function Recently, the problem of impulsive stabilization for differential equations has attracted many authors' attentions and some results have been published (see [1-10]). Impulses can make unstable systems stable. The problem of stabilizing the solutions by imposing proper impulse control has been used in many fields such as physics, pharmacokinetics, biotechnology, economics, chemical technology. However some authors have researched the impulsive stabilization problems for two kinds of 2th-order delay differential equations in [1-5], they proved that it also can be made exponentially continuous with respect to initial data by impulses on some interval $t_k$. And the presented references here ([7,8,10]) dealt with mostly the first-order delay differential equations (see[7]), In this paper , we consider third-order delay differential equations and deal with more general equations, the results we prove here generalize recent ones by Li and Weng [1]. This paper is about third-order delay differential equations and deal with more general equations. We also establish sufficient conditions for the stability of solutions by imposing proper impulse control.

This paper is organized as follows. In Section 2, we establish third-order delay deferential equations. In Section 3, by using Lyapunov function and analysis methods, we prove that the non-impulsive equations can be stabilized by the proper impulse control. In Section 4, two examples are discussed to illustrate the efficiency of the main results.

## 2. Preliminaries

We consider the following two equations with impulses:

$$\begin{cases} x'''(t)+c(t)x''(t)+b(t)x'(t)+a(t)x(t-\tau)=0, \quad t\ge t_0; \quad t\ne t_k, \quad k=1,2\cdots \\ x(t)=\varphi(t), \quad t_0-\tau\le t\le t_0; \quad x'(t_0)=y_0, \quad x''(t_0)=z_0 \\ x(t_k)=I_k(x(t_k^-)), \quad x'(t_k)=J_k(x'(t_k^-)), \quad x''(t_k)=U_k(x''(t_k^-)) \end{cases} \tag{1}$$

and

$$\begin{cases} x'''(t)+c(t)x''(t)+b(t)x'(t)+\int_{t-\tau}^{t}e(t-u)x(u)du=0, \quad t\ge t_0; \quad t\ne t_k, \quad k=1,2\cdots \\ x(t)=\varphi(t), \quad t_0-\tau\le t\le t_0; \quad x'(t_0)=y_0, \quad x''(t_0)=z_0 \\ x(t_k)=I_k(x(t_k^-)), \quad x'(t_k)=J_k(x'(t_k^-)), \quad x''(t_k)=U_k(x''(t_k^-)) \end{cases} \tag{2}$$

With the following assumptions:

$(H_1)$ $\tau>0,\ x(t):[t_0-\tau,+\infty]\to R$;

$(H_2)$ $x''(t),x'(t)$ denotes the right derivative of $x'(t),\ x(t)$, and $x''(t)=[x'(t)]'$, $x'(t)=\lim\limits_{h\to 0}\dfrac{x(t+h)-x(t)}{h}$. If $x(t)$ is piecewise continuous, then $x(s^-)$ and $x(s^+)$ denote, respectively, its left and right limits as $t$ tend to $s$;

$(H_3)$ $\varphi:[t_0-\tau,t_0]\to R$ has at most finite discontinuity points of the first kind and is right continuous at these points;

$(H_4)$ $a(t),b(t),c(t)$ are continuous on $[t_0,+\infty]$, $e(t)$ is continuous on $[0,\tau]$;

$(H_5)$ $0<t_1<t_2<\cdots<t_k<t_{k+1}<\cdots,\quad \lim\limits_{k\to\infty}t_k=+\infty$, with $\tau\le t_{k+1}-t_k\le l$, $t\in N$;

$(H_6)$ consider the impulses at times $t_k$, $k=1,2\cdots$

$$x(t_k)=I_k(x(t_k^-)),\ x'(t_k)=J_k(x'(t_k^-)),\ x''(t_k)=U_k(x''(t_k^-)$$

Where $I_k,J_k,U_K:R\to R$ are continuous and $I_k(0)=J_k(0)=U_k(0)=0,\ k\in N$;

The following definitions are slightly modified from [1]:

**Definition 2.1** A function $x:[t_0-\tau,t_0+\tau)\to R,\ \alpha>0$ is a solution of equation (1)(or(2)), though $(t_0,\varphi,y_0,z_0)$, if

(i) $x(t), x'(t), x''(t)$ are continuous on $\left[t_0, t_0+\alpha\right)\backslash\{t_k, k\in Z\}$;

(ii)

$$x(t)=\varphi(t), t\in\left[t_0-\tau, t_0\right], \quad x'(t_0)=y_0, x''(t_0)=z_0,$$

$x(t)$ satisfies the first equality of (1)(or(2) on $\left[t_0, t_0+\alpha\right)\backslash\{t_k, k\in Z\}$;

(iii) $x(t), x'(t), x''(t)$ all have two-side limits and right continuous at point $t_k$ , and $x(t_k), x'(t_k), x''(t_k)$ satisfy the third equality of (1)(or(2));

**Definition 2.2** The problem of equation (1)(or(2)) is said to be exponentially stabilized by impulses, if there exist $\alpha>0$, a sequence $\{t_k\}_{k\in N}$ satifying $[H_5]$, and sequences of continuous functions $\{I_k\}, \{J_k\}, \{U_k\}$. such that for all $\varepsilon>0$, there exists $\delta>0$, such that if a solution $x(t;t_0,\varphi,y_0,z_0)$ of (1)(or(2)) fulfills:

$$\sqrt{\|\varphi\|^2+y_0^2+z_0^2} \le \delta$$

Then

$$\sqrt{x^2(t)+x'^2(t)+x''^2(t)} \le \varepsilon\exp\left[-\alpha(t-t_0)\right],$$

$t\ge t_0$ ,

where $\|\varphi(s)\| = \sup\limits_{t_0-\tau\le s\le t_0}|\varphi(s)|$

## 3 Main Results

First we consider system (1)

**Theorem 3.1.** If there exist $A\ge 0, B\ge 0, C\ge 0$, such that $|a(t)|\le A,\ |b(t)|\le B,\ |c(t)|\le C,$ and

$$A\tau\le\exp\{-2(1+A+B+C)\tau\}, \tag{3}$$

The solution of system(1)can be exponentially stabilized by impulses.

**Proof.** By (3) there exist $\alpha>0$ and $l\ge\tau$ such that

$$A\tau\le\exp\left[-2\alpha(l+\tau)\right]\exp\left[-2(1+A+B+C)l\right]. \tag{4}$$

Let $\alpha, l$ be as in (4). every sequence $\{t_k\}_{k\in N}$ satisfying $(H_5)$ with $t_k-t_{k-1}\le l,\ k\in N,$ let

$$I_k(u)=J_k(u)=U_k(u)=d_ku. \quad k=1,2\cdots$$

$$d_k=\sqrt{\frac{P_k-A\tau}{2}}$$

$$p_k=\exp\left[-2\alpha(t_{k+1}-t_k+\tau)\right]$$
$$\exp\left[-2(1+A+B+C)(t_{k+1}-t_k)\right]$$

It is easy to verify that $d_k\le 1,$ and

$$d_k^2+A\tau=\frac{p_k+A\tau}{2}\le p_k .$$

For every $\varepsilon>0$, let

$$\delta=\frac{\varepsilon}{\sqrt{1+A\tau}}\exp\left[-\alpha(t_1-t_0)\right]$$
$$\exp\left[-(1+A+B+C)(t_1-t_0)\right]$$

We will prove that for each solution $x(t;t_0,\varphi,y_0,z_0)$ of (1), such that

$$\sqrt{\|\varphi\|^2+y_0^2+z_0^2}\le\delta$$

we have

$$\sqrt{x(t)+x'^2(t)+x''^2(t)}\le\varepsilon\exp\left[-\alpha(t-t_0)\right]$$

$t\ge t_0$

If $t\in\left[t_{k-1}, t_k\right), k\in N$, consider the Lyapunov functional

$$V(t)=x^2(t)+x'^2(t)+x''^2(t)+\int_{t-\tau}^{t}|a(s+\tau)|x^2(s)ds$$

and $V(t)$ satisfies:

(i) $V(t)\ge x^2(t)+x'^2(t)+x''^2(t);$

(ii) $V(t)\le x^2(t)+x'^2(t)+x''^2(t)+\|x\|_t^2\int_t^{t+\tau}|a(s)|ds$

$$\le x^2(t)+x'^2(t)+x''^2(t)+A\tau\|x\|_t^2$$
$$\le(1+A\tau)\left[\|x\|_t^2+x'^2(t)+x''^2(t)\right]$$ where

$$\|x\|_t = \sup_{t-\tau\le s\le t}|x(s)|;$$

(iii) $V'(t)=2x(t)x'(t)+2x'(t)x''(t)+2x''(t)x'''(t)$

$$+|a(t+\tau)|x^2(t)-|a(t)|x^2(t-\tau)$$
$$=2x(t)x'(t)+2x'(t)x''(t)$$
$$+2x''(t)\left[-c(t)x''(t)-b(t)x'(t)-a(t)x(t-\tau)\right]$$
$$+|a(t+\tau)|x^2(t)-|a(t)x^2(t-\tau)|$$
$$\le 2x(t)x'(t)+2x'(t)x''(t)+2c(t)x''^2(t)$$
$$+2b(t)x'(t)x''(t)+2a(t)\ x(t-\tau)x''(t)$$
$$+|a(t+\tau)|x^2(t)-|a(t)x^2(t-\tau)|$$
$$\le\left[x^2(t)+x'^2(t)\right]+\left[x'^2(t)+x''^2(t)\right]$$
$$+2Cx''^2(t)+B\left[x'^2(t)+x''^2(t)\right]$$
$$+A\left[x^2(t-\tau)+x''^2(t)\right]+Ax^2(t)$$
$$\le\left[2(1+A+B+C)\right]\left[x^2(t)+x'^2(t)+x''^2(t)\right]$$
$$\le\left[2(1+A+B+C)\right]V(t)$$

Solving $V'(t)\le\left[2(1+A+B+C)\right]V(t)$ , we obtain

$$V(t)\le V(t_0)\exp\left[2(1+A+B+C)(t-t_0)\right]$$

$(1^0)$ if $t\in\left[t_0, t_1\right)$, Integrating the above inequality from $t_0$ to $t_1$, we obtain

$$x^2(t)+x'^2(t)+x''^2(t)\le V(t)$$
$$\le V(t_0)\exp\left[2(1+A+B+C)(t_1-t_0)\right]$$
$$\le(1+A\tau)\left[\|x\|_{t_0}^2+x'^2(t_0)+x''^2(t_0)\right]$$
$$\exp\left[2(1+A+B+C)(t_1-t_0)\right]$$

$$\le (1+A\tau)\delta^2\exp[2(1+A+B+C)(t_1-t_0)]$$
$$\le \varepsilon^2\exp[-2\alpha(t_1-t_0)]$$
$$\le \varepsilon^2\exp[-2\alpha(t-t_0)]$$

Therefore

$$\sqrt{x^2(t)+x'^2(t)+x''^2(t)}\le\varepsilon\exp[-\alpha(t-t_0)],\qquad t\in[t_0,t_1)$$

Especially

$$\sup_{t_1-\tau\le t\le t_1}\left[x^2(t)+x'^2(t)+x''^2(t)\right]\le\varepsilon^2\exp[-2\alpha(t_1-t_0-\tau)]$$

$(2^0)$ If $t\in[t_1,t_2)$. by the expression of $V(t)$, we have

$$x^2(t)+x'^2(t)+x''^2(t)\le V(t)$$
$$\le V(t_1^+)\exp[2(1+A+B+C)(t-t_1)]$$

$$\le V(t_1^+)\exp[2(1+A+B+C)(t_2-t_1)]$$
$$\exp[2(1+A+B+C)(t_2-t_1)]$$

$$=\left\{x^2(t_1)+x'^2(t_1)+x''^2(t_1)+\int_{t-\tau}^{t}|a(s+\tau)|x^2(s)ds\right\}$$
$$\exp[2(1+A+B+C)(t_2-t_1)]$$
$$=\left\{\left[I_1(x^2(t_1^-))+J_1(x'^2(t_1^-))+U_1(x''^2(t_1^-))\right]\right.$$
$$\left.+\int_{t-\tau}^{t}|a(s+\tau)|x^2(s)\ \ ds\ \right\}$$
$$\exp[2(1+A+B+C)(t_2-t_1)]$$

$$=\left\{d_1^2\left[x^2(t_1^-)+x'^2(t_1^-)+x''^2(t_1^-)\right]+\int_{t-\tau}^{t}|a(s+\tau)|x^2(s)ds\right\}$$
$$\exp[2(1+A+B+C)(t_2-t_1)]$$
$$\le d_1^2\sup_{t_1-\tau\le t\le t_1}\left[x^2(t)+x'^2(t)+x''^2(t)\right]$$
$$\exp[2(1+A+B+C)(t_2-t_1)]$$

$$+\sup_{t_1-\tau\le t\le t_1}x^2(t)A\tau\exp[2(1+A+B+C)(t_2-t_1)]$$

$$\le(d_1^2+A\tau)\sup_{t_1-\tau\le t\le t_1}\left[x^2(t)+x'^2(t)+x''^2(t)\right]$$
$$\exp[2(1+A+B+C)(t_2-t_1)]$$
$$\le(d_1^2+A\tau)\varepsilon^2\exp[-2\alpha(t_1-t_0-\tau)]$$
$$\exp[2(1+A+B+C)(t_2-t_1)]$$

By the definitions of $d_1$ and $p_1$, we have

$$x^2(t)+x'^2(t)+x''^2(t)\le V(t)$$
$$\le\varepsilon^2(d_1^2+A\tau)\exp[-2\alpha(t_1-t_0-\tau)]$$
$$\exp[2(1+A+B+C)(t_2-t_1)]$$
$$=\varepsilon^2\left(\frac{p_1+A\tau}{2}\right)\exp[-2\alpha(t_1-t_0-\tau)]$$
$$\exp[2(1+A+B+C)(t_2-t_1)]$$
$$\le\varepsilon^2p_1\exp[-2\alpha(t_1-t_0-\tau)]$$
$$\exp[2(1+A+B+C)(t_2-t_1)]$$
$$=\varepsilon^2\exp[-2\alpha(t_2-t_0)]$$
$$\le\varepsilon^2\exp[-2\alpha(t-t_0)]$$

We obtain for $t\in[t_1,t_2)$,

$$\sqrt{x^2(t)+x'^2(t)+x''^2(t)}\le\varepsilon\exp[-\alpha(t-t_0)]$$

$(3^0)$ With analogous arguments,we can verify that for all $k\in N,\ t\in[t_{k-1},t_k),\ k=1,2\cdots$, we have

$$\sqrt{x^2(t)+x'^2(t)+x''^2(t)}\le\varepsilon\exp[-\alpha(t-t_0)]$$

Hence

$$\sqrt{x^2(t)+x'^2(t)+x''^2(t)}\le\varepsilon\exp[-\alpha(t-t_0)],$$

$t\ge t_0$,

The proof is complete.

Now we prove that the problem (2) can be exponentially stabilized by impulses.

**Theorem3.2** If there exist $E\ge0$, such that $|e(t)|\le E$, and

$$\frac{1}{2}E\tau^2<\exp\{-2(1+A+B+E\tau)\tau\},\qquad(5)$$

The solution of system (2) can be exponentially stabilized by impulses.

**Proof.** By (5), there exist $\alpha>0$ and $l>\tau$, such that

$$\frac{1}{2}E\tau^2<\exp[-2\alpha(l+\tau)]\exp[-2(1+A+B+E\tau)l]\qquad(6)$$

Let $\alpha,l$ be as in (6). every sequence $\{t_k\}_{k\in N}$ satisfying $(H_5)$ with $t_k-t_{k-1}\le l,\ k\in N,$ let

$$I_k(u)=J_k(u)=U_k(u)=d_ku.\quad k=1,2\cdots$$

$$d_k=\sqrt{\frac{P_k-\frac{1}{2}E\tau^2}{2}}$$

$$p_k=\exp[-2\alpha(t_{k+1}-t_k+\tau)]$$

$$\exp[-2(1+A+B+E\tau)(t_{k+1}-t_k)]$$

It is easy to verify that $d_k\le1,$ and

$$d_k^2+\frac{1}{2}E\tau^2=\frac{p_k+\frac{1}{2}E\tau^2}{2}\le p_k.$$

For every $\varepsilon>0$, let

$$\delta=\frac{\varepsilon}{\sqrt{1+\frac{1}{2}E\tau^2}}\exp[-\alpha(t_1-t_0)]$$
$$\exp[-(1+A+B+E\tau)(t_1-t_0)]$$

We will prove that for each solution $x(t;t_0,\varphi,y_0,z_0)$ of (1), such that

$$\sqrt{\|\varphi\|^2 + y_0^2 + z_0^2} \le \delta$$

we have

$$\sqrt{x(t) + x'^2(t) + x''^2(t)} \le \varepsilon \exp\left[-\alpha(t-t_0)\right],$$

$$t \ge t_0 ,$$

If $t \in [t_{k-1}, t_k),\ k \in N$ , consider the Lyapunov functional

$$V(t) = x^2(t) + x'^2(t) + x''^2(t)$$

$$+ \int_{t-\tau}^{t} \left[ \int_{u}^{t} |e(u-s+\tau)| x^2(s) ds \right] \mathrm{du}$$

and $V(t)$ satisfies

(i) $V(t) \ge x^2(t) + x'^2(t) + x''^2(t);$

(ii) $V(t) \le x^2(t) + x'^2(t) + x''^2(t) + \|x\|_t^2 \int_{t-\tau}^{t} \int_0^{\tau} |e(s)| ds du$

$$\le x^2(t) + x'^2(t) + x''^2(t) + \frac{1}{2} E\tau^2 \|x\|_t^2$$

$$\le (1 + \frac{1}{2} E\tau^2) \left[ \|x\|_t^2 + x'^2(t) + x''^2(t) \right]$$

where $\|x\|_t = \sup_{t-\tau \le s \le t} |x(s)|;$

(iii) $V'(t) = 2x(t)x'(t) + 2x'(t)x''(t) + 2x''(t)x'''(t)$

$$+ \int_{t-\tau}^{t} |e(u-t+\tau)| x^2(t) du - \int_{t-\tau}^{t} |e(t-s)| x^2(s) ds$$

$$= 2x(t)x'(t) + 2x'(t)x''(t) + 2x''(t)$$

$$\left[ -c(t)x''(t) - b(t)x'(t) - a(t)x(t-\tau) \right]$$

$$+ \int_{t-\tau}^{t} |e(u-t+\tau)| x^2(t) du - \int_{t-\tau}^{t} |e(t-s)| x^2(s) ds$$

$$\le 2x(t)x'(t) + 2x'(t)x''(t) + 2c(t)x''^2(t)$$

$$+ 2b(t)x'(t)x''(t) + 2\int_{t-\tau}^{t} e(t-u)x(u)x''(t) du$$

$$+ \int_{t-\tau}^{t} |e(u-t+\tau)| x^2(t) du - \int_{t-\tau}^{t} |e(t-s)| x^2(s) ds$$

$$\le \left[ x^2(t) + x'^2(t) \right] + \left[ x'^2(t) + x''^2(t) \right] + 2Cx''^2(t) + B\left[ x'^2(t) + x''^2(t) \right]$$

$$+ E\tau \left[ x^2(u) + x^2(t) \right] + E\tau x^2(t)$$

$$\le \left[ 2(1 + A + B + E\tau) \right] \left[ x^2(t) + x'^2(t) + x''^2(t) \right]$$

$$\le \left[ 2(1 + A + B + E\tau) \right] V(t)$$

Solving $V'(t) \le \left[ 2(1 + A + B + E\tau) \right] V(t)$ , we obtain

$$V(t) \le V(t_0) \exp\left[ 2(1 + A + B + E\tau)(t - t_0) \right]$$

$(1^0)$ if $t \in [t_0, t_1)$, Integrating the above inequality from $t_0$ to $t_1$, we obtain

$$x^2(t) + x'^2(t) + x''^2(t) \le V(t)$$

$$\le V(t_0) \exp\left[ 2(1 + A + B + E\tau)(t_1 - t_0) \right]$$

$$\le (1 + \frac{1}{2} E\tau^2) \left[ \|x\|_{t_0}^2 + x'^2(t_0) + x''^2(t_0) \right] \exp\left[ 2(1 + A + B + E\tau)(t_1 - t_0) \right]$$

$$\le (1 + \frac{1}{2} E\tau) \delta^2 \exp\left[ 2(1 + A + B + E\tau)(t_1 - t_0) \right]$$

$$\le \varepsilon^2 \exp\left[ -2\alpha(t_1 - t_0) \right]$$

$$\le \varepsilon^2 \exp\left[ -2\alpha(t - t_0) \right]$$

Therefore

$$\sqrt{x^2(t) + x'^2(t) + x''^2(t)} \le \varepsilon \exp\left[ -\alpha(t - t_0) \right], \quad t \in [t_0, t_1)$$

Especially

$$\sup_{t_1 - \tau \le t \le t_1} \left[ x^2(t) + x'^2(t) + x''^2(t) \right] \le \varepsilon^2 \exp\left[ -2\alpha(t_1 - t_0 - \tau) \right]$$

$(2^0)$ If $t \in [t_1, t_2)$. by the expression of $V(t)$, we have

$$x^2(t) + x'^2(t) + x''^2(t) \le V(t)$$

$$\le V(t_1^+) \exp\left[ 2(1 + A + B + E\tau)(t - t_1) \right]$$

$$\le V(t_1^+) \exp\left[ 2(1 + A + B + E\tau)(t_2 - t_1) \right]$$

$$= \left\{ x^2(t_1^+) + x'^2(t_1^+) + x''^2(t_1^+) + \int_{t-\tau}^{t} \left[ \int_{u}^{t} |e(u-s+\tau)| x^2(s) ds \right] du \right\}$$

$$\exp\left[ 2(1 + A + B + E\tau)(t_2 - t_1) \right]$$

$$= \left\{ x^2(t_1) + x'^2(t_1) + x''^2(t_1) + \int_{t-\tau}^{t} \left[ \int_{u}^{t} |e(u-s+\tau)| x^2(s) ds \right] du \right\}$$

$$\exp\left[ 2(1 + A + B + E\tau)(t_2 - t_1) \right]$$

$$=$$

$$\left\{ \left[ I_1(x^2(t_1^-)) + J_1(x'^2(t_1^-)) + U_1(x''^2(t_1^-)) \right] + \int_{t-\tau}^{t} \left[ \int_{u}^{t} |e(u-s+\tau)| x^2(s) ds \right] du \right\}$$

$$\exp\left[ 2(1 + A + B + E\tau)(t_2 - t_1) \right]$$

$$= \left\{ d_1{}^2 \left[ x^2(t_1^-) + x'^2(t_1^-) + x''^2(t_1^-) \right] + \int_{t-\tau}^{t} \left[ \int_{u}^{t} |e(u-s+\tau)| x^2(s) ds \right] du \right\}$$

$$\exp\left[ 2(1 + A + B + E\tau)(t_2 - t_1) \right]$$

$$\le d_1^2 \sup_{t_1 - \tau \le t \le t_1} \left[ x^2(t) + x'^2(t) + x''^2(t) \right]$$

$$\exp\left[ 2(1 + A + B + E\tau)(t_2 - t_1) \right]$$

$$+ \sup_{t_1 - \tau \le t \le t_1} x^2(t) \frac{1}{2} E\tau^2 \exp\left[ 2(1 + A + B + E\tau)(t_2 - t_1) \right]$$

$$\le (d_1^2 + \frac{1}{2} E\tau^2) \sup_{t_1 - \tau \le t \le t_1} \left[ x^2(t) + x'^2(t) + x''^2(t) \right]$$

$$\exp\left[ 2(1 + A + B + E\tau)(t_2 - t_1) \right]$$

$$\le (d_1^2 + \frac{1}{2} E\tau^2) \varepsilon^2 \exp\left[ -2\alpha(t_1 - t_0 - \tau) \right]$$

$$\exp\left[ 2(1 + A + B + E\tau)(t_2 - t_1) \right]$$

By the definitions of $d_1$ and $p_1$ , we have

$$x^2(t) + x'^2(t) + x''^2(t) \le V(t)$$

$$\le \varepsilon^2 (d_1^2 + \frac{1}{2} E\tau^2) \exp\left[ -2\alpha(t_1 - t_0 - \tau) \right]$$

$$\exp\left[2(1+A+B+E\tau)(t_2-t_1)\right]$$
$$=\varepsilon^2\left(\frac{p_1+\frac{1}{2}E\tau^2}{2}\right)\exp\left[-2\alpha(t_1-t_0-\tau)\right]$$
$$\exp\left[2(1+A+B+E\tau)(t_2-t_1)\right]$$
$$\le\varepsilon^2 p_1\exp\left[-2\alpha(t_1-t_0-\tau)\right]$$
$$\exp\left[2(1+A+B+E\tau)(t_2-t_1)\right]$$
$$=\varepsilon^2\exp\left[-2\alpha(t_2-t_0)\right]$$
$$\le\varepsilon^2\exp\left[-2\alpha(t-t_0)\right]$$

We obtain for $t\in\left[t_1,t_2\right)$,

$$\sqrt{x^2(t)+x'^2(t)+x''^2(t)}\le\varepsilon\exp\left[-\alpha(t-t_0)\right]$$

$\left(3^0\right)$ With analogous arguments,we can verify that for all $k\in N,\ t\in\left[t_{k-1},t_k\right),\ k=1,2\cdots$, we have

$$\sqrt{x^2(t)+x'^2(t)+x''^2(t)}\le\varepsilon\exp\left[-\alpha(t-t_0)\right]$$

Hence

$$\sqrt{x^2(t)+x'^2(t)+x''^2(t)}\le\varepsilon\exp\left[-\alpha(t-t_0)\right],$$

$t\ge t_0$,

The proof is complete.

## 4 Examples

**Example 4.1**. Consider the following equation:

$$\begin{cases} x'''(t)+0.33x''(t)-0.025x'(t)-0.5x(t)-x(t-0.01)=0, & t\ge 0\\ x(t)=\varphi(t), \quad -0.01\le t\le 0; & x'(0)=y_0, \quad x''(0)=z_0. \end{cases} \tag{7}$$

whose characteristic equation is

$$\lambda^3+0.33\lambda^2-0.025\lambda-0.5-e^{-0.01\lambda}=0$$

By Mathematica sofeware, we find a characteristic root of (7) with the positive real part. Hence the non-impulsive system (7) is unstable.

Consider

$A=1,\quad l=\tau=0.01,\quad \alpha=1/2,\quad B=C=0.5$, and we can verify that

$$A\tau\le\exp\left[-2\alpha(l+\tau)\right]\exp\left[-2(1+A+B+C)l\right]$$
$$<\exp\left[-2(1+A+B+C)\tau\right].$$

Considering the impulses at $t_k$, such that $t_k-t_{k-1}\equiv 0.01$ and

$$x(t_k)=I_k(x(t_k^-))=dx(t_k^-),\ \ x'(t_k)=J_k(x'(t_k^-)),$$
$$=dx'(t_k^-),\ \ x''(t_k)=U_k(x''(t_k^-))=dx''(t_k^-)$$

where $d=\sqrt{\dfrac{\exp(-0.08)-0.01}{2}}$, By Theorem 3.1 the unstable system (7) can be exponentially stabilized by impulses.

**Example 4.2**. Consider the following equation:

$$\begin{cases} x'''(t)-0.75x''(t)-x(t-0.0375)=0, & t\ge 0\\ x(t)=\varphi(t), \quad -0.0375\le t\le 0; & x'(0)=y_0, \quad x''(0)=z_0. \end{cases} \tag{8}$$

whose characteristic equation is

$$\lambda^3-0.75\lambda^2-e^{-0.0375\lambda}=0$$

By Mathematica sofeware, we find a characteristic root of (8) with the positive real part. Hence the non-impulsive system (8) is unstable.

Consider

$A=1,\quad l=\tau=0.0375,\quad \alpha=1/2,\quad C=0.75$, and we can verify that

$$A\tau\le\exp\left[-2\alpha(l+\tau)\right]\exp\left[-2(1+A+B+C)l\right]$$
$$<\exp\left[-2(1+A+B+C)\tau\right].$$

Considering the impulses at $t_k$, such that $t_k-t_{k-1}\equiv 0.0375$ and

$$x(t_k)=I_k(x(t_k^-))=dx(t_k^-),$$
$$x'(t_k)=J_k(x'(t_k^-))=dx'(t_k^-),$$
$$x''(t_k)=U_k(x''(t_k^-))=dx''(t_k^-),$$

where $d=\sqrt{\dfrac{\exp(-0.28125)-0.0375}{2}}$, By Theorem 3.1 the unstable system (8) can be exponentially stabilized by impulses.

## References

[1] A.Weng,J.Sun, Impulsive stabilization of second-order nonlinear delay differential systems,Appl.Math.Comput.214 (2009)95-101

[2] L.P. Gimenes, M. Federson, Existence and impulsive stability for second order retarded differential equations, Appl. Math.Comput.177 (2006)44–62.

[3] Xiang Li, Peixuan Weng, Impulsive stabilization of two kinds of second-order linear delay differential equations, J. Math. Anal. Appl. 291 (2004) 270–281.

[4] A.Weng,J.Sun, Impulsive stabilization of second-order delay differential equations, Nonlinear Anal,:Real Word Appl.8(2007) 1401-1420

[5] L.P. Gimenes, M. Federson, Impulsive stability for systems of second order retarded differential equations, Nonlinear Anal. 67 (2007) 545–553.

[6] Xinzhi Liu, George Ballinger, Existence and continuability of solutions for differential equations with delays and state-dependent impulses, Nonlinear Anal. 51 (2002) 633–647.

[7] L. Berezansky, E. Braverman, Impulsive stabilization of linear delay differential equations, Dynam. Systems,Appl. 5 (1996) 263–276.

[8] W. Feng, Y. Chen, The weak exponential asymptotic stability of impulsive differential system, Appl. Math. J. Chinese Univ. 1 (2002) 1–6.

[9] J. Shen, Z. Luo, X. Liu, Impulsive stabilization of functional differential equations via Liapunov functionals, J. Math. Anal.Appl.240 (1999)1-5

[10] X. Li, Impulsive stabilization of linear differential system, J. South China Normal Univ. Natur. Sci. Ed. 1(2002) 52–56.

# IBSEAD: - A Self-Evolving Self-Obsessed Learning Algorithm for Machine Learning

Jitesh Dundas[1] and David Chik [2]

[1]Scientist, Edencore Technologies,
Row House – 6, Opp Ambo Vihar, Tirupati Nagar-II, Off Unitech Road, Virar(w),
Maharashtra, Thane-401303, India
Email: - jbdundas@gmail.com

[2]Senior Scientist, Riken Brain Institute,
Dept of Robotics and Neuroscience, Riken Institute, Japan

***Abstract***: We present IBSEAD or distributed autonomous entity systems based Interaction - a learning algorithm for the computer to self-evolve in a self-obsessed manner. This learning algorithm will present the computer to look at the internal and external environment in series of independent entities, which will interact with each other, with and/or without knowledge of the computer's brain. When a learning algorithm interacts, it does so by detecting and understanding the entities in the human algorithm. However, the problem with this approach is that the algorithm does not consider the interaction of the third party or unknown entities, which may be interacting with each other. These unknown entities in their interaction with the non-computer entities make an effect in the environment that influences the information and the behaviour of the computer brain. Such details and the ability to process the dynamic and unsettling nature of these interactions are absent in the current learning algorithm such as the decision tree learning algorithm. IBSEAD is able to evaluate and consider such algorithms and thus give us a better accuracy in simulation of the highly evolved nature of the human brain. Processes such as dreams, imagination and novelty, that exist in humans are not fully simulated by the existing learning algorithms. Also, Hidden Markov models (HMM) are useful in finding "hidden" entities, which may be known or unknown. However, this model fails to consider the case of unknown entities which maybe unclear or unknown. IBSEAD is better because it considers three types of entities-known, unknown and invisible. We present our case with a comparison of existing algorithms in known environments and cases and present the results of the experiments using dry run of the simulated runs of the existing machine learning algorithms versus IBSEAD.

***Keywords***: Self-evolving algorithm; machine learning; decision-trees; learning algorithms, Hidden Markov Models

## 1. Introduction

One of the fundamental problems in AI is the capability of the robots to learn on their own. The manner in which learning is done by robots, will decide the actions that are taken by the same. The goal of machine learning is the ability of the machines to learn and interpret information like humans. Over the past decades, we made great progress in moving towards this goal. However, there are still issues in providing the accuracy in understanding and interpretation of the knowledge by the machines. We present here the learning algorithms that have till date, made a lot of impact in the field of artificial intelligence. However, these algorithms are falling short of providing learning capabilities (of the human level) to the robots**.**

We present IBSEAD - a learning algorithm that will allow the robots to learn, at a higher level, with humans. We then compare the existing learning algorithms and measure if IBSEAD scores better in complex situations and interactions, with the same efficiency as a normal human being.

## 2. Assumptions

The paper has the following assumptions:-

1) We believe that the computer brain is composed of the visual system, detection system and the CPU (Central Processing Unit) system that will process the information. The computer is a simulated example of the human being with the computer brain being similar to the human brain.

2) We call IBSEAD self-obsessed because it is concerned with its own interaction and wishes to improve its own survival rate. This algorithm tries to do what is best for itself, simulating what a normal human being tries to do in his/her life. Every action that is performed is a result of its manifestation of self-interests and self-centered perception of the environment in which the CB exists.

3) The environment is here divided into:-

a) The internal environment that is made up of the entities present in the computer.

b) The external environment is the environment that is made up of the entities present outside the CB or computer brain. This is the region where unknown entities are expected to be present the most.

c) The invisible entities are the entities which are not seen/visible/detected by the CB but still have an effect on the actions/decisions and perceptions of the CB directly or indirectly. These entities are in existence but are just invisible or a not directly available.

d) The unknown entities are the entities that have an effect on the system but their existence or any information about them is still unknown. For e.g. the distant galaxies are unknown to us but they do impact us when a space vessel travels in space for investigation. We do not have any information about them but their effect on the scenario is well accepted. The presence of such entities ensures that the risk estimation and the unknown reactions are taken care of.

e) Those entities that are detected and understood by the CB are called as known entities. Invisible entities are not visible but are understood by the CB. Unknown entities are neither visible not known but their absence is rules out.

## 3. IBSEAD Algorithm

Despite the recent advances in machine learning, the higher modes of human learning techniques still elude us in robotics. One of the most important reasons is that the failures on the hardware side are not properly handled by the robot in its learning process.

Secondly, the learning techniques do not consider the group based environments in which the measurements are taken for different states of each of the group entities and then a measurement of the needed trait taken. For e.g. we know that as the entities in the environment are arranged in groups, and are changing dynamically in several modes, each of these groups has an individual measurement and thus it has to be aggregated and averaged out, to get an average impact of the group's effect on the interaction with the environment. Similarly, all the groups in the scope of the observation scene have their own measurements. The CB is interacting with each of the entity groups and this complexity is not measured properly by the decision tree based learning methods.

Another point worth noting is that although an entity may be present in the scope of the CB, it may not be interacting with the same. Again, the interaction between the entities and the CB may be intended, unwanted or hostile. These interactions are not measured properly by the existing methods.

Thirdly, the unknown entities have an impact on the learning capabilities of the CB. These indirect entities are interacting directly with the CB or indirectly via the entities of the CB observation scenario. There are indirect effects of the actions of these unknown entities which are not recorded by the existing learning methods. The CB may not be aware fully, of the existing functionalities and impact, of the interactions of the unknown entities. Some of the existing methods do not have any provisions for such complex functionalities and thus are not able to higher levels of human learning capabilities.

All the deficiencies in the existing methods give a strong reason for the creation of a new algorithm that will deliver on such issues. IBSEAD is an effort in this direction.

The algorithm has the following steps:-

1) Scan through the problems and find all the entities within its physical scope

2) Scan and also consider the entities not in physical scope. Classify them as known, invisible or hidden and unknown entities.

3) Map the entities into groups, single or non-single entity, based on understanding of their group dynamics.

4) For each group, find their impact and track their connections to the CB.

5) For those conditions where the switch is yes in both the entities, the interaction is executed and learning started.

Please note that some of the steps have been removed to ensure the confidential nature of the current projects on this algorithm. The important steps have been shown here with the differences in the current algorithms like decision trees. Figure-1) explains the steps in detail with focus on the final picture as it will look in the learning process.

## 4. Background

A lot of work has been done on the learning algorithms in artificial intelligence. Decision-tree based learning techniques organize the entities of the environment, into tree like structures, so as to facilitate the flow of information between each of them. There are several algorithms that have helped in making machines learn and evolve.

Learning is roughly classified into supervised and unsupervised learning. .Fisher proposed the first learning algorithm for pattern recognition. Hidden Markov models [19] proposed the use of hidden states of entities to consider such scenarios but could not explain further regarding the different attributes of the entities and the interaction conditions involved. Moreover, there was a need to explain the quality of communication in the same. There is a need to quantify intangible entities which is missed by Hidden Markov Models. IBSEAD is a step in this direction. Hidden Markov models (HMM) are useful in finding "hidden" entities, which may be known or unknown. However, this model fails to consider the case of unknown entities which maybe unclear or unknown. Also, IBSEAD is considers three types of entities- known, unknown and invisible while HMM considers the hidden and known entities only. Boltzmann [20] machine based equations also misses out on such similar issues and is known to be very theoretical in nature. Bayesian statistics depends [22] on the ability to measure the correctness of a hypothesis. However, it is clear that the absence of information of any entity will make it difficult to present a hypothesis of it. However, IBSEAD takes the use of interaction of the surrounding entities, along with the environment, internal or external, in which the unknown entity is most expected to be present, as key parameters. Bayesian based algorithms seem to miss out on the other three features of IBSEAD, which play an important role in accurate learning algorithms. Case based [23] reasoning and Inductive Logic Programming [24] requires past experience of the scenarios in order to learn about the present. However, this can be time consuming and prone to higher error rates as unknown entities may not be simple and their interaction random. IBSEAD handles this situation better as it considers unknown entities and the presence of unknown entities is considered beforehand and no unwanted scenarios are expected.

One of the serious problems in Gaussian process based algorithms is that the values will give incorrect answers in the case of dependent entities and dependent interactions [25]. Consider the case of two entities A and B, where the interaction of QC (A->B) is influencing the interaction of QC (B<-->CB). Clearly, there is an issue in which the above Gaussian process based algorithms will give inaccurate values. Moreover, the points Xi are needed to give us values of the desired result dataset, in which we assume that the points Xi will always give us correct values. However, if the behavior of the entity changes and the points thus plot wrong values (or even changes are seen) then we find that the obtained values are very wrong. Also, this algorithm expects prior knowledge of the Gaussian functions for correct estimation. Thus, if the unknown entities are not known, then their effects are difficult to measure. This method is limited only till the "Hidden" or "Invisible" entities as per the complex scenario used by the IBSEAD algorithm in this paper. Group method of data handling [26, 27] (GMDH) is very good application for polynomial based multilayered neural network based algorithms. Again, we miss out the unknown entities and the cases where fuzzy or no information is available.

All the above algorithms miss out on the quality of communication and the switch needed for allowing the communication.

## 5. Methodology

We studied the methods present in machine learning for scenarios that involved complex human interactions. We then presented our algorithm IBSEAD and then measure the performance with other existing algorithms on the scenarios presented below. Finally we implemented our algorithm in a simulation environment and deduced conclusions from the same.

Please note the following scenarios:-

1) Optimizing stock market gain: - In most of the times, existing algorithms will tell us specific formulae that seem to be very static in their consideration. Certain parameters are hard-coded into the scenario and then the equation is executed. However, in a stock market, the value of the share price depends on several known and unknown entities. Several algorithms can tell us how a company share price is performing based on the known entities such as market price, share price trend, company accounts, etc. However, there are several entities that are not considered. Some of these include insider trading, environmental conditions that may affect the region, natural and artificial calamities, the sudden death of the promoters or feud between them, gossip, influence of negative people, etc. Such entities are not considered in any of the learning algorithms and thus fail to deliver the accuracy and impact needed. IBSEAD takes care of this problem as it covers such invisible entities (we call these as invisible as they do not seem to be detected directly but do have an impact on the resulting interaction) and thus will deliver a much higher and better accuracy on the same. Again, we see that each of the invisible entities will interact directly or indirectly with the computer brain (assuming that the computer is doing the trading on the market). Again, each of the entity’s interaction will be possible only when the switch of each of the entity (which decides whether to interact with the other entity or entities or not. If this interaction not present between the entities in consideration, then this means that one or both of the entities are having this switch as No. This state can be due to ignorance, presence of blockage agents like noise or even just perception, individual decision, etc). Such a complex environment cannot be learnt with the existing algorithms. IBSEAD answers many of the complexities mentioned above and thus surely gives a higher accuracy and better risk management of the stock market scenario.

2) Go Game Problem: - In the Go game problem, each player is expected to use intuition besides other skills to be able to understand and make winning moves against the opponent. However, the go game requires observation as well as if possible, the capability to understand the opponent too. The existing learning algorithms do not implement the presence of essential entities such as opponent behaviour, intuition, etc and thus may not give the expected results efficiently. IBSEAD considers the coverage of such entities and interactions and thus gives better results too. For e.g.) IBSEAD will consider opponent behaviour also as anger or tension of the opponent may give insights into the mental state and thus the expected performance level of the opponent.

3) Moving Trains & the underlying complexities: - The environment in which the train travelled from City A to City B was rainy. Thus the train reached late and also some of its engine parts (even the rails on the path) were rusted. Now such third party interactions – from the past and present, affect the decision of the CB of travelling by the train. The CB might never know of such detail but these interactions between the unknown entities (rails) and the external known entities (the train) exists and has an affect on the CB’s existence. Such details are considered by IBSEAD and thus account for better results than decision tree based and other types of algorithms.

4) Visual Recognition: - Consider the case wherein we have 3 objects: dog, cat and table. The training set has 20 images

each. While the test set has 10 images each. We now compare standard neural networks VS IBSEAD in the above scenario. We know that IBSEAD considers invisible entities as well and thus “NOISE” is also an entity here. The computer brain entity may not be aware of the entity creating the noise but the noise does reach the computer. Thus, it becomes an entity itself in this case (though it may be a different case wherein the entity may be visible and noise will be a distraction or blockage of interaction. Still IBSEAD considers better coverage (by 20-30 %) of entities and state of their being in such cases while neural networks don’t do so). Also, IBSEAD helps in gaining higher levels of understanding such as concentration and ignorance. Standard neural networks are found to be 40-50% correct while IBSEAD were found to be 70-80% accurate. The reason is that in standard neural networks, information lost as "noise" whereas in IBSEAD, "noise" is considered as unknown entity.

5) Loans Risk Assessment: - We collected the datasets (simulated versions) in the format as prescribed as in the paper by Xavier et al. The existing dataset had factors including Income, Advance EMI, Rent, Qualifications, Dependents, Experience. The paper claims 98% accuracy. Hidden layers are shown but they don't consider the quality of data, availability or intangible or invisible entities as ibsead does. We now consider IBSEAD for the same problem. We modified it to include parameters such as influence of customer in the bank, corruption, business feasibility, regulatory environment, etc. The final modified dataset had 20% new cases of extremely volatile kind that could cause issues. We got the following results:-

5.1) Coverage :- We considered hidden entities (and unknown entities) like black money income, power/ influence on loan process, viability of business , trustworthiness of this loan for the customer, economic conditions of the market, bank solvency, future trends, etc .
5.2) Quality of communication: - Some of the details obtained maybe crooked or forged. Is the client ready to give his consent to the communication? Do we need to verify case in background from other banks/institutions/people, etc? These are some of the factors considered.
5.3) Switch:-A switch field for each attribute (0-10) to tell if the values are valid or not is missing. What if the entities or attributes aren't giving the information e.g. sensitive information about business? Ignorance or hiding details causes switch to become NO.
5.3) Software errors/human errors/corruption/natural calamities are to be factored here.
5. 4) Pattern search does not reveal corruption or future trends or manager intuition & trust. However, these are considered in IBSEAD while keeping a track of patterns in loans.
Addition of these causes the Neural Network to give reduced 60-65% accuracy in the modified dataset. IBSEAD gives more accuracy & thus 90% accuracy was obtained.

## 6. Existence of Multiple Concurrent Connections between Entities in the scenario

We define a connection as an interaction between two steps (or entities). Say in a decision tree, A and B are two steps, with A being above and B being below. How can we consider that A will always interact with B? There are several issues that need investigation:-
The connections may be stopped because of ignorance. We will call the consent and openness of each of the entities (i.e. A and B) to be very necessary to be able to pursue the interaction or communication between the entities. Some of the agents of such blockages or interrupts are noise, darkness or ignorance. Each of these conditions, if present in the concerned entity or entities, can create issues in the interaction. Obstacles in the path of connection between the entities are a source of concern or blockage for the scenario. It is possible that the blockage may be intentional or unintentional, beneficial or harmful. The value of interaction between two entities A and B will be positive only when the switch between the two entities is set to true. This is like the AND condition based interaction (Figure -2) switching wherein the interaction is allowed only when all cases are true. Thus, in this case, if more than 2 entities are concurrently involved, then all the entities should have the switch set to true to allow interaction. One interaction at a time is what the brain can handle to give optimum performance. The decision-based algorithms fail to handle these conditions. There is a need to consider focus and concentration also in the learning algorithms to be able to handle complex scenarios such as chess and Go game. This is missing in existing algorithms such as decision tree based algorithms, neural networks, etc. They consider the states to be static in such complex environment whereas the IBSEAD algorithm considers this as dynamic. The decision tree based algorithms consider one assumption: - They always believe that all the entities are connected to each other. We know that the human brain is the best entity at learning and most of the algorithms have basis with it. However, the human brain cannot handle more than one connection at a time. How can we assume that all the connections will be active and also connected to each other, just because they are in the scope of the learning environment of the computer brain?
Consider a scenario where a person is sitting in a train. He is then watching the scenario, looking at the buildings when he finds a train coming in the opposite track. The user is surprised by this entity’s presence. If we consider the Decision Tree based algorithms, then there is no way that this knowledge based connection and the train as an entity would be considered. Moreover, there is no provision of a switch which will tell if the user or train is interacting with each other. There is absence of a condition for checking states such as ignorance, blockages to interactions like darkness, miscommunication, etc.
Another major issue in this is the handling of the context of the scenario in order to achieve the meaning and the intended observation.

## 7. Advantages

This algorithm takes into account the non-visible entities that do affect the interactions and learning process of the robot.

1) Decision tree based systems do not account for scenarios where the entities may not be interacting in a tree like fashion. The tree based structure is invalid when the interactions at the second and lower levels come into picture. What if there are interactions without any such sub-levels.

2) The decision tree algorithms do not consider horizontal and backward interactions, something which is so common and essential in any learning process. IBSEAD fills the gap in this direction.

3) IBSEAD gives a more comprehensive and accurate picture than its predecessors.

4) IBSEAD can answer the problems in adding consciousness and awareness in robots, something which current algorithms fail to add.

5) This program considers entities as individuals and not as groups or sub-systems (with common goals), which seems to be the case with most of the living and dynamic environment entities. In a scenario (in which the robot is supposed to learn about walking into a railway train), it has to interact with people, some in group while some walk alone. Some of the entities may be even trains. Such a scenario may involve unknown (or invisible) entities that cannot be seen by the robot. The robot can only feel its effect. For e.g. here it considers the rainfall and the supervisors who control the route to the train as invisible entities (or unknown entities). Such complex scenarios are not given by decision tree algorithms nor do any of the existing algorithms give the accuracy as IBSEAD.

6) IBSEAD is relatively complete, easy to use and deeply, compared to a hierarchical structure based decision trees.

7) The ability of the algorithms to implement higher levels of human consciousness and learning are also not convincing. IBSEAD is a positive step in this direction.

## 8. Conclusion

We have found that IBSEAD has a better performance and accuracy in learning of robots, when compared to existing methods such as decision-tree based learning methods, in certain scenarios.

IBSEAD accounts for invisible entities and their interaction and effects, something which existing algorithms fail to deliver. There is a switch to ensure that the entities are ready to communicate (flag set to "Y" is set). There is a better coverage of entities and the other deeper details of the learning process and communication, something which existing algorithms fail to deliver.

## 9. Future Scope

We wish to propose that IBSEAD be used to handle complex situations that are novel and not falling as per the "learn from existing entities and knowledge" type of situations. In cases where no past experience is available, IBSEAD performance might get slowed down. We wish to pursue this in the future scope of this algorithm.

## 10. Financial Interests

There was no clash of interest found in the above research work.

## Acknowledgement

Many thanks to Prof Uma Srinivasan who helped in pursuing this work. This work was not supported by any organization and has no clashes of interest.

## References

[1] Support Vector Networks. Cortes C. & Vapnik V. Machine learning, 1995 – Springer. Online:- http://www.springerlink.com/content/k238jx04hm87j80g/

[2] A weighted nearest neighbour algorithm for learning with symbolic features. Cost et al. Machine Learning Volume 10, Number 1, 57-78, DOI: 10.1023/A:1022664626993 .Online:- www.springerlink.com/index/r21k03320q127784.pdf

[3] Learning Information Extraction Rules for semi structured and free text. Soderland S. Machine Learning .Volume 34, Issue 1-3 (February 1999). Special issue on natural language learning. Pages: 233 – 272. 1999 ISSN: 0885-6125.

[4] Very Simple Classification Rules Perform Well on Most Commonly Used Datasets. Robert C. Holte. Machine Learning Volume 11, Number 1, 63-90, DOI: 10.1023/A:1022631118932

[5] Machine Learning for Information Extraction in Informal Domains. Dayne Freitag. Machine Learning. Volume 39, Numbers 2-3, 169-202, DOI: 10.1023/A:1007601113994

[6] Self-Improving Reactive Agents Based On Reinforcement Learning, Planning and Teaching. Long-Ji Lin. Machine Learning, 8, 293-321 (1992) © 1992
Kluwer Academic Publishers, Boston. Manufactured in The Netherlands.

[7] MML Inference of Decision Graphs with Multi-Way Joins and Dynamic Attributes. Tan P and Dowe D. Online: - http://www.csse.monash.edu.au/~dld/Publications/2003/Tan+Dowe2003_MMLDecisionGraphs.pdf as on 23Aug 2010.

[8] Supervised clustering using decision trees and decision graphs: An ecological comparison. M.B. Dalea, P.E.R.

Dalea, and P. Tan B. Ecological Modeling. Volume 204, Issues 1-2, 24 May 2007, Pages 70-78

[9] Instance-Based Learning Algorithms. AVID W. AHA et al. Machine Learning, 6, 37-66 (1991) © 1991 Kluwer Academic Publishers, Boston. Manufactured in The Netherlands. Online:- http://www.springerlink.com/content/g4qv6511520x3041/fulltext.pdf

[10] Learning logical definitions from relations. J. R. Quinlan. Machine Learning. Volume 5, Number 3, 239-266, DOI: 10.1007/BF00117105

[11] AI 2003: advances in artificial intelligence: 16th Australian Conference on AI, Perth, Australia, December 3-5, 2003: proceedings .Tamás D. Gedeon, Lance Chun Che Fung. Online:- http://books.google.co.in/books?id=4ClFWeqtNAC&lpg=PA270&ots=zMyau1MoZf&dq=7)%09MML%20Inference%20of%20Decision%20Graphs%20with%20MultiWay%20Joins%20and%20Dynamic%20Attributes.&pg=PA270#v=onepage&q=7)%09MML%20Inference%20of%20Decision%20Graphs%20with%20Multi-Way%20Joins%20and%20Dynamic%20Attributes.&f=false

[12] Unsupervised Learning: Foundations of Neural Computation--A Review. DL Wang - 2001. Online: - www.aaai.org/ojs/index.php/aimagazine/article/download/1565/1464.

[13] Comparing supervised and unsupervised category learning. BRADLEY C. LOVE. Psychonomic Bulletin & Review 2002, 9 (4), 829-835. Online: - http://pbr.psychonomic-journals.org/content/9/4/829.short.

[14] Supervised Machine Learning: A Review of Classification Techniques. S. B. Kotsiantis. Informatica .31 (2007) 249-268 249. Online:- http://www.informatica.si/PDF/31-3/11_Kotsiantis%20-%20Supervised%20Machine%20Learning%20-%20A%20Review%20of...pdf

[15] http://en.wikipedia.org/wiki/Category:Classification_algorithms

[16] http://en.wikipedia.org/wiki/List_of_machine_learning_algorithms

[17] Learning in the presence of drifts and hidden concepts. Widmer et al. Machine Learning.Vol-23, Issue-1, (April 1996), Pages: 69 - 101.1996.ISSN:0885-6125

[18] Xavier et al. improving prediction accuracy of loan default- A case in rural credit. Online: - http://www.ifmr.ac.in/pdfs/Improving prediction accuracy. PDF

[19] Ephraim Y, Merhav N (June 2002). "Hidden Markov processes". *IEEE Trans. InformTheory* 48: 1518–1569. doi: 10.1109/TIT.2002.1003838.

[20] Ackley, D. H.; Hinton, G. E.; Sejnowski, T. J. (1985). "A Learning Algorithm for Boltzmann Machines". *Cognitive Science* 9: 147–169. doi: 10.1207/s15516709cog0901_7. http://learning.cs.toronto.edu/~hinton/absps/cogscibm.pdf.

[21] Bretthorst, G. Larry, 1988, *Bayesian Spectrum Analysis and Parameter Estimation* in Lecture Notes in Statistics, 48, Springer-Verlag, New York, New York.

[22] David MacKay (2003). Information Theory, Inference, and Learning Algorithms. Cambridge University Press.

[23] Aamodt, Agnar, and Enric Plaza. "Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches" Artificial Intelligence Communications 7, no. 1 (1994): 39-52.

[24] S.H. Muggleton and L. De Raedt. Inductive logic programming: Theory and methods. Journal of Logic Programming, 19, 20:629-679, 1994.

[25] Williams, Christopher K.I. (1998). "Prediction with Gaussian processes: From linear regression to linear prediction and beyond". In M. I. Jordan. Learning in graphical models. MIT Press. pp. 599–612.

[26] A.G. Ivakhnenko. Heuristic Self-Organization in Problems of Engineering Cybernetics. Automatica 6: pp.207–219, 1970.

[27] H.R. Madala, A.G. Ivakhnenko. Inductive Learning Algorithms for Complex Systems Modeling. CRC Press, Boca Raton, 1994.

## Author Biographies

**Mr. Jitesh B. Dundas** is working as a Software Engineer in a reputed IT MNC in Mumbai. He is also a Scientist with Edencore Technologies (www.edencore.net). He has completed his Masters in Computer Applications from Pune University in 2007. He has found the concept of the “Law of Connectivity in Machine Learning” – a paper which is published in IJSST in Dec 2010 as well as a co-author of this IBSEAD algorithm paper (http://en.wikipedia.org/wiki/IBSEAD). More details on his research work and contact details can be found on his research homepage at http://openwetware.org/wiki/Jitesh_Dundas_Lab .

**Dr. David Chik** is a Senior Scientist at Riken Institue, Japan. He received is PhD from the University of Hong Kong and has held several reputed positions in top universities and research institutions in the past. He is a genius with several reputed publications in his name, with this IBSEAD paper being just of them. His further details can be found at his homepage at one at http://www.brain.riken.jp/common/cv/d_chik.pdf

# Using H-Algorithm to Find the Study of Multi-Server Wireless Multicast System

**Dr.A.Arul Lawrence selvakumar[1], N. K.Prema[2]**
Director / Department of CA, Adhiparasakthi Engineering College[1],
Asst.Professor/Department of CSE, IFET College of Engineering[2],
Anna Technical University, Chennai, INDIA, [1, 2]
Aarul72@hotmail.com[1], premasenthi@gmail.com[2]

***Abstract:*** *In order to minimize the overall network traffic in a multi-server wireless multicast system, the number of users served by each server (and hence the group size) should remain constant. As the underlying traffic fluctuates, a split and merge scheme is implemented in a physical server to achieve load balancing. Minimizing the number of servers during the merge operation is NP hard and to achieve these two algorithms namely FFD bin packing algorithm and LL algorithm are proposed to find the near optimal values of destination servers. The performance of these algorithms are analyzed and compared based on several parameters. Results show that LL algorithm outperforms FFD algorithm.*

***Keywords:*** heuristic algorithm, load balancing, dynamic split and merge, destination servers, response time.

## 1. Introduction

The number of users in a multicast group tends to fluctuate due to frequent user join/leave. In order to handle key management efficiently and reduce the join/leave latency a dynamic split and merge scheme is suggested [5], [7]. If the number of users in a server is greater than $Ø_{max}$, the server is split into several logical servers for which the number of users in each server is as close as possible to the optimal sizes of different servers, and bins are destination servers.

An important parameter to study the performance of server packing algorithms is the server response time. For a server packing algorithm to exhibit good convergence, response time is not expected to increase drastically. For example in a M/M/1 queuing model, let $\delta$ be the utilization, and $1/\mu$ be the service time, which is the minimum response time observed when a single request has been processed; then, the response time is expressed as $1/\mu(1- \delta)$. The service time $1/\mu$ of most applications running efficiently on existing servers are sufficiently short and further reduced on the destination server whose performance may be several times higher than that of the existing servers. The response time cannot be more than a certain number of times longer than such a small $1/\mu$. For example, a response time is five times as long as $1/\mu$ if $\delta = 0.8$ (80%).

Thus we need a better heuristic algorithm for finding a near-optimal solution to the server packing problem in reasonable time. Numerous algorithms have already been proposed for one and two-dimensional bin packing problems and First-Fit Decreasing (FFD) is one of the best. FFD and its family are greedy, i.e., items are packed as much as possible into currently prepared bins, and new bin added if an item cannot be packed into any of the current bins. Therefore, the FFD family unbalances the load between bins that are added group size $\rho/g$. If there are some servers in which the total number of users is less than $Ø_{min}$, the groups are merged into a single logical server with the goal of getting as close as possible to $\rho/g$. The problem of finding proper groups to be merged is NP-hard. NP is the set of problems such that, when given a solution, whether it is a true(ly optimal) solution or not can be verified in polynomial time, i.e., O ($n^c$) time, where n is the problem size (the number of items in the packing problem) and c is a constant. Naturally, finding an optimal solution needs more time, for example, exponential time O ($c^n$), and is impossible in practice for not a small n. Even if c = 2 and n = 100, the exponential time will be almost $10^{30}$. The "server" merging problem is also NP hard and the number of destination servers is required to be as small as possible from the point of view of cost reduction and manageability. This minimization can be formalized as a bin packing problem well known in the field of operations research. We are given items of different sizes in the bin packing problem and asked to pack them all into a minimum number of bins with a given capacity. Items for server consolidation are existing servers, item sizes are group early and late. This is why we compared FFD with the least loaded (LL), a load-balancing algorithm widely used in request-based systems. The load balancing approach is more favorable for performance but has not yet been considered within the context of the packing problem. The rest of the paper is organized as follows. Section 2 outlines some of the related work in group key management. Section 3 describes a dynamic merge and split scheme. The detailed explanation on FFD and LL algorithms are given in section 4. The results of the analysis and discussion are given in Section 5. Concluding remarks are provided in section 6.

## 2 RELATED WORKS

Much of the previous work on server optimization has been done without considering the dynamic nature of the multicast group members. This body of work includes dynamic split and merge scheme for large scale wireless multicast. Our work is based on the scheme given in [6] and [7], and we model and analyze it. Previous works address mainly reducing number of existing servers and has considered neither a dynamic split and merge scheme nor the comparison between FFD and LL algorithms. Yong Meng Teo (2001) focuses on an experimental analysis of the performance and scalability of cluster-based web servers.

The three dispatcher- based scheduling algorithms analyzed are: round robin scheduling, least connected based scheduling and least loaded based scheduling. The least loaded algorithm is used as the baseline (upper performance bound) in the analysis and the performance metrics include average waiting time, average response time, and average web server utilization. It is found that the least connected algorithm performs well for medium to high workload.

G. Shen et al (2001) present heuristic algorithms that may be used for light-path routing and wavelength assignment in optical WDM networks under dynamically varying traffic conditions. They considered both the situations where the wavelength continuity constraint is enforced or not enforced along a light-path. The performance of these algorithms has been studied through simulations. A comparative study on their performance with that of a simpler system that uses fixed shortest-path routing has been performed. The proposed algorithms provided lower blocking probabilities and are simple enough to be applied for real time network control and management. They have also studied that the heuristic algorithms are computationally simple and efficient to implement and provide good wavelength utilization leading to efficient usage of the network's resources.

Türkay Dereli and G. Sena Daş (2002) studied a hybrid simulated-annealing (SA) algorithm for the two-dimensional (2D) packing problem. A recursive procedure has been used in the proposed algorithm to allocate a set of items to a single object. The problem has been handled as a permutation problem and the proposed recursive algorithm is hybridized with the simulated annealing algorithm. The effectiveness of the algorithm has been tested on a set of benchmark problems. The computational results have shown that the algorithm gives promising results.

Yao Zhao and Fangchun Yang (2006) proposed an accumulated k-subset algorithm (AK algorithm) to balance load in distributed SLEE. Based on a model of resource heterogeneity and load vector, they have found that the AK algorithm improves the k-subset algorithm by accumulating load information within every update interval. Experiments on different update intervals and request arrival rates suggested AK further reduces herd effect due to stale load information, and outperforms k-subset algorithm by 5%-10%. F. Clautiaux et al (2007) proposed a new exact method for the well-known two-dimensional bin-packing problem. It is based on an iterative decomposition of the set of items into two disjoint subsets. They have tested the efficiency of this method against benchmarks of the literature.

## 3. DYNAMIC SPLIT AND MERGE SCHEME

Since the number of users in a multicast group tends to fluctuate, the system can have variable number of servers. During a busy period when more number of users join the group, number of servers can be more and during a quiet period, the number of servers can be less in order to handle the key management efficiently. We therefore fix a threshold $Ø_{max}$, for the maximum number of users in a group and $Ø_{min,}$ for minimum number of users a server can have at a particular period of time. This is due to the fact that more number of servers adds to the complexity of the system.

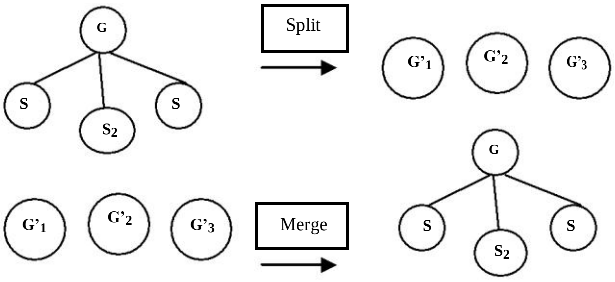


Figure 1: Splitting and Merging for K=3

The number of servers the system needs at a particular period of time is decided by the following procedure.

Step 1: Fix a threshold for $Ø_{max}$ and $Ø_{min}$
Step 2: If u > $Ø_{max,}$ Split the group
Step 3: If u < $Ø_{min,}$ Merge the group

Merging a group with some other group is done in such a way that the total number of users in the merged group does not exceed $Ø_{max}$. Therefore, before merging a group we must find the possible groups that can be merged. Where, $Ø_{max}$ and $Ø_{min}$ represent maximum and minimum number of users in a group respectively.Initially there will be a single server and when more number of users join the group multiple servers are introduced into the system. We use the LKH for generation and distribution of group keys.

Figure 1 shows an example of merging and splitting for K=3. If there is a group in which the total number of users, *u*, is greater than $Ø_{max}$, the group is split into three sub groups and the original subgroup keys, $S_{1,}$ $S_2$ and $S_3$ become the new group keys, $G'_{1,}$ $G'_2$ and $G'_3$, for these three new groups respectively. Whereas, if there are three groups in which *u* is less than $Ø_{min}$, the groups are merged and generate a new group key is generated.

The original group keys, $G'_{1,}$ $G'_2$ and $G'_3$, become subgroup keys, $S_{1,}$ $S_2$ and $S_3$, which can be used to encrypt the new group key, *G* that is sent to these three groups. Hence, the new merged group will have three sets of message overhead, one for each subgroup.

In order to tackle this problem several algorithms have been proposed in the bin packing context for consolidating items into minimum number of bins. In this paper first-fit decreasing (FFD) bin-packing algorithm and the least loaded (LL) are used. Both these algorithms are given the same input and the results are compared for various *n* values.

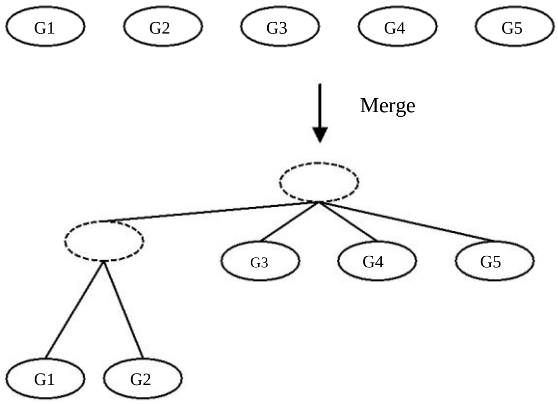


Figure 2: An example of server merge

## 4. ALGORITHMS

In this section, we present two algorithms that are evaluated in our experiments. We study the performance of FFD bin packing algorithm and the LL algorithm. These algorithms were chosen because they are some of the mostly used algorithms in this field, and are fairly simple to implement and do not add redundant delays in the system.

In the FFD algorithm, items are first sorted in decreasing order of size [6]. The FFD algorithm to address the server packing problem is shown in Figure 3. There are a number of empty bins of size with increasing index. The items are placed into the bins one by one, placing each item in the first bin in which it will fit (i.e., the total size of items in the bin does not exceed ) in a round-robin manner. The time complexity of FFD algorithm is shown to be *O(n* log *n),* where *n* is the number of items.

FFD algorithm is applied for merging servers. Each server is considered as an item with its group size as the item size. Assuming that there are many bins with size of $Ø_{min}$, packing operation is done in such a way that, the number of nonempty bins is very close to the optimal number of servers. Therefore, each bin should be filled as much as possible. After packing the groups into the bins, the groups can be merged in a bin into a new larger group served by a single logical server.

The following example demonstrates a simple method to merge the servers. In order to keep the load as balanced as possible, the server with more number of users are added into higher level nearer to the root (i.e., level) while the server with less number of users are added into the lower level. Fig. 2 illustrates a case of merging five servers with a branching factor of 4. If G1 and G2 are two groups with lesser number of users, these two groups are added into the second level and the larger groups are added into the first level. The dotted ovals represent the new nodes created after merging.

The FFD algorithm to address the server packing problem is shown in Figure 3. FFD receives *n* existing servers and sorts them in descending order of utilizations of a certain resource.

TABLE 1
NOMENCLATURE USED IN THE PAPER

| Symbol | Definition |
|---|---|
| $\rho$ | *Total average number of concurrent users* |
| $g$ | *Number of multicast groups* |
| $Ø_{max}$ | *Maximum threshold for the number of users in a server* |
| $Ø_{min}$ | *Minimum threshold for the number of users in a server* |
| $\delta$ | *Utilization* |
| $\mu$ | Service rate |
| $n$ | Number of existing servers |
| $m$ | Number of destination servers |
| $u$ | Number of users in the server |
| K | Branching factor |

The sorting is carried out for the largest (peak) utilizations within a time period even if time – series data are used. After the algorithm is executed, we obtain server accommodations $X_j$(j = 1, ...., *m*), where m is the number of destination servers. The function packable ($X_j$, $s_i$) returns true if packing existing server $s_i$ into destination server $s_j$ satisfies the constraints (i.e.., the utilization of $s_j$ does not exceed a threshold for any resource); otherwise it returns false.

FFD sequentially checks if all existing servers $s_1$, ...., $s_n$ can be packed into one of m current destination servers. FFD then packs $s_i$ into a destination server first found to be able to accommodate it. If $s_i$ cannot be packed into any current destination server, the (*m*+1) – th destination server is added and accommodates it. The complexity of this FFD algorithms is O($n_2$) because *m* is almost proportional to *n*. Here, we assumed the utilizations of no existing servers were beyond thresholds. Note that the binary search technique can reduce this complexity to O(*n* log *n*), but the sequential search is better for actual problems with time – series data.

```
Sort existing servers to $\{s_1,...,s_n\}$ in descending order;

$m \leftarrow 1$; $X_1 \leftarrow \{\}$;
for $i \leftarrow 1$ to n do
    for $j \leftarrow 1$ to m do
        if packable($X_i$, $s_i$) then
            $X_j \leftarrow X_j \cup \{s_i\}$;
            break
        fi
    end for;
    if $j=m+1$ then          /* If fail to pack $s_i$ */
        $m \leftarrow m+1$;   /* a new server is added */
        $X_m \leftarrow \{s_i\}$    /* to have $s_i$ */
    fi
end for
```

Figure 3: FFD algorithm

## 4.2 Least Loaded algorithm

The LL algorithm works on the principle of load balancing. The LL algorithms attempts to balance the load between servers by assigning incoming jobs to the least – loaded server. In server packing, an existing server with a high utilization is packed into a destination server with a low utilization. Figure 4 shows the LL algorithm that addresses the server packing problem. The function LB ($\{s_1, .....s_n\}$) in the figure returns the theoretical lower bound for the number of destination servers that accommodate existing servers $\{s_1, ..... s_n\}$. The lower bound is the smallest integer of numbers larger than the sum of the utilizations divided by a threshold. The lower bound for the CPU is $LB_c$ = $[\sum_{i=1}^{n} \rho_{c_i} /R_c]$ while that for the disk is $LB_d$ = $[\sum_{i=1}^{n} \rho_{d_i} /R_d]$ Function LB ($\{s_1, .....s_n\}$) returns the larger integer of the two lower bounds.

Figure 4: LL algorithm

```
sort existing servers to {$s_1$,.....$s_n$} in descending order;

m ← LB ({$S_1$,....$S_n$});
while true do
for j ← 1 to m do
Xj ← {} /* initilization *?
end for;
for i ← 1 to n do
sort destination servers to {$X_1$,...$X_m$} in ascending
order;
for j ← 1 to m do
if packable ($X_j$, $S_i$) then $X_j$ ← $X_j$ ∪{$S_i$};
break
fi
end for;

if j = m + 1 then /* If fail to
pack $s_i$, */
m ← m +1; /* a new server is added */
break
fi
end for;
if i = n + 1 then /* all packed */
break
fi
end while
```

The complexity of LL is O($d$ . $n^2$log $n$), where d is the difference between the lower bound and the final number m of destination servers. This complexity can be reduced to O($d$ . $n^2$) if we efficiently sort destination servers. The sorting does not actually require O($n$ log $n$) time but O($n$) because only the utilizations of a destination server that has accommodated $s_i$ is updated in iterations with i.

## 5. RESULTS AND DISCUSSIONS

In this section, we present the results from an extensive set of experiments to investigate the performance of the algorithms under study.

The algorithms are systematically evaluated across the wide spectrum of distribution parameter values for virtual server load and node capacity to give a clear view of the performance of the algorithms.

The performance metrics considered are:

- Consolidation Efficiencies
- Destination Servers
- Convergence Time
- Total Workload moved
- Success Ratio and
- Response Time

In a multiserver network of computing hosts, the performance of the system depends crucially on dividing up work effectively across multiple server nodes. The random arrival of users at each server is likely to bring about uneven server loads in such a system. Dynamic load balancing algorithms compared to bin packing algorithms have the potential to perform well under heavy loads. Naturally dynamic load balancing strategies are more complex and the overheads involved are much more. But one can not negate their benefits. Load balancing is found to reduce significantly the mean and standard deviation of job response times, especially under heavy and/or unbalanced workload. The performance is strongly dependent upon the load index. The reduction of the mean response time increases with the number of hosts, but levels off beyond a few tens of hosts.

| *n* | Algorithm | *m* | *m/LB* | Convergence Time (sec) |
|---|---|---|---|---|
| 50 | FFD | 39.6 | 1.34 | 0.061 |
| | LL | 37 | 1.12 | 0.073 |
| 100 | FFD | 87.3 | 1.26 | 0.069 |
| | LL | 84.2 | 1.11 | 0.078 |
| 150 | FFD | 131.7 | 1.19 | 0.082 |
| | LL | 127 | 1.09 | 0.188 |
| 200 | FFD | 188 | 1.14 | 0.127 |
| | LL | 171 | 1.09 | 0.284 |
| 250 | FFD | 217 | 1.08 | 0.142 |
| | LL | 203 | 1.05 | 0.323 |

Table 2: Comparison of average number m of destination servers offered by FFD and LL for various n values

The values m / LB closer to 1.00 mean higher efficiencies. The rightmost column indicates the average execution times for the algorithms. The algorithms have been implemented in java language (JDK 1.5). The results show that while m increases linearly with n, LL algorithm results in the better m values compared to FFD algorithm. Similarly, the convergence time for LL algorithm is better than FFD.

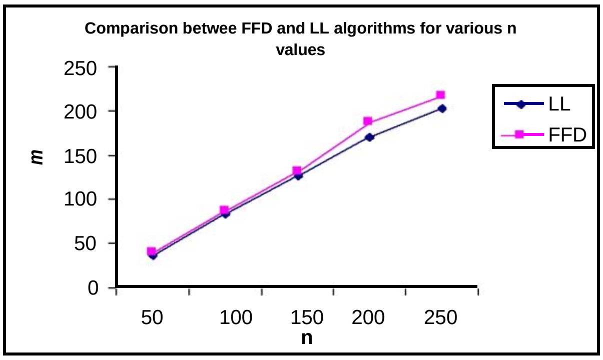


Figure 5: Comparison of FFD and LL based on m

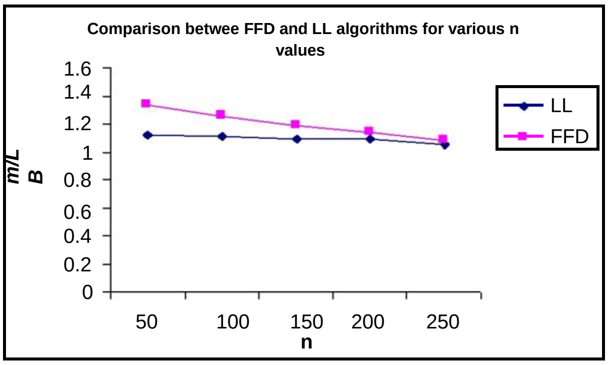


Figure 6: Comparison of FFD and LL based on m/LB

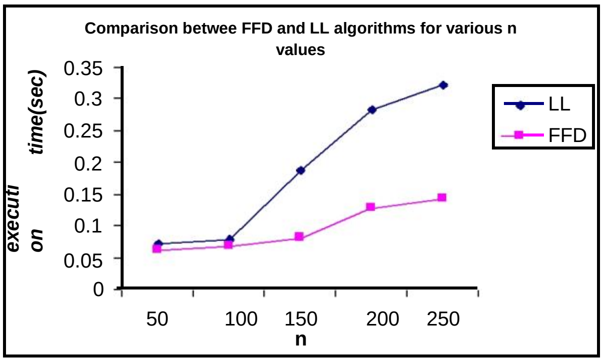


Figure 7: Comparison of FFD and LL based on convergence time

Response time is a function of the CPU requirements of the components comprising the application, the number of remote messages exchanged by these components, and the current load on required resources. Our objective function is to improve performance which is simply defined as minimization of the system average response time. For this, the number of application remote messages exchanged has to be kept as low as possible. For each remote message exchanged, we model the cost incurred by adding a delay to the time consumed by the message originator. Application performance (response time) is modeled as described in Fig.5.

In the figures, the workload moved performance result is plotted in the form of the total load moved as a fraction of the total workload of the system when the algorithms successfully terminate. The success ratio is defined to be the percentage of the problem instances for which feasible solutions are found among all problem instances.

Given a server S, comprising n users: $n_1 \cdots n_n$:
response time
(S(i))=wait_time(S(i))+cpu_required(S(i))+itc_cost(S(i))+system_cpu(S(i))
where: wait_time(S(i))=start_time(S(i))-arrival_time(S(i))
itc_cost(S(i))=number_of_key_exchanges(S(i))
[itc-information transfer cost]
system_cpu(S(i))=cpu time „stolen" for system activities during the components" execution.

Given a server with a heuristic algorithm (HA) and n users, performance (HA) = response_time

$$(HA) = 1/n \sum_{k=1}^{n} \text{response\_time(server } S_k)$$

Figure 8: Calculation of response time

| n | Algorithm | Success Ratio (%) | Moving Workload (%) | Response Time(ms) - Split | Response Time(ms) - Merge |
|---|---|---|---|---|---|
| 50 | FFD | 100 | 19 | 18.5 | 11.3 |
| | LL | 99.8 | 20.3 | 11.2 | 10.7 |
| 100 | FFD | 99.3 | 19.8 | 19.1 | 11.9 |
| | LL | 99.6 | 20.8 | 11.9 | 10.9 |
| 150 | FFD | 98.7 | 21 | 19.9 | 12.4 |
| | LL | 99.5 | 21.6 | 12.3 | 11.4 |
| 200 | FFD | 98.5 | 21.3 | 20.3 | 12.6 |
| | LL | 99.5 | 21.9 | 13.1 | 11.8 |
| 250 | FFD | 98.5 | 21.5 | 21.5 | 13 |
| | LL | 99.5 | 22.1 | 13.5 | 12.1 |

Table 3: Comparison of average number m of destination servers offered by FFD and LL for various n values

Load balancing is still very effective when a large portion of the workload is immobile. All servers, even those with light loads, benefit from load balancing. System instability is possible, but can be easily avoided. The Least-Loaded algorithm produced average response times representing 34.8% of the average response times produced by the FFD bin packing algorithm.

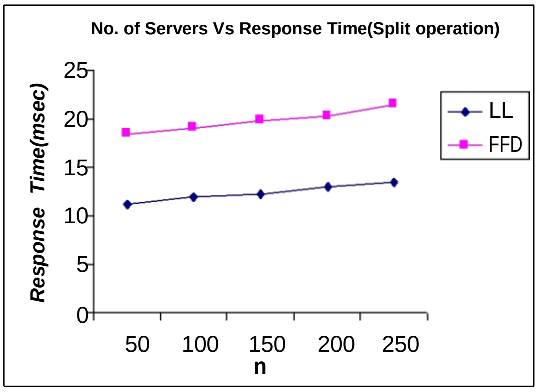


Figure 9: Comparison of FFD and LL based on response time (split operation)

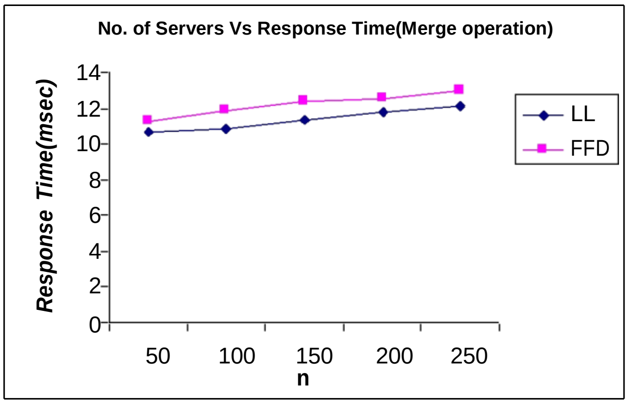


Figure 10: Comparison of FFD and LL based on response time (merge operation)

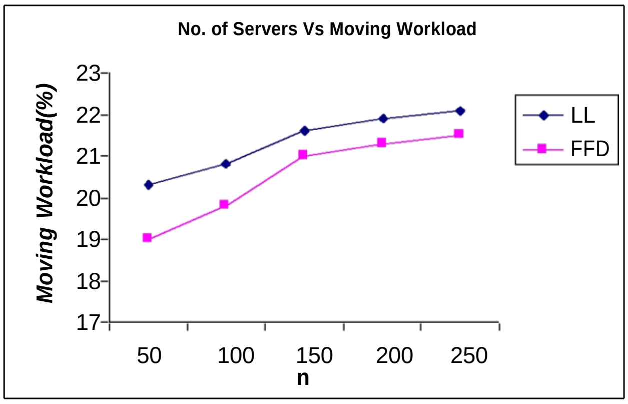


Figure 11: Comparison of FFD and LL based on success ratio

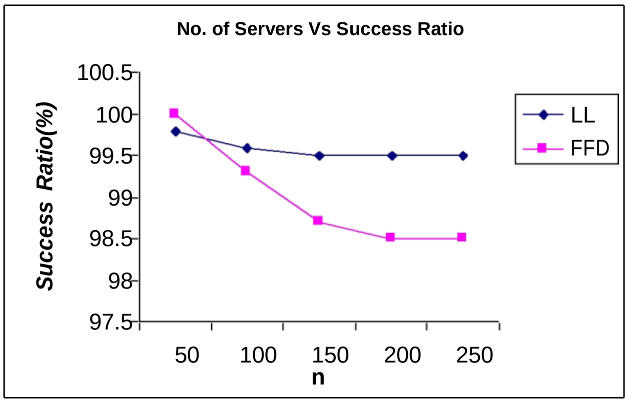


Figure 12: Comparison of FFD and LL based on moving work load

## 6. Conclusion

In order to efficiently handle the frequent membership change in a multicast system, a dynamic split and merge technique has been proposed. Two algorithms, FFD and LL, have been suggested to get near optimal values for number of destination servers during the merge operation. Comparison between FFD and LL algorithm shows that the convergence time is lower for FFD, whereas LL algorithm performs well in getting the number of destination servers very close to the optimal value and balances the load better than FFD.

## REFERENCES

*[1].Valeria Cardellini, Michele Colajanni, Philip S. Yu, "Redirection Algorithms for Load Sharing in Distributed Web-server Systems," icdcs, pp.0528, 19th IEEE International Conference on Distributed Computing Systems (ICDCS'99), 1999*

*[2].Yao Zhao; Fangchun Yang **,** **“**A Dynamic Load Balancing Algorithm for Distributed SLEE in Mobile Service Provisioning**”,** International Conference on Wireless Communications, Networking and Mobile Computing, 2006 WiCOM 2006, Volume 1, Issue 4, 22-24 Sept. 2006 Page(s):1 – 4*

***[3].Yong Meng Teo, “**Comparison of Load Balancing Strategies on Cluster-based Web Servers, SIMULATION, Vol. 77, No. 5-6, Pages 185-195, 2001**.***

*[4].G. Shen, S. K. Bose, T. H. Cheng, C. Lu and T. Y. Chai, —Efficient heuristic algorithms for light-path routing and wavelength assignment in WDM networks under dynamically varying loads , Computer Communications, Volume 24, Issues 3-4, Pages 364-373, 2001.*

*[5] F. Clautiaux, J. Carlier, and A. Moukrim. A new exact method for the two-dimensional bin packing problem with fixed orientation operations research letters, Vol. 35, No.3, pp. 357-364, 2007*

*[6].A Caprara and P. Toth. Lower bounds and algorithms for two dimensional vector packing problem. Discrete Applied Mathematics, 111:231-262, 2001*

*[7].Lodi, S. Martello, and D. Vigo. Recent Advances on two dimensional vector packing problem Discrete Applied Mathematics, 123:379-396, 2002.*

*[8].Spellmann, K. Erickson, and J. Reynolds. Server consolidation using performance modeling. IT Professional, 5:31-36, 2003.*

*[9]. D L Eager , E D Lazowska , J Zahorjan, —A comparison of receiver-initiated and sender-initiated adaptive load sharing, Performance Evaluation, v.6 n.1, p.53-68, 1986.*

*[10]. Arthur P. Goldberg, Gerald J. Popek , Stephen S. Lavenberg, —A Validated Distributed System Performance Model, Proceedings of the 9th International Symposium on Computer Performance Modelling, Measurement and Evaluation, p.251-268, May 25-27, 1983*

*[11].Haċ,"A Distributed Algorithm for Performance Improvement through Replication and Migration," Proc. IEEE Computer Networking Symposium November 17-18, 1986, Washington, D.C., pp. 163--168.*

*[12] Asser N. Tantawi, Don Towsley, Optimal static load balancing in distributed computer systems, Journal of the ACM (JACM), vol.32, no.2, pp .445-465, 1985.*

*[13].Y.-T. Wang, and R. J. T. Morris, "Load Sharing in Distributed Systems," IEEE Transactions on Computers, Vol. C-34, No.3, pp. 204—217, 1985.*

*[14].B. S. Baker, —A new proof for the first-fit decreasing bin-packing algorithm, J. Algorithms, vol. 6, pp. 49–70, 1985.*

# Proposing LT based Search in PDM Systems for Better Information Retrieval

Zeeshan Ahmed

University of Wuerzburg Germany
Vienna University of Technology Austria

## Abstract

PDM Systems contain and manage heavy amount of data but the search mechanism of most of the systems is not intelligent which can process user's natural language based queries to extract desired information. Currently available search mechanisms in almost all of the PDM systems are not very efficient and based on old ways of searching information by entering the relevant information to the respective fields of search forms to find out some specific information from attached repositories. Targeting this issue, a thorough research was conducted in fields of PDM Systems and Language Technology. Concerning the PDM System, conducted research provides the information about PDM and PDM Systems in detail. Concerning the field of Language Technology, helps in implementing a search mechanism for PDM Systems to search user's needed information by analyzing user's natural language based requests. The accomplished goal of this research was to support the field of PDM with a new proposition of a conceptual model for the implementation of natural language based search. The proposed conceptual model is successfully designed and partially implementation in the form of a prototype. Describing the proposition in detail the main concept, implementation designs and developed prototype of proposed approach is discussed in this paper. Implemented prototype is compared with respective functions of existing PDM systems .i.e., Windchill and CIM to evaluate its effectiveness against targeted challenges.

***Keywords:*** Product Data Management System; Language Technology; Search

## 1. Introduction

In early 1970s there was no such system to automate the process of data management, then in 1980s Computer Integrated Manufacturing was introduced but seemed not to be successful in product data management. With emergence of CAD technologies PDM Systems are introduced and used to manage engineering data, activities and processes through better control of engineering data, activities, changes and product configurations. PDM products mainly manage information about design and manufacturing of products including technical operations and running projects. PDM is also renowned as Engineering Data Management and Engineering Document Management Systems because it provides better management and control over engineering data, activities, and changes related to design and manufacture of product. Product Lifecycle Management (PLM) is another acronym for PDM to manage the entire development life cycle of the product by integrating people, data, processes and business systems.

PDM provides a backbone for the controlled flow of engineering information throughout the product life cycle by using engineering data, such as CAD, ERP and field service. Moreover PDM also supports product teams and techniques by providing Concurrent Engineering in improving engineering workflow. PDM systems address issues such as control, quality, reuse, security and availability of engineering data. PDM performs five main functions to integrate and manage all applications, information, and processes during the associated product life cycle i.e., Data vault and document management, Workflow and process management, Product structure management, Parts management and Program management [8].

The major objectives of PDM are to reduce the cost of engineering, reduce effort in product development life cycle, reduce time in change handling and new product development, improve the quality and services of the product, deliver and support products at the given time, improve team coordination, increase customization of products, maintain product configuration based information, manage large volumes of data generated by computer based systems, reduce engineering environment based problems, provide better access to information, provide better reuse of design information, provide common data warehousing, secure engineering data's originality, prevent error creation and propagation and make a strong effect on market shares. Moreover PDM is also supposed to handle business process work flows, change management, revision control, product configurations, product structure management, project tracking and resource planning.

To meet the aforementioned objectives of PDM, the concepts turned into the real time applications called PDM systems. These systems are developed to manage product data throughout enterprise, ensuring the availability of right information for the right person at the right time and in the right form [9]. PDM systems are mainly used by project managers, designers, engineers, administrators, manufacturing, sales, marketing, purchasing and other personal in the companies. Product related information controlled by PDM systems includes part definitions and other design data, engineering drawings, project plans, software components of products, product specifications, NC programs, analysis results, correspondence, bills of materials etc. Commercial PDM systems have been developed and are used in companies for more than a decade now. Many companies today have realized strategic importance of a PDM system implementation and usage. But the implementations have often been associated with problems and large costs for the companies. Still there is lot of work to be done in order to improve PDM systems functionality and to develop methods for their proper implementation and use in different areas of the product development and the sales delivery process.

PDM Systems plays an important role in tracking products among different engineering groups by reducing time to market, increasing product quality and reducing total cost. Furthermore PDM System controls, manages and distributes product data automatically to the needed people. A PDM system is typically used within enterprise to organization to access and control data related to its products and to manage the life cycles of those products. PDM Systems are capable of providing user directed and utility functions. User directed functions are i.e. Data vault and document management for storage and retrieval of product information, workflow and process management procedures for handling product data and providing of a mechanism to drive a business with information, Product structure management handling of bills of material, product configurations, and associated versions and design variants, Parts management providing of information on standard components and facilitating reuse of designs. Program management provides work breakdown structures and allows coordination between product related processes, resource scheduling and project tracking. Where as the utility functions are the □□Communication and notification capabilities such as links to email provide support for information transfer and events notification,□ □Data transport tracking of data locations and moving of the data from one location or application to another, □□Data translation file exchange in the proper format. □□Image services, storage, access, viewing and markup of product information, □□System administration system control and monitoring of operation and security.

In this paper; in section 2 a clear vision to the targeted problems is provided, then in section 3 related research work is discussed, section 4 presents the over all proposed approach where as section 5 presents the solution towards the targeted and discussed problem. Section 6 discusses implementation designs of proposed approach to develop it into the form of software application, discussed in section 7. Proving the effectiveness against targeted problem, implemented prototype is compared with some existing PDM Systems in section 8. Discussion is concluded in section 10 after the presentation of some existing limitations in prototype development in section 9.

## 2. Problem Definition

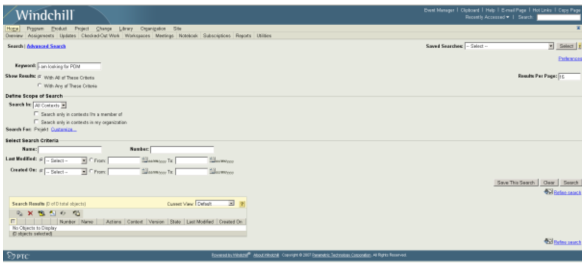
PDM Systems contain and manage heavy amount of data, which itself is a big achievement but on the other hand the problem starts when user needs to find out some information out of this heavy data. The search mechanism of most of the available PDM Systems is limited because it provides limited structured options to find out the information. If user needed information is available amongst those provided search options then it is fine but in case user needs some information which can't be found using provided options then the outcome will be limited. Moreover it also consumes time that at first the user needs to read the given search options, then to provide the needed information by filling forms and then performing search.

No doubt available search mechanisms of PDM Systems capable of searching results with high probability but it is time consuming, heavily structured, static and limited. There are some serious problematic issues in search mechanisms of existing PDM systems which are needed to be resolved. The search mechanism of most of the existing PDM systems is;

1. Based on a static way of searching information like filling forms and making search.
2. Not capable of processing user's structured / unstructured natural language based queries to search information.

3. Not capable of retrieving information by extracting Meta data out of data.
4. Not capable of providing geometrical search for graphical documents e.g. a user is interested in finding some CAD documents and he wants to make the search with some geometrical information like size of screw etc. but using existing PDM System's search mechanism it's not possible.
5. Not capable of performing system spanning based search e.g. a user is interested in finding some information needed to design a CAD document but that is not available in the default System, then it will not look for other web based available relevant information sources for CAD document designing. In short the existing PDM systems do not provide the multiple database connectivity and search; it only looks in its own connected database for the retrieval of the required information.
6. Not capable of weighting extracted results like some other search engines e.g. Google etc.

Currently available search mechanisms of some of the PDM Systems e.g. Windchill [1] as shown below in Fig. 1 and CIM Data Base [2] as shown below in Fig. 2 etc. are limited because they provide restricted options to find out the information.


Keyword: I am looking for PDM

Show Results: ● With All of These Criteria ○ With Any of These Criteria

Search In: All Contexts

Search For: Projekt

Last Modified: -- Select --

Created On: -- Select --

Results Per Page: 15

Search Results (0 of 0 total objects)

Current View: Default

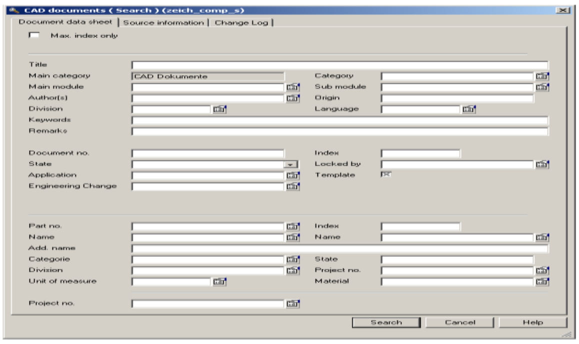
No Objects to Display

(0 objects selected)

Windchill
Event Manager | Clipboard | Help | E-mail Page | Hot Links | Copy Page
Recently Accessed | Search
Home Program Product Project Change Library Organization Site
Overview Assignments Updates Checked-Out Work Workspaces Meetings Notebook Subscriptions Reports Utilities
Search | Advanced Search
Saved Searches: -- Select --
Select
Preferences
Define Scope of Search
□ Search only in contexts I'm a member of
□ Search only in contexts in my organization
Customize...
Select Search Criteria
Name:
Number:
From: To:
Save This Search
Clear
Search
Refine search
Number Name Actions Context Version State Last Modified Created On
Refine search
PTC
Powered by Windchill About Windchill Copyright © 2007 Parametric Technology Corporation. All Rights Reserved.

Fig. 1. Windchill Advanced Search Form

As shown in Fig.1, the presented advanced search mechanism of Windchill consists of a form with several different options to make search e.g. if a user is looking for a documents then one is required to enter the following information like name, number, last modified date etc and only then he can look for the required information. Windchill offers customization of the search form e.g. a user can enhance search form by adding or deleting some options from the main option list and can improve the search mechanism. Though Windchill is providing an efficient form based search mechanism, at the same time it is restricting the user to the existing search with provided options, consuming user's time in filling a form for finding some information etc. This Windchill search form also offers an option to make full text search but that can only work for one keyword e.g. "CAD". It does not allow to make search by entering natural language based short conditional objective statements e.g. "want CAD" or simple conditional objective statements e.g. "I am looking for CAD" or multiple conditions based objective statements e.g. "I am looking for CAD where document type is doc".


CAD documents ( Search ) (zeich_comp_s)

Main category: CAD Dokumente

Fig. 2. CIM Database CAD Document Search Form

As shown in Fig. 2, the presented search form to search CAD documents using CIM Database consists of several different options like Title, Category, Author, Origin, Language etc. Like Windchill the search mechanism of CIM Database is also limited and time consuming. In both the cases (CIM and Windchill) user at first needs to fill some form to make search. Moreover both the systems are incapable of providing intelligent search, Meta data based full text search, geometrical search and multiple database spanning search. Moreover another deficiency in CIM Database search mechanism is that it is case sensitive by default. To make case insensitive search a user at first has to locate and change options. Like Windchill, CIM Database search forms also offers an option to make full text search but that can only work for one keyword e.g. "PDM". It does not allow to make search by entering natural language based short conditional objective statements e.g. "need PDM" or simple conditional objective statements e.g. "I am looking for PDM" or multiple conditions based objective statements e.g. "I am looking for PDM where document type is doc and pdf".

## 3. Related Research Work - Language Technology

Targeting the objective of this research; introduction and implementation of a new way for the implementation of a natural language based search mechanism to extract desired results from database by processing and modeling natural language based user requests, a thorough is conducted. Meeting the goals of this research, the filed of Language Technology is selected and explored.

Language Technology is a linguistic based field of computer science, also called as Human Language Technology or Natural Language Processing (NLP) [3]. Language technology is about to make machine capable of reading, listening, understanding and analyzing human (natural) language. The main objective of Language technology is to teach machines, how to communicate and help humans by communicating (listening and speaking) with them [4]. To meet aforementioned goal of natural language processing and make machines understand natural language, language technology is composed of two steps i.e., Tokenization and Parsing. Tokenization is also called lexical analysis, during the process of lexical analysis, natural language based instruction tokenizes in possible number of tokens by a lexer. Then these tokens are matched with the dictionary of used natural language for processing to identify valid and invalid tokens. These tokens consist of letters, digits and symbols. Then at the end of the process of lexical analysis a stream of tokens is generated by lexer for further processing. Generated token stream by lexer, then is considered by the parser to semantically evaluate the instruction. Every parser of any language programming or natural language consists of a set of rules. These set of rules are the combinations of tokens of dictionary of the language. To evaluate the semantic of a statement it is compulsory to first evaluate the tokens from dictionary and then the combinations of used valid tokens to understand the meaning or semantic of statement. To meet parsing goals, parsers are divided in to two types i.e., LL Parser and LR Parser. The LL parser constructs left most derivation and LR constructs eight most derivation during parsing, in simple words LL parser starts parsing by replacing nonterminals from left side and LR from right side. Moreover parser creates the sequences of tokens to put them into an Abstract Syntax Tree (AST).

In the domain of language technology, following the concepts of language technology, many approaches have been introduced by many researchers which are providing lots of values in the implementation of natural language processing i.e., Analyzing English Grammar [5], Layerd Domain Class [6] and Another tool for language recognition [7] etc.

### *3.1. Analyzing English Grammar[5]*

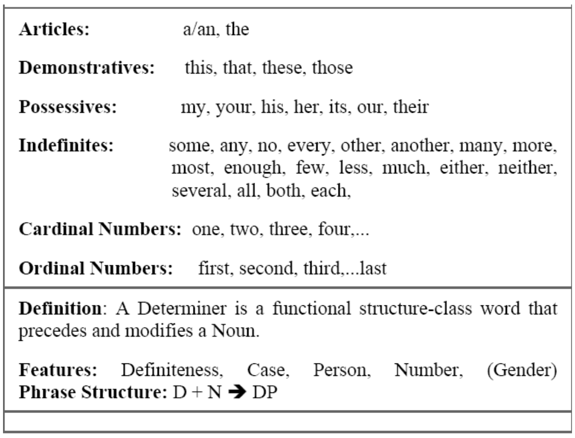
Author has discussed an approach to analyze English grammar. Approach is divided into following three steps i.e., Division of grammar, Tokenization of Sentences using Lexer, Seven Steps Structuring and Categorization.

1. Division of grammar; Grammar is divided into two inter related studies: Morphology and Syntax. Morphology is to form the words in smaller units called morphemes, for example, the word "books" here would have two morphemes (i) the root/stem "book", and (ii) the inflectional morpheme {s} showing number [+Plural]. Syntax is to string words together to form (Partial) Phrases, Clauses, and (full) Sentences. For example, as presented above, the Determiner Phrase (DP) is formed from out of the string D+N.

2. Tokenization of Sentences using Lexer; Tokenizing sentences is by categorizing in two categories i.e., Class words and Functionals. Class words include Nouns, Verbs, Adjectives, and Adverbs while functional includes Determiners, Auxiliary/Modals, Pronouns, Complementizers, and Qualifiers.

3. Structuring and Categorization, as shown in Fig.3. To further structure the tokens of input sentences, a seven step guide is followed by author.

   a. Determining (DP) Articles, a/the; Demonstratives, this/that/these/those; Genitives my/our/your/their, etc.) precede Nouns: e.g., The book
   b. Determining Adjectives (AdjP) (modifiers of Nouns e.g., red, good, fast, etc.) precede and generally describe nouns [(Det)+Adj+N] (e.g., (The) read shoes
   c. Determining Main Verbs (MVP) (Tensed Verbs such as goes/went, walks/walked, keeps/kept, etc.) typically follow the subject of declarative sentences (adhering to the English SVO Subject Verb Object word order).
   d. Determining Auxiliary/Modals (AuxP) serve to introduce Main Verbs (MVPs). All functional features associated with Verbs {Tense, and Agreement features of Person/Number} are borne out of the Aux
   e. Determining Verb Phrase (VP) (Infinitive Phrase) unlike the MVP is a Non Tensed Verb Phrase. Such VPs tend to project after an already positioned MVP. These phrases include all three Infinitive types/forms e.g., I like to cook (=Infinitive 'to'), I like cooking (Infinitive 'ing'), I can cook (Infinitive 'bare verb stem' ).



**Articles:** a/an, the

**Demonstratives:** this, that, these, those

**Possessives:** my, your, his, her, its, our, their

**Indefinites:** some, any, no, every, other, another, many, more, most, enough, few, less, much, either, neither, several, all, both, each,

**Cardinal Numbers:** one, two, three, four,...

**Ordinal Numbers:** first, second, third,...last

**Definition**: A Determiner is a functional structure-class word that precedes and modifies a Noun.

**Features:** Definiteness, Case, Person, Number, (Gender)
**Phrase Structure:** D + N ➔ DP

Fig. 3. Summary of Determiners [5]

f. Determining Adverb Phrase (AdvP), like adjectives for nouns, modifies verbs e.g, softly touched, quickly ran, etc., (Adv+V).

g. Determining SVO/Head Initial Phrase: In addition to English being an SVO word order, English stipulates that the Head of a Phrase must be in the first initial position within the phrase (i.e., that word which labels the phrase such as Determiner, Adjective, Main Verb, etc. must come first in forming the phrase.

Analyzing English Grammar is highly relevant to the goal of this research. In this natural language processing module, the same mechanism is used to parse the sentences like tokenizing sentences, identifying relevant and irrelevant tokens with respect to the used grammar and then trying to analyze semantics of the input sentence.

### *3.2. Layered Domain Class (LDC) [6]*

Authors developed software called Layer Domain Class (LDC) for parsing and deep semantic processing of English language based sentences. LDC consists of two major components and an external retrieval module. The first component, which was called "Prep," obtains information about a new domain and the language to be used in discussing it. The second, "user phase," component of LDC resembles an ordinary NL processor. To process English languages based sentences the whole process of parsing in LDC is divided into following four steps i.e., Scanning, Parsing, Semantic Processing and Output Generation as shown below in Fig. 4.

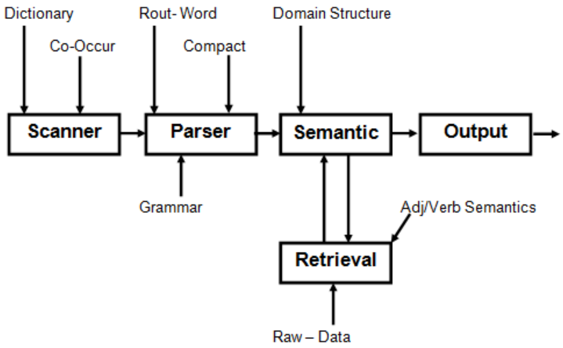


Fig. 4. Process of Phrase Parsing in LDC [6]

1. Scanner is to identify each word of the typed or spoken input and retrieve information about it from the dictionary file, which will have been created by Prep.
2. Parser is to determine, from the information provided by the scanner, the syntactic structure of the input. In a computational domain, especially one for retrieval rather than programming, the syntactic complexity of most inputs lies in the complexity of their noun phrases. For this reason relative clause verb forms are considered as basic, and sentence level verbs as derived.
3. Semantics module is to translate the tree like parse structures into an internal form that is referred to as "bubble structures". These structures, which can be interpreted directly or can be translated into a formal query for the external retrieval component, possess at least three user desirable properties.
4. Output generator converts the top level datarep produced by semantics into a human readable form.

### *3.3. Another Tool for Language Recognition (ANTLR) [7]*

ANTLR is a tool, developed in 1983 by Professor Terence Parr and his colleagues to write grammar of languages, although this tool is only used for writing programming languages grammars but it also has the capability to write natural language grammar as well. ANTLR contain frameworks for compilers, recognizers and translators. It is implemented in Java but it can generate source code in Java, C, C++, C#, Objective C, Python and Ruby. ANTLR use EBNF (Extended Backus Naur Form) for the grammars, which is very formal way to describe the grammar. ANTLR provides a standard editor for grammar writing and generating lexer and parser. Till now this tool has been used for programming language's grammar writing but I am considering for natural language processing by writing natural language's grammar and generating lexer and parser to make the machine to understand it. ANTLR allows for generation of parsers, lexers, tree parsers and combined lexer parsers. Parsers can automatically generate Abstract Syntax Trees which can be further processed with tree parsers. ANTLR provides a single consistent notation for specifying lexers, parsers and tree parsers. This is in contrast with other parser/lexer generators and adds greatly to the tool's ease of use. By default ANTLR reads a grammar and generates a recognizer for the language defined by the grammar.

ANTLR has many belonging applications and opportunities to extensibilities. One of the biggest benefits is the grammar syntax; it is in EBNF form, which is a Meta syntax notation. Each EBNF rule has a left hand side (LHS) which gives the name of the rule and a right hand side (RHS) which gives the exact definition of the rule. Between the LHS and RHS there is the symbol ":" (colon), which separates the left from the right side and means "is defined as". One another benefit is the graphical grammar editor and debugger called ANTLRWorks, written by Jean Bovet and gives us the possibility to edit, visualize, interpret and debug any ANTLR grammar. It is based on a grammar editor with an interpreter for rapid prototyping and a language agnostic debugger for isolating grammar errors. ANTLRWorks also helps in eliminating grammar nondeterminisms by highlighting nondeterministic paths in the syntax diagram associated with a grammar. ANTLRWorks helps in making grammars more accessible to the average programmer by improve maintainability and readability of grammars and providing excellent grammar navigation and refactoring tools. To meet aforementioned goal of this research and development with respect to the implementation of an intelligent search by processing natural language, ANTLR can be used to take help in writing lexer and parser for own written natural language grammar.

## 4. Proposed Approach

Keeping eyes on above discussed major currently faced challenges of Product Data Management Systems, we can say, right now PDM community is in need of a very convincing and strong approach for its clients to win their confidence over the PDM Systems. Moreover PDM community also needs a new approach which can be very helpful in implementing the concepts of Product Data Management in the form of a web based software application capable of providing a user friendly graphical user interface which can also intelligently handle user's structured and unstructured natural language based requests for fast, optimized and efficient information retrieval or search mechanism

Targeting some of existing Product Data Management System development issues, proposed an approach, which was first conceptually modeled then converted in to implementation designs and which was then developed in the form of a prototype application. Proposed approach consists of four different modules i.e. Flexible GUI, NLP Search, Data Manager and Data Representer. Proposed approach is mainly for the development of a PDM system capable of providing a flexible web based graphical user interface, identifying user's structured and unstructured natural language based requests, processing natural language based user's requests to extract results from attached repositories, manage data in database management system and represent system outputted information as the result of user input in user's understandable format. Residing within the scope of this paper only discussing the NLP Search module in detail, as it is designed and implemented targeting the above mentioned unresolved issue of unintelligent search in PDM Systems.

## 5. NLP Search

Currently available data search mechanism in almost al PDM systems is not very efficient and based on old ways of searching information by entering the relevant information to the respective fields to search some specific information from attached repositories. As shown in Figure 5, if a user in need of some information, whether it is available in attached different networks sources or it is a CAD design or it is a document based information, in all the cases user has to spend some time in administrating the PDM system with the information he has about the thing which he is looking for e.g. to look for a document with specific type, date and author, user has to enter these relevant information entities in respective fields of search form and then system will check for the availability of that information. In case user does not know the complete or relevant information about the object he is looking for then it can be time consuming by increasing the manual efforts at user end. Targeting the problem of this unintelligent and old fashioned way of searching data in PDM Systems, I propose a natural language based search mechanism i.e. NLP Search for PDM System development.

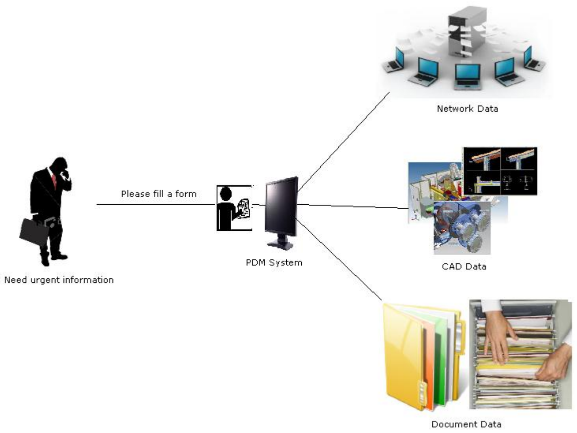


Fig. 5. Search in mechanism in PDM Systems

NLP Search is the second most important module of the proposed approach. This module is proposed by targeting the problem of unintelligent search in PDM Systems. The proposed job of this module is to take natural language based instructions from user to extract or search user needed information from attached repository. The proposed mechanism in this module is based on the extracted knowledge obtained as the result of conducted research in the field Language Technology and Semantic web.

To meet the aforementioned goal of a NLP search mechanism implementation, we need to propose a human machine communication system capable of translating user's natural language based instruction to machine understandable format. This data translation can be performed by structuring and categorization of data. To achieve this goal, a grammar is needed to be written based on the dictionary and rules of natural language used for user system communication. Residing within the scope of this research, a grammar is written which is based on small dictionary of keywords and set of rules. The designed grammatical view of the grammar is shown in figure 6. The grammatical view is designed with respect to English language grammar (rules) and consists of three parts i.e. A-Subject, B-Verb and C-Object, the simplest grammatical rule of used language i.e., "English".

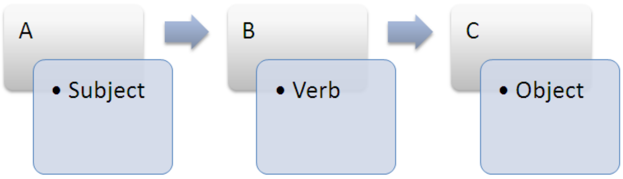


Fig. 6. Abstract View of Grammar

A is the subject; the set of tokens representing Grammatical Persons e.g. "I, We, He, she etc". B is the representation of Verbs and Helping verbs e.g. "is, need, want, look, give etc" and C is the Object; a set of tokens representing Nouns like "document, project, life cycle, person etc". These three parts most of the times combine to create a natural language based sentence e.g. "I am looking for PDM Systems" etc. Furthermore six different conditions are also introduced and used in this grammar i.e. Between, Euqal, Greater, Less, Greater Than, Less Than and With, as shown in Figure 7.

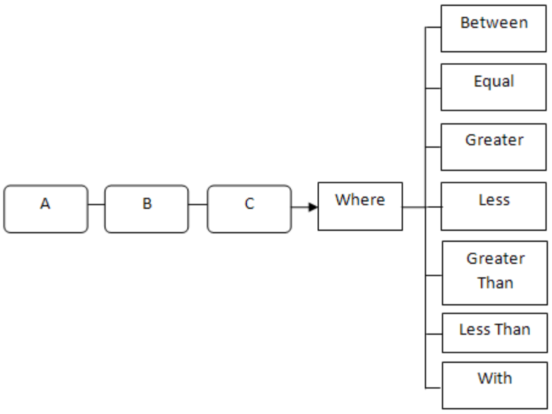


Fig. 7. Grammatical View with Conditions

Using this grammar four kinds of natural language based instructions can be processed i.e., Keyword based instructions, Short conditional objective statements, Simple conditional objective statements and Multiple Conditions based Objective statements.

- Keyword based instructions; user can simply enter any of these or similar words and can look for relevant information e.g. "PDM", "CAD", and "Documents" etc.
- Short conditional objective statements; where user is looking for some information by just combining some keywords e.g. "CAD Designs" and "PDM Documents" etc.
- Simple conditional objective statements; where user is looking for some information by writing simple unconditional defined natural language grammar based instructions e.g. "I am looking for PDM Document", "I need CAD Designs" etc.
- Multiple Conditions based Objective statements; where user is looking for some information by writing defined natural language grammar based instructions with more than one conditions e.g. "I am looking for PDM Documents where Document Type is doc", "Give me CAD designs of Car parts"

Following context and rules of the grammar, lexer and parser are also written. Lexer consists of the following sets of tokens i.e., Digits, Numbers, Subject, Verb, Object, Blanks and Conditions.

- Digits are the numbers from 0 to 9

- Numbers are the combinations of digits e.g. 123 etc.
- Subject is the set of tokens representing Grammatical Person
- Verbs is a set of tokens representing Verbs and Helping verbs
- Object is a set of tokens representing
- Nouns the possible actual objects
- Blanks are the empty spaces
- Conditions are the tokens representing conditional words e.g. where, between, Equal, And, these words can be used to make conditional statements like "I am looking for CAD where document equals to Screw", "We are looking for Project details between Date 01-09-08 and 01-09-09", "I want Document where Author equal to Michael" and "I need Product and Project Document" etc.

Whereas Parser consist of the following sets of five direct search rules i.e. astmt, bstmt, cstmt, stmt1, stmt2 and four conditional search rules i.e., condbt, condeq, condweq. Condeqbt have been created which are as follows;

- rule astmt represents only Subject from in process natural language based search queries like "I, We etc."
- rule bstmt represent only Verbs and Helping verbs from in process natural language based search queries like "need, look, give, am, are etc.",
- rule cstmt represent only Objcet in process natural language based search queries, etc "document, person , project etc.".
- rule stmt1 is the combination of Subject, Verbs and Helping Verbs and Object, to analyze natural language based search queries like "I am looking for CAD document" etc.
- rule stmt2 is the combination of Verbs and Helping Verbs and Object to analyze natural language based search queries like "give CAD desing" etc.
- rule condbt is the combination of Subject, Verbs and Helping Verbs, Object and Condition "between" e.g " I am looking for CAD Design between Number 100 and 200" etc.
- rule condeq is the combination of Subject, Verbs and Helping Verbs, Object and Condition "equal" e.g. " I need CAD with name MotorEngine and type BMP" etc.
- rule condweq is the combination of Subject, Verbs and Helping Verbs, Object and Conditions "with" and "equal" e.g. "I want Project with PDM name PDMDatabase" etc.
- rule condqbt is the combination of Subject, Verbs and Helping Verbs, Object and Conditions "with", "equal" and "between" e.g. "looking for a Project where PDMDatabase name is between 2000 to 2009 Date" etc.

After having a grammar, the overall job of NLP Search module is divided into five steps .i.e., data reading, tokenization, parsing, semantic modeling and query generation, each step requires intensive effort in design and development. The main concept behind the organization of these five steps is to first read the user input natural language based instruction and to understand the semantic hidden in the context of natural language based instruction by lexing and parsing it. Then generate a query (SQL) to extract the desired results from attached repository. The implementation designs are constructed and discussed for prototype implementation of NLP Search using the concepts and technologies for Language Technology. The grammar is written using Antlr, lexer, parser and rest of the NLP Search module is developed by constructing implementation designs, and with the user of Java programming language.

## 6. Implementation Designs of Proposed Approach

To implement the proposed approach as a prototype (software) application, taking advantage from the observed information as the result of conducted research in the field of PDM System Development, I have designed a classical tier architecture consisting of three layers i.e., Presentation Layer, Business Logic and Database, as shown in figure 8.

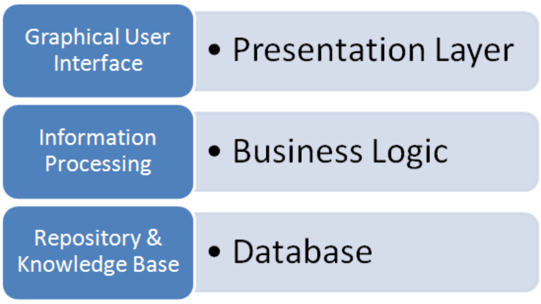


Fig. 8. Classical Three Tier Architecture

Presentation layer is the Graphical user interface of the prototype, carrying the jobs of user system communication. Business Logic is the information processing module of the prototype, carrying the jobs of transferring the user and system data between graphical user interface and database by implementing a communication system. Third layer is the Database, the main repository of the prototype, proposed to store, secure and manage data. This layer is the back end database management system, used to store and manage product attribute data and documentary information, as well as the relationships between data. This DBMS is usually a relational database system which provides complete functionalities to manage the product implemented using MySQL database management system.

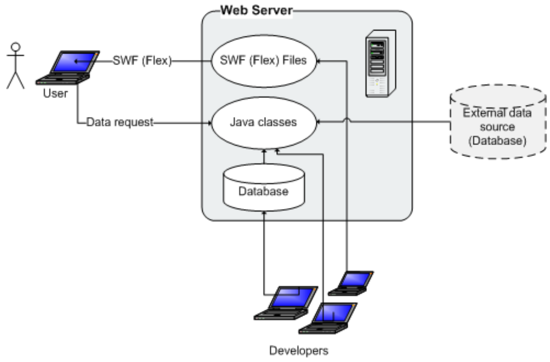


Fig. 9. Design Methodology

Following three tier application model of proposed approach, I have designed implementation methodology for the development of proposed prototype, as shown in Figure 9. The current version of proposed approach will be implanted with the use of Java (servlets and JSP) to handle user input, manage and retrieve data from the database. Tomcat is used as the main web server and middleware of the program. Users can access the web pages with the given URL and then can build graphical user interface or search the data after successful identity authorization. The data communication between three tiers is managed by Action Message Format (AMF) using the Simple Object Access Protocol (SOAP). AMF based client requests are delivered to the web server using Remote Procedure Call (RPC). The use of RPC allows presentation tier to directly access methods and classes at the server side. When data is request from user then a remote call is made from the user interface in the remote services' (via the server side includes) class members and the result is sent as an object of a Java class. A web browser is mainly needed to access the developed application with a user of a specified universal resource link (URL). User will send a request to the web server through Hypertext Transfer Protocol (HTTP), the web server will pass the request to the application components. These application components are implemented using servlet/JSP, designed to handle user request coming from web server with the use of java remote classes. Then used servlets or JSP classe talks to the database server, perform the data transactions and send the response to the client. To increase flexibility of graphical user interface at client end, the development of front end is performed using Flex Flex (Builder 3 IDE), earlier discussed in section section 3.2. Relational database is designed and implemented using MySQL 5.

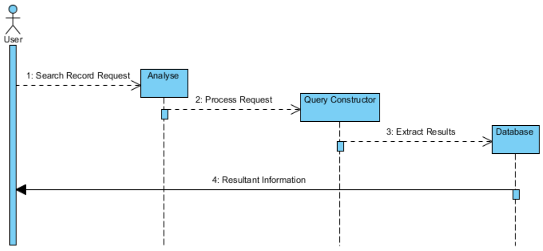
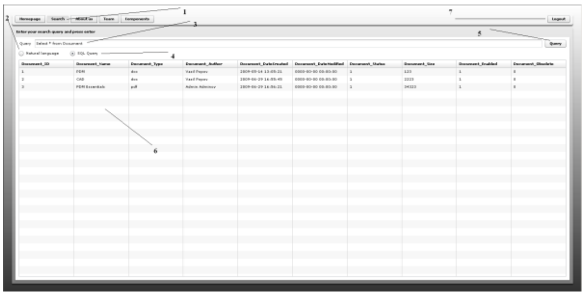


Fig. 10. System Sequence Design

As shown in figure 10, the System Sequence design consists of the three main components .i.e., Analyze, Query Constructor and Database. These three components perform certain jobs, the job of Analyze is to check if the it's a natural language based required from user to search record, then Query Constructor creates a SQL query to extract data from database and provide that to the user. Process and Model Information component first will store the information in to Repository, then will process the information by lexing, parsing and semantically modeling In the end Process and Model Information component will first save the information in Repository and then return the final system output to Graphical Interface.

## 7. NLP Search Prototype

This is the search module of developed Web Application. It can be accessed using Search Link from the links on the main. This prototype version is NLP Search, residing with in the limited scope, providing a search mechanism capable of processing natural language based user requests to extract desired information. The over all job of this module is explained in table 1 and shown in Figure 11 and 12.



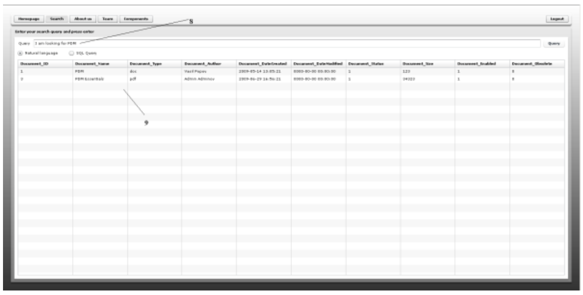
Enter your search query and press enter

Query: Select * from Document

○ Natural language ● SQL Query

| Document_ID | Document_Name | Document_Type | Document_Author | Document_DateCreated | Document_DateModified | Document_Status | Document_Size | Document_Enabled | Document_Obsolete |
|---|---|---|---|---|---|---|---|---|---|
| 1 | PDM | doc | Vasil Popov | 2009-05-14 13:05:21 | 0000-00-00 00:00:00 | 1 | 123 | 1 | 0 |
| 2 | CAD | doc | Vasil Popov | 2009-06-29 16:05:45 | 0000-00-00 00:00:00 | 1 | 2223 | 1 | 0 |
| 3 | PDM Essentials | pdf | Admin Adminov | 2009-06-29 16:06:21 | 0000-00-00 00:00:00 | 1 | 34323 | 1 | 0 |

Fig. 11. I-SOAS Prototype- SQL based Search



Enter your search query and press enter

Query: I am looking for PDM

● Natural language ○ SQL Query

| Document_ID | Document_Name | Document_Type | Document_Author | Document_DateCreated | Document_DateModified | Document_Status | Document_Size | Document_Enabled | Document_Obsolete |
|---|---|---|---|---|---|---|---|---|---|
| 1 | PDM | doc | Vasil Popov | 2009-05-14 13:05:21 | 0000-00-00 00:00:00 | 1 | 123 | 1 | 0 |
| 2 | PDM Essentials | pdf | Admin Adminov | 2009-06-29 16:56:21 | 0000-00-00 00:00:00 | 1 | 34323 | 1 | 0 |

Fig. 12. I-SOAS Prototype- Natural Language based Search

## 8. Comparison with CIM and Windchill

Residing within the scope, meeting the goals of this research work and to evaluate the effectiveness of NLP Search module, some comparisons has been performed amongst developed Prototype, Windchill and CIM. During comparison earlier discussed four kinds of natural language search instructions i.e., Keyword based instructions, Short conditional objective statements, Simple conditional objective statements and Multiple Conditions based Objective statements, have been compared with the existing search module of CIM Database.

### *8.1. CIM and Prototype Comparison*

In this section the search module of CIM Database has been compared with the search module of implemented prototype version. During comparison earlier discussed four kinds of natural language search instructions i.e., Keyword based instructions, Short conditional objective statements, Simple conditional objective statements and Multiple Conditions based Objective statements, have been compared with the existing search module of CIM Database.

### *8.1.1. Keyword based instruction based comparison*

A search was made with only one keyword

"PDM"

using both the search module of CIM Database as shown in Fig. 13 and prototype as shown in Fig. 14. As the result, both the applications performed accurately and the resultant information was presented by both systems, which proves that both the systems are equally efficient while making one keyword based search.

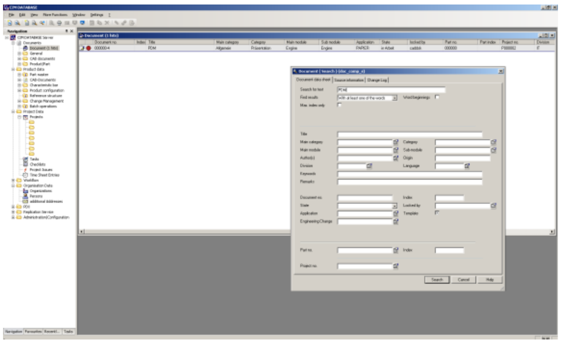
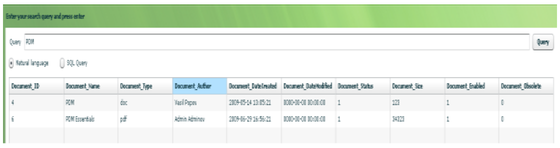


Fig. 13. CIM Database– Keyword based Search



Enter your search query and press enter

Query: PDM

● Natural language ○ SQL Query

| Document_ID | Document_Name | Document_Type | Document_Author | Document_DateCreated | Document_DateModified | Document_Status | Document_Size | Document_Enabled | Document_Obsolete |
|---|---|---|---|---|---|---|---|---|---|
| 4 | PDM | doc | Vasil Popov | 2009-05-14 13:05:21 | 0000-00-00 00:00:00 | 1 | 123 | 1 | 0 |
| 6 | PDM Essentials | pdf | Admin Adminov | 2009-06-29 16:56:21 | 0000-00-00 00:00:00 | 1 | 34323 | 1 | 0 |

Fig. 14. Prototype – Keyword based Search.

### *8.1.2. Short Objective Statement based comparison*

A search was made with short objective statement,

"want PDM"

using both the search module of CIM Database as shown in Fig. 15and Prototype as shown in Fig.16. The resultant Fig. 15 of CIM Database clearly shows that the CIM Database was unable to find any output even when the result existed, moreover CIM Database search doesn't allow user to enter small alphabets because it is case sensitive. Whereas the resultant Fig. 16 of Prototype NLP Search module is presenting the obtained results. This comparison clearly proves that the text based search mechanism of CIM Database is unable to compile short objective statement whereas the Prototype NLP Search module successfully compiled short objective statement and presented the results.

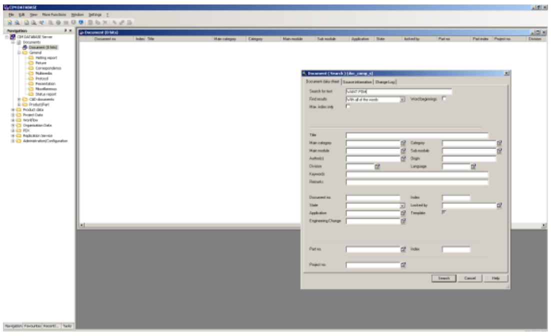
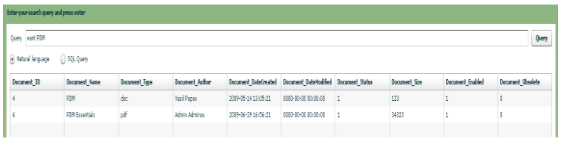


Fig. 15. CIM Database – Short Objective Statement based Search

.



Enter your search query and press enter

Query: want PDM [Query]

● Natural language ○ SQL Query

| Document_ID | Document_Name | Document_Type | Document_Author | Document_DateCreated | Document_DateModified | Document_Status | Document_Size | Document_Enabled | Document_Obsolete |
|---|---|---|---|---|---|---|---|---|---|
| 4 | PDM | doc | Vasil Popov | 2009-05-14 13:05:21 | 0000-00-00 00:00:00 | 1 | 123 | 1 | 0 |
| 6 | PDM Essentials | pdf | Admin Adminov | 2009-06-29 16:56:21 | 0000-00-00 00:00:00 | 1 | 34323 | 1 | 0 |

Fig. 16. Prototype– Short Objective Statement based Search.

### *8.1.3. Simple Objective Statement based comparison*

A search was made with simple objective statement,

“I am looking for PDM”

using both the search module of CIM Database as shown in Fig. 17 and Prototype as shown in Fig. 18. The resultant Fig. 17 of CIM Database clearly shows that the CIM Database was unable to find any output even when the result existed, whereas the resultant Fig. 18 of Prototype NLP Search module is presenting the results. This comparison clearly proves that the text based search mechanism of CIM Database is unable to compile simple objective statement whereas the Prototype NLP Search module successfully compiled simple objective statement and presented the results.

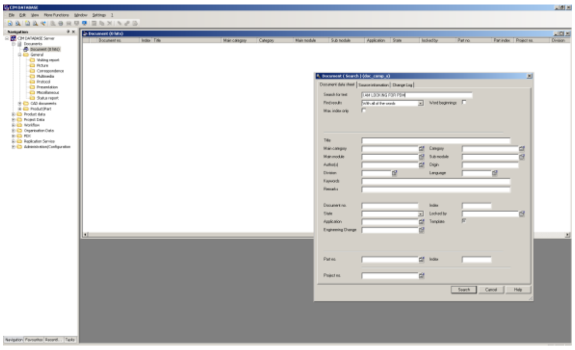
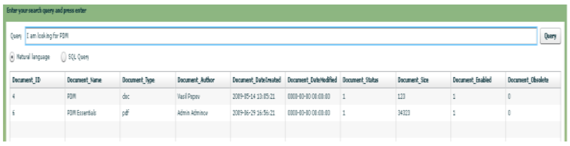


Fig. 17. CIM Database – Simple Objective Statement based Search.



Enter your search query and press enter

Query: I am looking for PDM [Query]

● Natural language ○ SQL Query

| Document_ID | Document_Name | Document_Type | Document_Author | Document_DateCreated | Document_DateModified | Document_Status | Document_Size | Document_Enabled | Document_Obsolete |
|---|---|---|---|---|---|---|---|---|---|
| 4 | PDM | doc | Vasil Popov | 2009-05-14 13:05:21 | 0000-00-00 00:00:00 | 1 | 123 | 1 | 0 |
| 6 | PDM Essentials | pdf | Admin Adminov | 2009-06-29 16:56:21 | 0000-00-00 00:00:00 | 1 | 34323 | 1 | 0 |

Fig. 18. Prototype – Simple Objective Statement based Search

### *8.1.4. 4. Multiple Conditions based Objective Statement based comparison*

A search was made with multiple conditions based objective statement using both the search module of CIM Database as shown in Fig. 19 and Prototype NLP Search as shown in Fig. 20. This time the search module of both the applications was compared with the same kind of search methodology (natural language based full text search) like in previous comparisons. Because we have already concluded from previous comparisons that the natural language based full text search module of CIM Database is unable to compile short and simple objective statements to search the information, so it would be unwise to expect CIM Database search module to compile natural language based multiple conditions based objective statements.

So this time the provided search form based options were used to find some results by filling options with relevant information (conditional information). As shown in Fig. 19 that there are several different options on the Document search form of CIM Database to find out the information about a document e.g. with some specific type or category etc. All required is to fill the search form by entering or selecting required information e.g. type as shown in Fig. 19. Whereas as shown in Fig. 20, in case of Prototype NLP Search module you just need to enter the natural language based search query .e.g.

“I am looking for PDM where Document Type is doc”

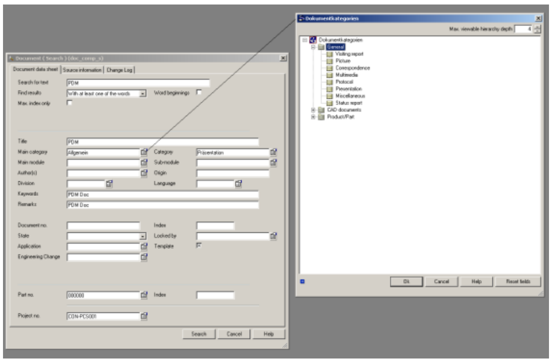
and the relevant result will appear. In this comparison both the search modules of the applications bring results but the major difference is not of the results but of the ease in use of the software. In case of CIM Database a user at first needs to open relevant search form e.g. document search or product search etc. and then has to enter relevant information in respective text boxes and then might also need to open some other forms and select some other information e.g. category etc. and then he can make a search. But in case of Prototype NLP Search instead of providing all this information user only needs to enter a simple natural language based multi conditional statement, which not only provides the ease in use but also reduces the time and effort of the user in searching required information.

As earlier mentioned that the Prototype NLP Search module also allows a professional database user to enter the SQL based query of his own choice to search the results, Fig. 21 presents the obtained results using SQL query

“select * from document where Document_Type = "doc" and Document_Name = "PDM"”

The resultant information not only satisfies the user but also confirms the result of above used natural language based multiple condition based object statement. This search with Multiple Conditions based Objective statement comparison proves that the required effort needed to search multi conditional information is very little in

Prototype NLP Search module as compared to the CIM Database search module.



Document ( Search ) (doc_comp_s)

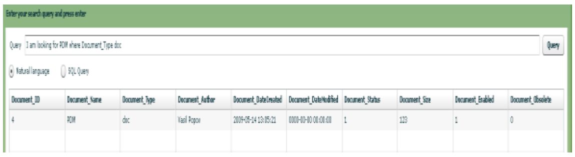
Document data sheet | Source information | Change Log

Search for text: PDM
Find results: With at least one of the words
Word beginnings: □
Max index only: □
Title: PDM
Main category: Allgemein
Category: Präsentation
Main module:
Sub-module:
Author(s):
Origin:
Division:
Language:
Keywords: PDM Doc
Remarks: PDM Doc
Document no.:
Index:
State:
Locked by:
Application:
Template: ☒
Engineering Change:
Part no.: 000000
Index:
Project no.: CON-PCS001

Dokumentkategorien

Max. viewable hierarchy depth: 4

Fig. 19. CIM Database – Multiple Conditions Search



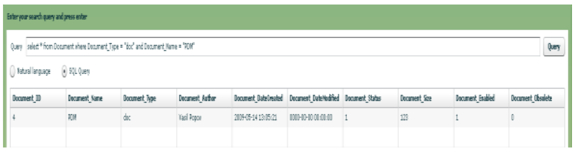
Enter your search query and press enter

Query: I am looking for PDM where Document_Type doc

● Natural language ○ SQL Query

| Document_ID | Document_Name | Document_Type | Document_Author | Document_DateCreated | Document_DateModified | Document_Status | Document_Size | Document_Enabled | Document_Obsolete |
|---|---|---|---|---|---|---|---|---|---|
| 4 | PDM | doc | Vasil Popov | 2009-05-14 13:05:21 | 0000-00-00 00:00:00 | 1 | 123 | 1 | 0 |

Fig. 20. Prototype – Multiple Conditions based Search



Enter your search query and press enter

Query: select * from Document where Document_Type = "doc" and Document_Name = "PDM"

○ Natural language ● SQL Query

| Document_ID | Document_Name | Document_Type | Document_Author | Document_DateCreated | Document_DateModified | Document_Status | Document_Size | Document_Enabled | Document_Obsolete |
|---|---|---|---|---|---|---|---|---|---|
| 4 | PDM | doc | Vasil Popov | 2009-05-14 13:05:21 | 0000-00-00 00:00:00 | 1 | 123 | 1 | 0 |

Fig. 21. Prototype – SQL Search Query based Conditional Search

## *8.2. Windchill and Prototype Comparison*

In this section of paper the search module of Windchill has been compared with the search module of implemented prototype version of Prototype NLP Search. Likewise CIM Database's comparison with developed Prototype, I have compared four kinds of natural language based search instructions i.e., Keyword based instructions, Short conditional objective statements, Simple conditional objective statements and Multiple Conditions based Objective statements, with the existing search module of Windchill and Prototype NLP Search.

### *8.2.1. 1. Keyword based instruction based comparison*

A search was made with only one keyword,

"CAD"

using both the search module of Windchill as shown in Fig. 22 and Fig. 23, and Prototype NLP Search as shown in Fig. 24. As show in figure Fig. 22 when the keyword CAD is entered in search text box of the main Home page of the Windchill and searched, a new page "Advanced Search" is appeared as shown in Fig. 23. The following result shown in Fig. 23 is obtained when CAD keyword is searched using advanced search page. Likewise as shown in the Fig. 24, when the keyword CAD is searched using Prototype NLP Search module then the CAD information is obtained. As the result both the applications performed accurately and the resultant information was presented by both systems, which proves that both the systems are equally efficient while making one keyword based search.

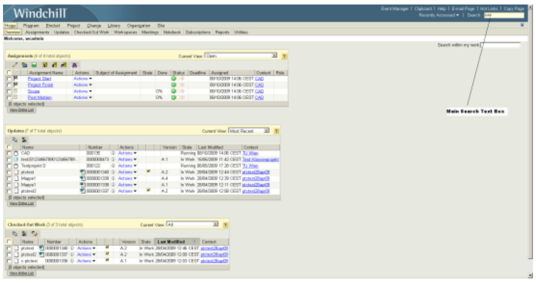


Fig. 22. Windchill Search –Main

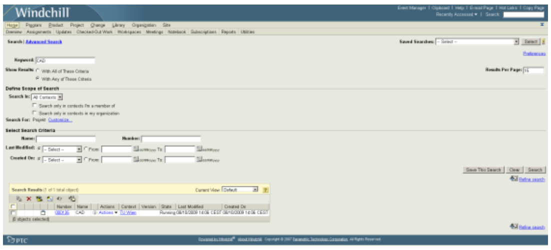
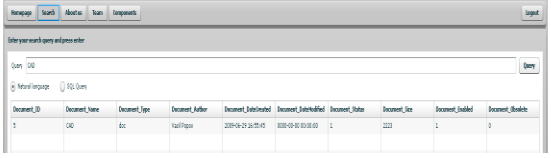


Fig. 23. Windchill – keyword based Search



Enter your search query and press enter

Query: CAD

● Natural language ○ SQL Query

| Document_ID | Document_Name | Document_Type | Document_Author | Document_DateCreated | Document_DateModified | Document_Status | Document_Size | Document_Enabled | Document_Obsolete |
|---|---|---|---|---|---|---|---|---|---|
| 5 | CAD | doc | Vasil Popov | 2009-06-29 16:55:45 | 0000-00-00 00:00:00 | 1 | 2223 | 1 | 0 |

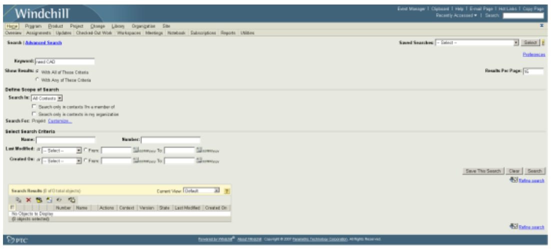
Fig. 24. Prototype – Keyword based Search

### *8.2.2. Short Objective Statement based comparison*

A search was made with short objective statement,

"need CAD"

using both the search module of Windchill as shown in Fig. 25 and Prototype NLP Search as shown in Fig. 26. The resultant Fig. 25 of Windchill clearly shows that likewise CIM Database Windchill is also unable to find any output even when the result existed. Whereas the Fig. 26 of Prototype NLP Search module presents the results. This comparison clearly proves that the text based search mechanism of Windchill is unable to compile short objective statement whereas the Prototype NLP Search module successfully compiles short objective statement and gives the results.


Windchill
Event Manager | Clipboard | Help | E-mail Page | Hot Links | Copy Page
Recently Accessed ▾ | Search:
Home Program Product Project Change Library Organization Site
Overview Assignments Updates Checked-Out Work Workspaces Meetings Notebook Subscriptions Reports Utilities
Search | Advanced Search
Saved Searches: -- Select --
Select
Preferences
Save This Search
Clear
Search
Refine search
PTC
Powered by Windchill® About Windchill Copyright © 2007 Parametric Technology Corporation. All Rights Reserved.

Keyword: need CAD

Show Results: ● With All of These Criteria ○ With Any of These Criteria

Results Per Page: 15

Define Scope of Search

Search In: All Contexts

□ Search only in contexts I'm a member of
□ Search only in contexts in my organization

Search For: Project Customize...

Select Search Criteria

Name: Number:

Last Modified: ● -- Select -- ○ From: mm/dd/yyyy To: mm/dd/yyyy

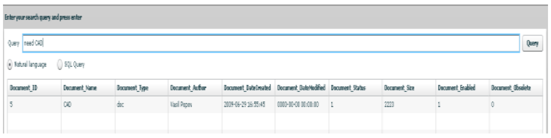
Created On: ● -- Select -- ○ From: mm/dd/yyyy To: mm/dd/yyyy

Search Results (0 of 0 total objects) Current View: Default

□ | Number | Name | Actions | Context | Version | State | Last Modified | Created On

No Objects to Display
(0 objects selected)

Fig. 25. Windchill – Short Objective Statements



Enter your search query and press enter

Query: need CAD

● Natural language ○ SQL Query

| Document_ID | Document_Name | Document_Type | Document_Author | Document_DateCreated | Document_DateModified | Document_Status | Document_Size | Document_Enabled | Document_Obsolete |
|---|---|---|---|---|---|---|---|---|---|
| 5 | CAD | doc | Vasil Popov | 2009-06-29 16:55:45 | 0000-00-00 00:00:00 | 1 | 2223 | 1 | 0 |

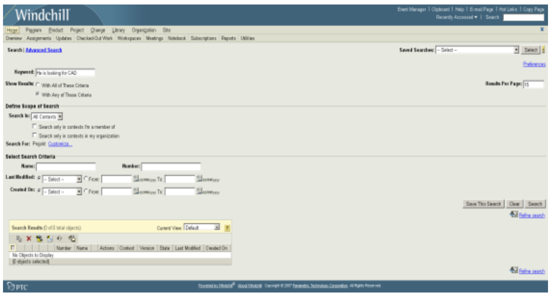
Fig. 26. Prototype Short Objective Statements based Search

### *8.2.3. Simple Objective Statement based comparison*

A search was made with simple objective statement,

"He is looking for CAD"

using both the search module of Windchill as shown in Fig. 27 and Prototype NLP Search as shown in Fig. 28. The resultant Fig. 27of Windchill clearly shows that Windchill is again unable to find any output using simple objective statement with using both the options of text i.e., With All of These Criteria and With Any of These Criteria, even when the result existed whereas the resultant Fig. 28 of Prototype NLP Search module presents the results. This comparison clearly proves that the text based search mechanism of Windchill is unable to compile simple objective statement whereas the Prototype NLP Search module successfully compiles simple objective statement and presents the results.


Windchill
Event Manager | Clipboard | Help | E-mail Page | Hot Links | Copy Page
Recently Accessed ▾ | Search:
Home Program Product Project Change Library Organization Site
Overview Assignments Updates Checked-Out Work Workspaces Meetings Notebook Subscriptions Reports Utilities
Search | Advanced Search
Saved Searches: -- Select --
Select
Preferences
Save This Search
Clear
Search
Refine search
PTC
Powered by Windchill® About Windchill Copyright © 2007 Parametric Technology Corporation. All Rights Reserved.

Keyword: He is looking for CAD

Show Results: ○ With All of These Criteria ● With Any of These Criteria

Results Per Page: 15

Define Scope of Search

Search In: All Contexts

□ Search only in contexts I'm a member of
□ Search only in contexts in my organization

Search For: Project Customize...

Select Search Criteria

Name: Number:

Last Modified: ● -- Select -- ○ From: mm/dd/yyyy To: mm/dd/yyyy

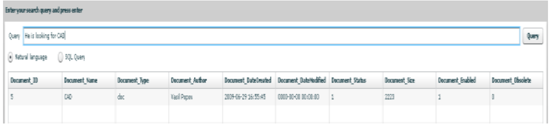
Created On: ● -- Select -- ○ From: mm/dd/yyyy To: mm/dd/yyyy

Search Results (0 of 0 total objects) Current View: Default

□ | Number | Name | Actions | Context | Version | State | Last Modified | Created On

No Objects to Display
(0 objects selected)

Fig. 27. Windchill – Simple Objective Statement based Search



Enter your search query and press enter

Query: He is looking for CAD

● Natural language ○ SQL Query

| Document_ID | Document_Name | Document_Type | Document_Author | Document_DateCreated | Document_DateModified | Document_Status | Document_Size | Document_Enabled | Document_Obsolete |
|---|---|---|---|---|---|---|---|---|---|
| 5 | CAD | doc | Vasil Popov | 2009-06-29 16:55:45 | 0000-00-00 00:00:00 | 1 | 2223 | 1 | 0 |

Fig. 28. Prototype NLP Search Query – Simple objective statement

### *8.2.4. Multiple Conditions based Objective Statement based comparison*

A search was made with multiple conditions based objective statement using both the search modules of Windchill as shown Fig. 29 and Prototype NLP Search as shown in Fig. 31. Like the conditions based objective statement comparison of CIM Database and Prototype NLP Search, here again this time I am not comparing both application's the search module with same kind of search methodology like I did in previous comparisons. As we have already concluded from previous comparisons that the Search module of Windchill is unable to compile short and simple objective statements to search the information, so this time I am not comparing search statements. But I am using search form based options to find some results by filling options with relevant information (conditional information).

As shown in Fig. 29 that there are some options on the advanced search form of Windchill to find out the information. To obtain any information by making multiple conditions based search, user is required to first fill some form based options, and if options of user interest are not available in the search form then user must click on "Customized" option to first choose the relevant options to create appropriate search form, which is again a hectic task. For example user is interested in finding out a document named CAD with some specific type e.g. "doc" and "pdf". As shown in Fig.29, user starts the search using word name CAD but the resultant information is based on CAD project. If user is not satisfied with this output then he needs to first customize the search page as shown in Fig. 29. Moreover by looking at this search output a user might think that there is no such document exists with the name CAD, but as clearly shown in Fig. 30 that there is a document with name CAD in the system. Whereas in case of Prototype NLP Search module user just needs to enter the natural language based search query .e.g.

"She need PDM with Document Type doc and Pdf"

As shown in Fig. 31, and the relevant results are presented. This search with Multiple Conditions based Objective statement comparison proves that the required effort needed to search some multi conditional information is very little in Prototype

NLP Search module as compared to the Windchill search module. Moreover in case of Windchill, for making different kinds of search like searching documents, projects, products etc. each time advanced search form will appear, which requires user to train himself for performing different kinds of search consuming time and effort where as in case of Prototype NLP Search module, user does not need to learn to use multiple options based search form to carry out different searches, moreover he can use simple natural language based queries to search required information.

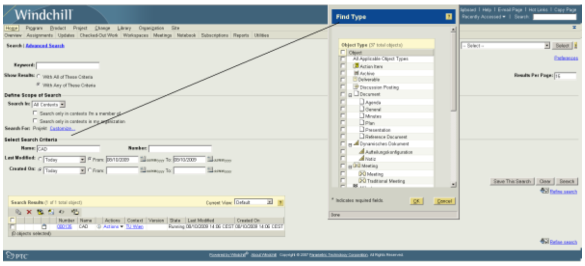


Fig. 29. Windchill – Multiple Conditions based Search

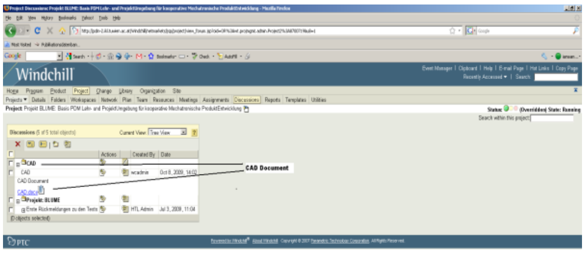
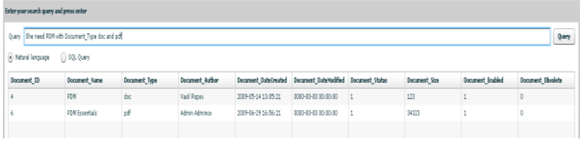


Fig. 30. Windchill – CAD Document



Query: She need PDM with Document_Type doc and pdf

| Document_ID | Document_Name | Document_Type | Document_Author | Document_DateCreated | Document_DateModified | Document_Status | Document_Size | Document_Enabled | Document_Obsolete |
|---|---|---|---|---|---|---|---|---|---|
| 4 | PDM | doc | Vasil Popov | 2009-05-14 13:05:21 | 0000-00-00 00:00:00 | 1 | 123 | 1 | 0 |
| 6 | PDM Essentials | pdf | Admin Adminov | 2009-06-29 16:56:21 | 0000-00-00 00:00:00 | 1 | 34323 | 1 | 0 |

Fig. 31. Prototype – Multiple Conditions based Search

## 9. Limitations

We have proposed and written a new grammar for natural language processor implementation for PDM Systems but due to the limited scope of this research and development work, the word length of proposed dictionary for lexer is small and the number of rules designed for parser are also a few. At the moment only seven different set of tokens for lexer i.e. Digits, Numbers, Subject, Verb, Objects, Blanks and Conditions, and five rules are defined for parser implementation.

## 10. Conclusion

Targeting the challenge of proposition of natural language based search, a thorough research has been conducted in the field of Language Technology and findings are presented in this paper. Taking help from observed information from conducted research in respective field and using person research and development experience I have proposed an approach. We have designed conceptual and implementation designs of proposed approach and implemented it using some software tools and technologies of present time i.e. Flex, Java, Antlr, MySQL, and presented developed prototype solutions. In the end concluding the research and development efforts, I can say that the proposed approach can put some values in enhancing PDM System development process. The inclusive implementation of the newly proposed idea in PDM System development can put some values in increasing the market values of PDM Systems by increasing its acceptability in industry by improving its use amongst managerial, technical and office staff.

## 11. Acknowledgments

I am thankful to University of Wuerzburg Germany and Vienna University of Technology Austria for giving me the opportunity to keep working on this research project. I am thankful to Prof. Dr. Detlef Gerhard for his supervision during this research and pay my gratitude to Prof. Dr. Thomas Dandekar for his generous support. I also thanks to my beloved wife Mrs. Saman Zeeshan (Majeed) for her support during this research, development and technical documentation.

## References

[1] Windchill, Reviewed 06 November2008<http://www.ptc.com/WCMS/files/56909/en/2757_Windchill_bro_ViewONLY.pdf>

[2] CIM DATABASE 2.9.5, User Manual by CONTACT Software GmbH

[3] Liddy, E.D. Natural Language Processing. In Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc. 2001

[4] Hans Uszkoreit, DFKI-LT - What is Language Technology?, Reviewed 03 September 2008, <http://www.dfki.de/lt/lt-general.php>

[5] Joseph Galasso, Analyzing English Grammar: An Introduction to Feature Theory - A Companion Handbook, California State University, Northridge, Draft May 5 2002,

[6] Bruce W. Ballard and John C. Lusth, An English-language processing system that "learns" about new domains, Pages 39-46, Year of Publication: 1983, ISBN ~ ISSN: 0095-6880, 0-88283-039-2, ACM New York, NY, USA

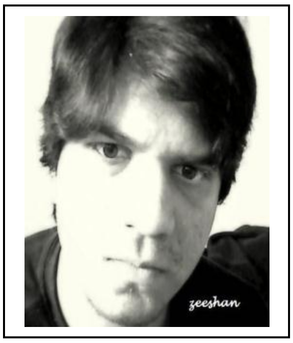
[7] ANTLR, Last reviewed 20 July 2009, <http://www.antlr.org/>

[8] C. S. Sung, Sam Joon Park, "A component-based product data management system", In Springer-Verlag London Limited 2006, Received: 18 April 2005 / Accepted: 25 October 2005/ / Published online: 2006

[9] Miller, E., "What's PMD ?" Computer- Aided Enineering Magzine, September 1997.

## Author Bibliography

Zeeshan Ahmed is a Software Research Engineer, presently working in the Department of Bioinformatics, Biocenter, University of Wuerzburg Germany. He has on record more than 12 years of University Education and more than 8 years of professional experience of working within different multinational organizations in the field of Computer Science with emphasis on software engineering of product line architecture based artificially intelligent systems. He also has more than 4 years experience of teaching as lecturer and supervising research thesis to graduate and undergraduate students in different institutes and universities.



# A short history of web based learning including GIS

Gilbert Ahamer

Austrian Academy of Science, Institute for GIScience,
gilbert.ahamer@oeaw.ac.at

***Abstract*:**

This paper suggests a short history of web based learning in three generations according to the usage of web based functionalities while presenting practical cases. The idea is to show how (1) content, (2) communication and (3) assessment have evolved in steps which are referred to as "generations of web learning". A fourth and a fifth step is proposed, making use of multi-perspectivism and geographic Information Systems (GIS).

The reader is offered a stepwise description of both didactic foundations of university lectures and a practical implementation of a widely available web platform. The relative weight of directive elements has gradually decreased through the "three generations", whereas characteristics of self-responsibility and self-guided learning have gained in importance.

1. Content was in early stages presented and expected to be learned but later on it was expected to be constructed for examples using case studies.

2. Communication meant in early stages delivering assignments to the lecturer but later on forming teams, exchanging standpoints and reviewing mutually.

3. Assessment initially consisted in marks invented and added up by the lecturer but was later enriched by peer review, mutual grading and voting procedures.

How much "added value" can the web provide for teaching, training and learning? Twelve years of experience suggest: mainly insofar as new (collaborative and self-directed) didactic scenarios are implemented.

***Keywords*:** e-learning history, .web-based learning, communication, content, assessment.

## 1. History of the "three initial generations of web based learning"

The target of this paper is to compare several strategies of assessing students' academic performance in cases where there is "more than one truth".

This text discerns three phases of web based teaching / training / learning (WBT) according to how didactic objectives and concepts are transposed (Bork, 2001, Prensky, 2001). Over the last years e-learning activities have increasingly made use of technological possibilities offered by current web platforms. In a number of cases, this enabled strife for student-centered and problem-based learning. Earlier work of the author is taken as an example for defining the three "generations" of web based learning (see Fig. 1).

## 2. Three initial generations of web support in practical examples

### 2.1. First generation: content and quiz

Very often, "putting one's lecture onto the web" means in practice to provide students with written online documents which replace many printed pages. Such content-centered understanding of "web based teaching" intrigues lecturers due to the decrease of administrative work that is expected as a result of pasting a link to a PDF file into an existing university web page. Such an approach might recall former times.

As a case for the 1st generation, since 1999 three *interdisciplinary* courses are held at an Austrian University of Applied Science (FH Joanneum FHJ), namely "Technology Assessment", "Systems Theory and Biology", and "Environmental Technology" (see cover pages in Fig. 2).

With the kind and helpful initial support of FHJ's Centre for Multimedia and Learning (CML) and its founder, several *functionalities* of the then newly acquired web platform Web Course Tools (WebCT, 2004) were employed in order to

1. present content to students and to allow students to study independently of time and place (Lo et al., 1999)
2. assess students' specific interests and preferences at the outset of the lecture in an "initial survey"
3. provide several case studies as topics for students' written assignments, allowing for differentiated anonymous personal choice
4. provide a discussion forum, where individual students could submit their resulting essays and where they would receive the lecturer's evaluation
5. require traditional results of cognitive learning (quiz equaling the written exam) and inform about exam results
6. ask students for their overall feedback after the end of the courses in a "final survey",

which is graphically represented in the left part of
Fig. **5**.

*Content* provided on the web platform was hierarchically structured into

1. one list of links representing the table of contents of the course
2. a set of 50 transparencies (in doc file format) used for face-to-face teaching
3. a multitude of 100's of text files and links covering details of all subject matters

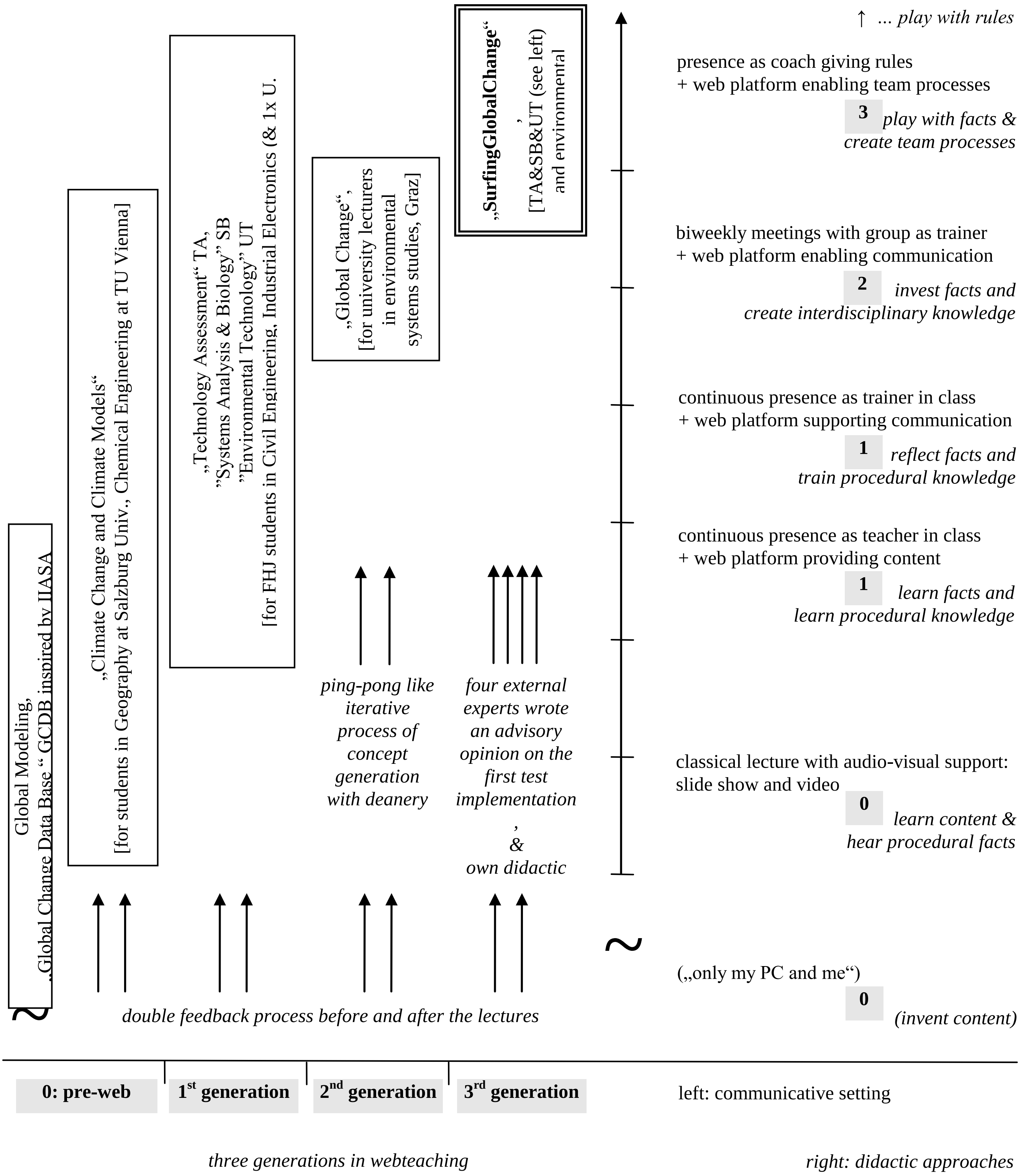

Global Modeling,
„Global Change Data Base “ GCDB inspired by IIASA
„Climate Change and Climate Models“
[for students in Geography at Salzburg Univ., Chemical Engineering at TU Vienna]
„Technology Assessment“ TA,
”Systems Analysis & Biology” SB
”Environmental Technology” UT
[for FHJ students in Civil Engineering, Industrial Electronics (& 1x U.
„Global Change“,
[for university lecturers
in environmental
systems studies, Graz]
„SurfingGlobalChange“
,
[TA&SB&UT (see left)
and environmental
ping-pong like iterative process of concept generation with deanery
four external experts wrote an advisory opinion on the first test implementation , & own didactic
double feedback process before and after the lectures
~
~
↑ ... play with rules
presence as coach giving rules
+ web platform enabling team processes
3 play with facts & create team processes
biweekly meetings with group as trainer
+ web platform enabling communication
2 invest facts and create interdisciplinary knowledge
continuous presence as trainer in class
+ web platform supporting communication
1 reflect facts and train procedural knowledge
continuous presence as teacher in class
+ web platform providing content
1 learn facts and learn procedural knowledge
classical lecture with audio-visual support: slide show and video
0 learn content & hear procedural facts
(„only my PC and me“)
0 (invent content)
0: pre-web
1st generation
2nd generation
3rd generation
left: communicative setting
three generations in webteaching
right: didactic approaches

**Fig. 1:** History of three generations of web based learning as based on the author’s earlier scientific work starting from the “Global Change Data Base” GCDB. Years indicate summer semesters; generations indicate steps in implementing communicative structures; arrows denote inputs; the right hand side shows the conceptual basis (communication and didactics).

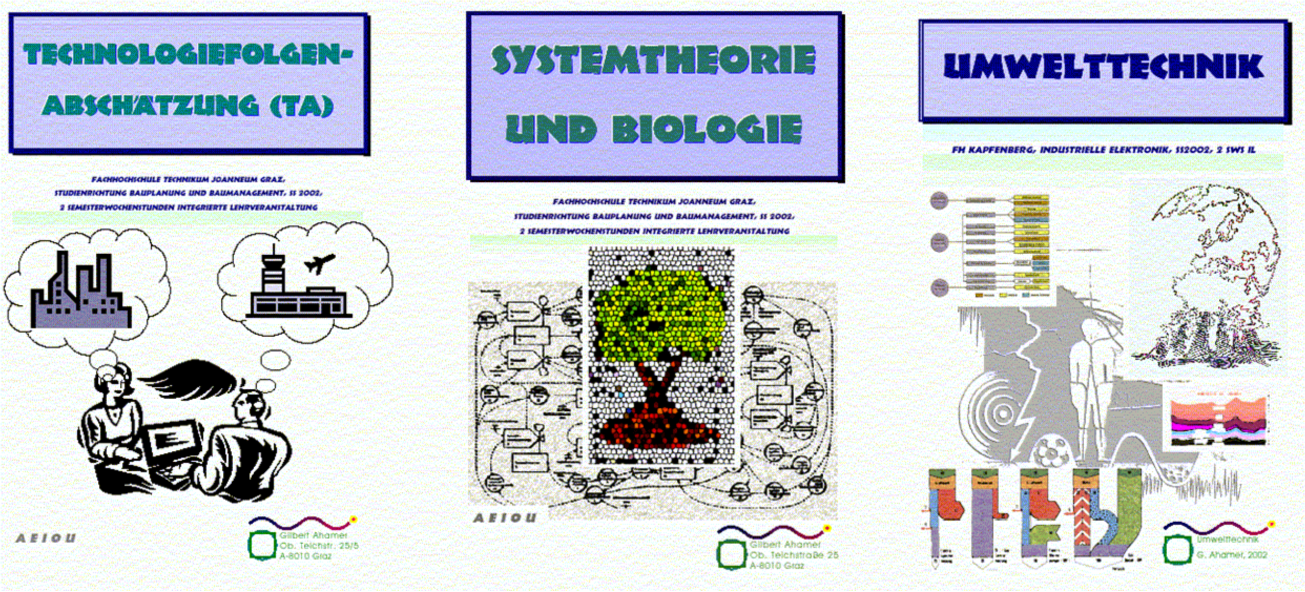

TECHNOLOGIEFOLGEN-ABSCHÄTZUNG (TA)
FACHHOCHSCHULE TECHNIKUM JOANNEUM GRAZ, STUDIENRICHTUNG BAUPLANUNG UND BAUMANAGEMENT, SS 2002, 2 SEMESTERWOCHENSTUNDEN INTEGRIERTE LEHRVERANSTALTUNG
AEIOU
Gilbert Ahamer Ob. Teichstr. 25/5 A-8010 Graz
SYSTEMTHEORIE UND BIOLOGIE
FACHHOCHSCHULE TECHNIKUM JOANNEUM GRAZ, STUDIENRICHTUNG BAUPLANUNG UND BAUMANAGEMENT, SS 2002, 2 SEMESTERWOCHENSTUNDEN INTEGRIERTE LEHRVERANSTALTUNG
AEIOU
Gilbert Ahamer Ob. Teichstraße 25 A-8010 Graz
UMWELTTECHNIK
FH KAPFENBERG, INDUSTRIELLE ELEKTRONIK, SS2002, 2 SWS IL
Umwelttechnik G. Ahamer, 2002

**Fig. 2:** First generation 1999: Three cover pages representing content delivered to students via both a web platform and a paper manuscript: “Technology Assessment” (TA), “Systems Analysis and Biology” (SB) and “Environmental Technology” (UT) at FH Joanneum over a period of six years. Source: Ahamer (1999).

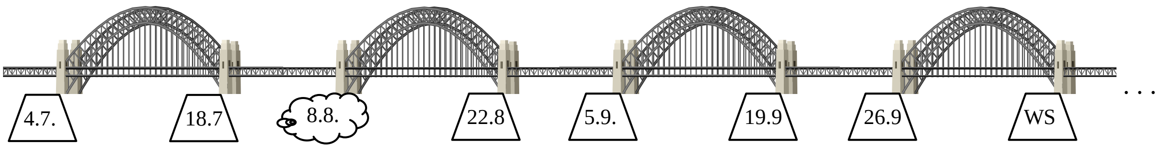


**Fig. 3:** Second generation 2002: time structure of 8 face-to-face meetings with online phases in between. Only one real meeting was replaced by a virtual one (cloud).

The *final grade* for these three courses (TA, SB, UT) consists of several components (
Fig. **5** left) that reflect both cognitive and creative abilities of the students, namely the

1. individual written online exam administered during lecture time in class while being supervised by the lecturer (max. 30 or 50 points for compulsory share plus max. 20 points for optional share)
2. “short” case study (1 page/person) on a general topic like ethics (written and oral performance); in earlier years with an oral presentation in class and in later years with directed mutual peer reviews among students via a platform
3. “long” case study (5 pages/team) as preparation for role-play in class representing a negotiation of a construction project as used in the Environmental Impact Assessment (EIA, 1997 and EIA, 2000).

Teaching occurred face-to-face because at that early stage no administrative high-level support for tele-teaching seemed realistic. Also, all three lectures had a strong component of individuals’ aims and of the ethical orientation that seemed to necessitate personal contact. In line with the experience of the author, here web tools played at best a supportive role. Later on the term “*blended learning*” was coined for such a combined teaching style.

### 2.2. Second generation: communication and construction

After four years of such a relatively simple architecture in web teaching, any interested actor would have felt a notable increase in

1. general awareness of didactic implications, e.g. by activities in the Austrian “Forum Neue Medien” (BMBWK, 2000) or in individual universities (NML, 2002)
2. community-building among web-trainers, e.g. three informal Austrian meetings on web didactics and seminars (Gierlinger, 2002) organized by the author
3. structures for professional formation (e.g. the multiple course schedules “Train-the-Trainer”) organized by FH Joanneum and others (CML, 2002).

In order to push ahead the target percentage of realistically implemented “web based training”, the vice-deanery at Graz University tasked the author with holding a summer course from July to October 2002 with *three distinct objectives:*

1. to train university teachers to utilize the WebCT platform
2. to create samples of online course material for later usage

3. to train lecturers in interdisciplinary collaboration.

The course *schedule* envisioned one face-to-face meeting every second week and online work in between (Ahamer, 2002; Ahamer and Carstensen, 2002), akin to a bridge with pillars (Fig. 3).

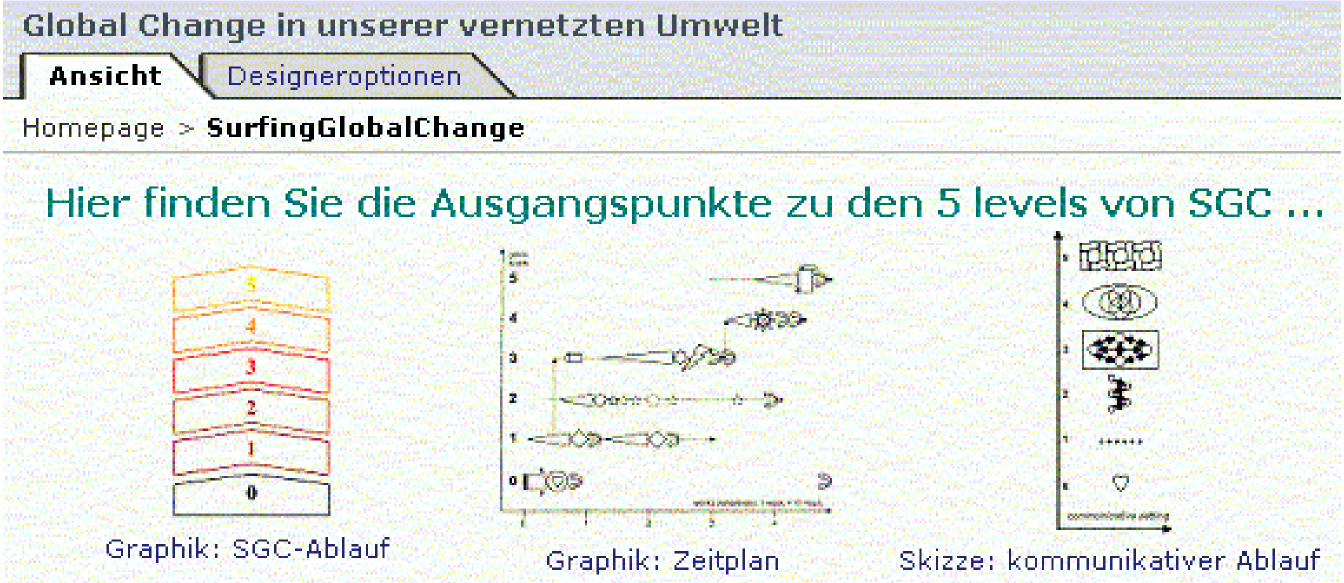
The architecture (Ahamer, 2002) comprised *6 phases* of ca. two weeks each:

1. concept and media (kick-off meeting, team building and planning)
2. collection of materials (creation of content pool and mutual commenting)
3. didactic processing of materials and condensing into web media (90min/team)
4. trial and evaluation (mutual teaching as test, subsequent documentation)
5. analysis and revision (reworking of web media, mutual commenting)
6. an entire interdisciplinary course is implemented in team teaching.

How well were initial objectives attained? *Evaluation* is of essential value (Barz et al., 1997; Carstensen & Reissert, 1997). A critical reflection and monitoring (Carstensen, 2002) states that 12 weeks time is too short for three ambitious goals. Encountered difficulties (like different activity level, high time consumption, decrease in motivation) are believed to be typical for future web teaching implementations by course members. In the view of the author, targets were reached according to Tab. 1.

**Tab. 1:** Monitoring of the degree to which the targets of the summer course have been reached according to the personal view of the author.

| **target according to initial concept of summer course** | **attainment of target after course** |
|---|---|
| usage of web platform for communication | 85% ☺ |
| authoring of concept and scenario for lecture | 80% ☺ |
| generation of module of web content | 80% ☺ |
| collaboration (independent of time and space) | 80% ☺ |
| technically mastering WebCT | 75% 😐 |
| didactic sense for implementation of web based training | 75% 😐 |
| team generation and group formation | 70% 😐 |
| interdisciplinary dialogue inside the teams | 60% 😐 |
| usage & interpretation of the Global Change Data Base | 30% ☹ |
| interdisciplinary dialogue between the teams | 30% ☹ |



Global Change in unserer vernetzten Umwelt

Homepage > **SurfingGlobalChange**

Hier finden Sie die Ausgangspunkte zu den 5 levels von SGC ...

Graphik: SGC-Ablauf

Graphik: Zeitplan

Skizze: kommunikativer Ablauf

**Fig. 4:** Third generation 2003-2005: Welcome screen of SurfingGlobalChange SGC.

The *iterative character* of the course and its successors comprises the years 2002-05:

1. the trainees of the first step (= summer course 2002) build up the structure of a web based "interdisciplinary course for Environmental Systems Sciences" (IPK-USW) in 2002/03
2. this course comprising 6 weekly hours is implemented via WebCT (Ahamer et al., 2002); students are required to merge technological, ecological and economic views and produces a number of written and reflected standpoints by using the game "SurfingGlobalChange"
3. innovative students from this first course propose a second implementation of SGC with different case studies focusing on the EU enlargement process (Florian, 2004). Thus the web based material will be annually expanded.

### 2.3. Third generation: collaboration and mutual assessment

Based on experiences described earlier, an original web based negotiation game "SurfingGlobalChange" (
Fig. **4**) was invented and implemented (Ahamer, 2004a).

This role-play is inspired by the conviction that equilibrium between two major complementary groups of skills has to be reached for successful professional life, namely *competition and consensus*.

Until 2010, SGC was implemented 25 times for Graz University (USW, 2010) and FH Joanneum in interdisciplinary courses for advanced semesters: Resulting social dynamics was monitored by a number of independent experts invited and financed by the author (e.g. Rauch, 2003). Moreover, a subset of the game idea of level3 was delivered as input to an EU project "UniGame"; additionally, a didactically founded game concept for the Graz contribution to this project was provided (Ahamer, 2003). Furthermore, a game scenario was developed in collaboration with FHJ members (Ahamer et al., 2003), which serves as a basis for a game that has been renamed in the meantime "UniGame: Social skills and knowledge training".

Detailed statistical evaluation of students' results has shown that cognitive performance (e.g. measured by quiz grades), skills of authoring academic articles, skills of reviewing them, and skills of discussion are to a large extent uncorrelated with each other and could be seen as independently varying. For the time being the conclusion is made that such skills have to be measured and assessed separately from each other in order to draw a complete picture of a personality.

## 3. Comparison of characteristics in three generations

### 3.1. Is there a trend in web platforms' functionalities used?

The *three main functionalities* of the web platforms, namely content, quizzes and communication are employed across the three generations, while the clear main trend is a *shift* away from the usage of content-oriented towards the usage of communication-oriented functionalities in the web platforms. The sharply increasing hit frequency underlines such a view and suggests that for students a discussion forum is a tool to create public space for members.

Digital media may serve as a *vehicle* for self-guided learning in thematically and communicatively open structures. Didactic deliberations and fundaments are largely available in Gierlinger et al.(2004) and Ahamer (2004). Web platforms are able to create *public space* as an easily accessible "home" for newly forming groups and as mentally comfortable living room for learners.

The overall trend regarding assessments consists in a *shift of roles*: initially only the lecturer has the power to grade, later on well-defined sub-portions of grading tasks are performed by peer students. Such development is well in line with a finding for another professional field, namely that for the assessment of university studies both internal and external evaluation is necessary (Reissert & Carstensen 1998).

### 3.2. How did assessment and grading develop?

Fig. **5** comprises the development of course units from the first to the third generation taking the described lectures as an example. It is visible that the invention of the web based negotiation game "SurfingGlobalChange" by the author equals further development of two earlier interdisciplinary web based lectures.

### 3.3. Which didactic method is chosen?

Based on the result of three generations of web based learning, SurfingGlobalChange is grounded in didactic deliberations made earlier (Ahamer, 2004) and

- builds on a tradition of simulation and gaming (Klabbers, 2001)
- relies on ethics of negotiation (e.g. Fischer-Kowalski et al., 1995)
- is inspired by constructing realities (Foerster, 2003; Kerres, 2001a)
- does not attempt to mathematically simulate complex realities (Meadows, 2001; Burns, 2002)
- but is simulative for real-life processes (Myers, 1999)
- is founded on systems thinking (Richmond, 1993; Ossimitz, 2000)
- allows for pragmatic strategies (Reilly, 2003)
- and uses environmental topics as trigger for the emerging global responsibility of humanity (Rauch, 2000, 2002, 2002a).

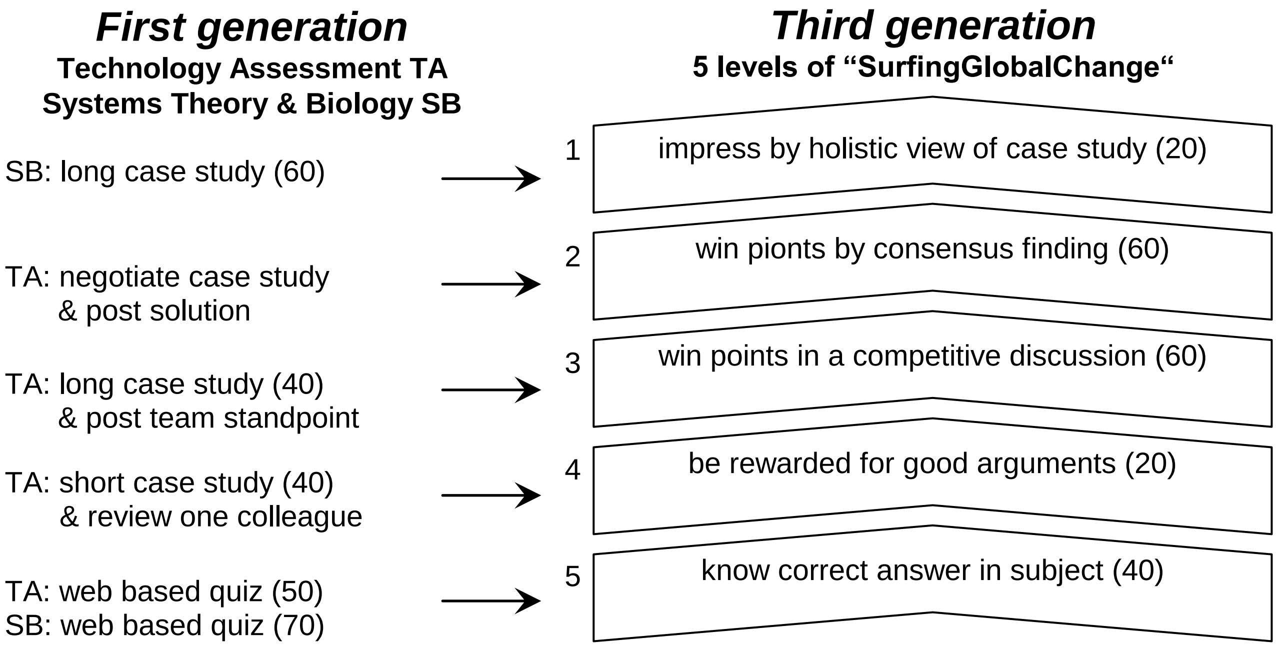

First generation
Technology Assessment TA
Systems Theory & Biology SB
SB: long case study (60)
TA: negotiate case study & post solution
TA: long case study (40) & post team standpoint
TA: short case study (40) & review one colleague
TA: web based quiz (50)
SB: web based quiz (70)
Third generation
5 levels of “SurfingGlobalChange“
1 impress by holistic view of case study (20)
2 win pionts by consensus finding (60)
3 win points in a competitive discussion (60)
4 be rewarded for good arguments (20)
5 know correct answer in subject (40)

**Fig. 5:** Development of course components comprising 4 weekly hours from classical web teaching in the first generation (left) to SGC as the third generation (right). Maximum rewards in the single levels are added in parentheses.

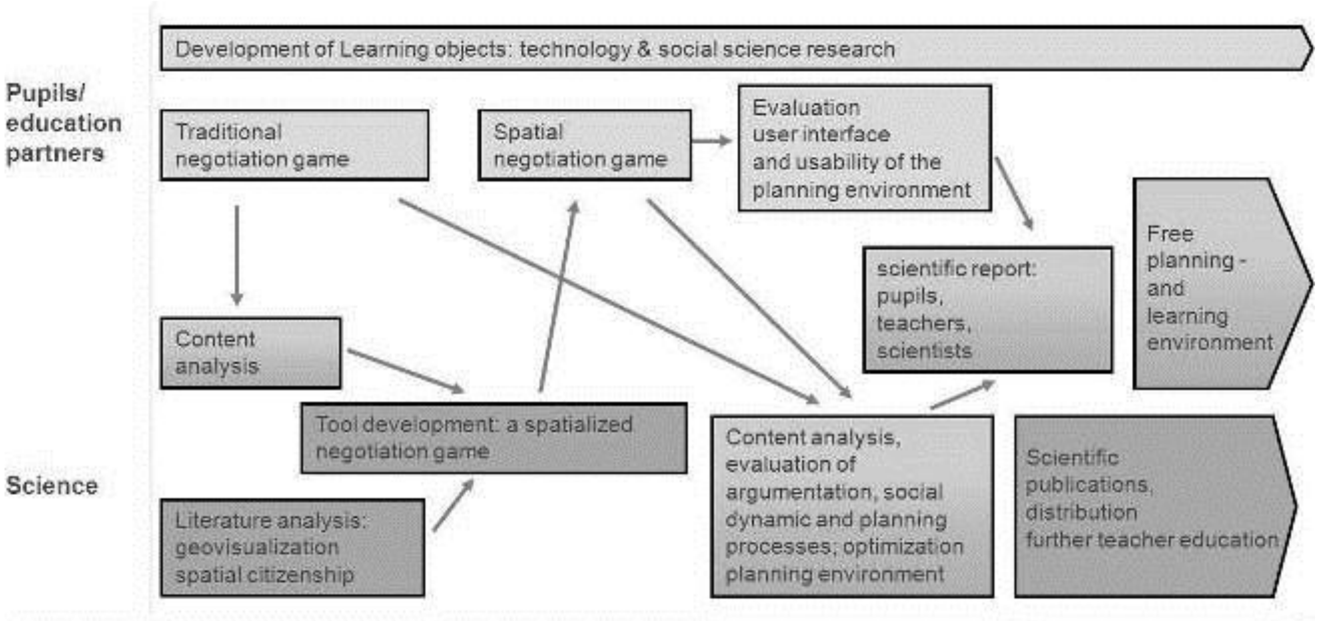

Development of Learning objects: technology & social science research
Pupils/ education partners
Traditional negotiation game
Spatial negotiation game
Evaluation user interface and usability of the planning environment
scientific report: pupils, teachers, scientists
Free planning - and learning environment
Content analysis
Tool development: a spatialized negotiation game
Content analysis, evaluation of argumentation, social dynamic and planning processes; optimization planning environment
Scientific publications, distribution further teacher education
Science
Literature analysis: geovisualization spatial citizenship

**Fig. 6:** The project structure of GEOKOM-PEP (Jekel et al., 2009, Vogler et al., 2010) supports decision making through visualization and mapping.

## 4. Game based learning as the fourth generation

### 4.1. Gaming – through a theoretical lens

The 3rd generation approach of collaboration and mutual assessment opens into focusing on "perspectives" as the main constituents of reality. Multiperspectivism (GS, 2010; IE, 2010) is an approach that specially suits international projects, intercultural approximation and peace efforts.

On the didactic level, this paper suggests to depart from "fact" to "view". Students are invited to take roles and implement them along a series debates backed by previously written and academically supported standpoint papers (levels 3 onwards in SGC). Perry and Sanderson (1998) inspires to such a "theatre of arguments" where bundles of arguments can be handed over from player to player. On the stage of the social spaces defined by the rules of SGC, students can tentatively take roles, adopt, share, transmit and perceive views and fight for them or mould them into a greater consensus. The substrate of action is "views", not "facts".

Views can be handed over to colleagues and slipped into, similar to clothes on a stage. Students explore the argumentative potential of diverse conglomerates of ethic convictions mixed into scientifically backed argumentative approaches.

Such a panel for gaming permits border conditions that are loose enough to allow for readaptation of own convictions along the learning process and tight enough to structurize an ordered debate. Human explanatory constructs are traded among participants and their explanatory value is counterchecked. Each participant feels stimulated to adapt previously adopted aggregates of world views in order to optimize their potential in finding allies.

Coming from the science of "design", Bucciarelli (1998) deals with such "designing of social processes"; Heaton (2002) introduces the notion of „cultural frame", expanding on the idea of "technological frame" and "frame of meaning". MacGregor's (2002) core method to increase appropriate levels of "awareness" throughout the design process is to encourage for "switching" of roles (as does SGC). SGC's rhythms of social interaction implement permanent refaming in the "space of meaning.

Restrepo and Christiaans (2004) in a very good article stress the importance of Underdeterminism and a sufficient amount of degrees of freedom for learning (= adopting new world views) – which is in fact "gaming".

### 4.2. Gaming – the practice

According to the rules of the web-based five-level negotiation game "Surfing Global Change" four to five tables in the lecture room symbolize the views which interact vividly during hour-long structured discussions (Fig. 7). Literally, the lecture room becomes a material manifestation of world views.

Students outside the ring of tables have the task to monitor the performance of their colleagues during discussion and to provide written feedback, thus introducing an element of "reflection in action" and "peer review" into the social process. This allows actors to deepen their understanding by adding an outside view to their actions.

## 5. Geographic Information Systems enable the fifth generation

### 5.1. Public Participation Geographic Information Systems (PPGIS)

Collaborative learning as developed until the 4th generation calls for suitable technological tools to manage the underlying complex fact-based and opinion-driven procedures of communication. During the last years, interactive tools for geo-information (GI) based learning environments has created vast new possibilities.

Both planning and learning processes can be seen as compatible – see Jekel (all years), Mayer et al. (2004, 2004a) – because they are socially embedded in a constructivist model (Foerster 2003, Vygotsky 1978, 1986) and open to Vygotsky's "psychology of play".

First evaluations of GI-based globes yield very positive results regarding the inclusion of learners and the quality of results (Strobl, 2007). Hereby they fulfil the requirements that are requested from constructivist-oriented multimedia learning environments (cf. e.g. Baumgartner 1995).

What do Geographic Information Systems (GIS) provide? Essentially a "viewing tool", a "macroscope": Geographic Information Systems (GIS) under the form of Google Earth or Bing Maps have entered virtually every living room when it comes to finding hotels in an unknown city or discussing about the optimal route to the next holiday site. After holiday, GIS may provide another (idealized) view of lived reality as in the four examples of

Fig. **8**: a bird's view during summer containing the route undertaken by car, a collection of downhill ski routes undertaken during on the first day together with a panorama photo, a documentation of the first beginners' attempts to learn snowboarding on a baby lift and then outreaching to more difficult slopes (together with another author's photo of these slopes), and finally a summer view of the entire ski resort "Marilleva 2000" through the lens of a satellite and of a fotographer posted near a church from opposite.

In this understanding, "geography" is understood as the science providing views. Geographic Information Systems (GIS) does the same, only quicker.

**Fig. 7:** Setting while gaming during discussions of “Surfing Global Change” 2007.

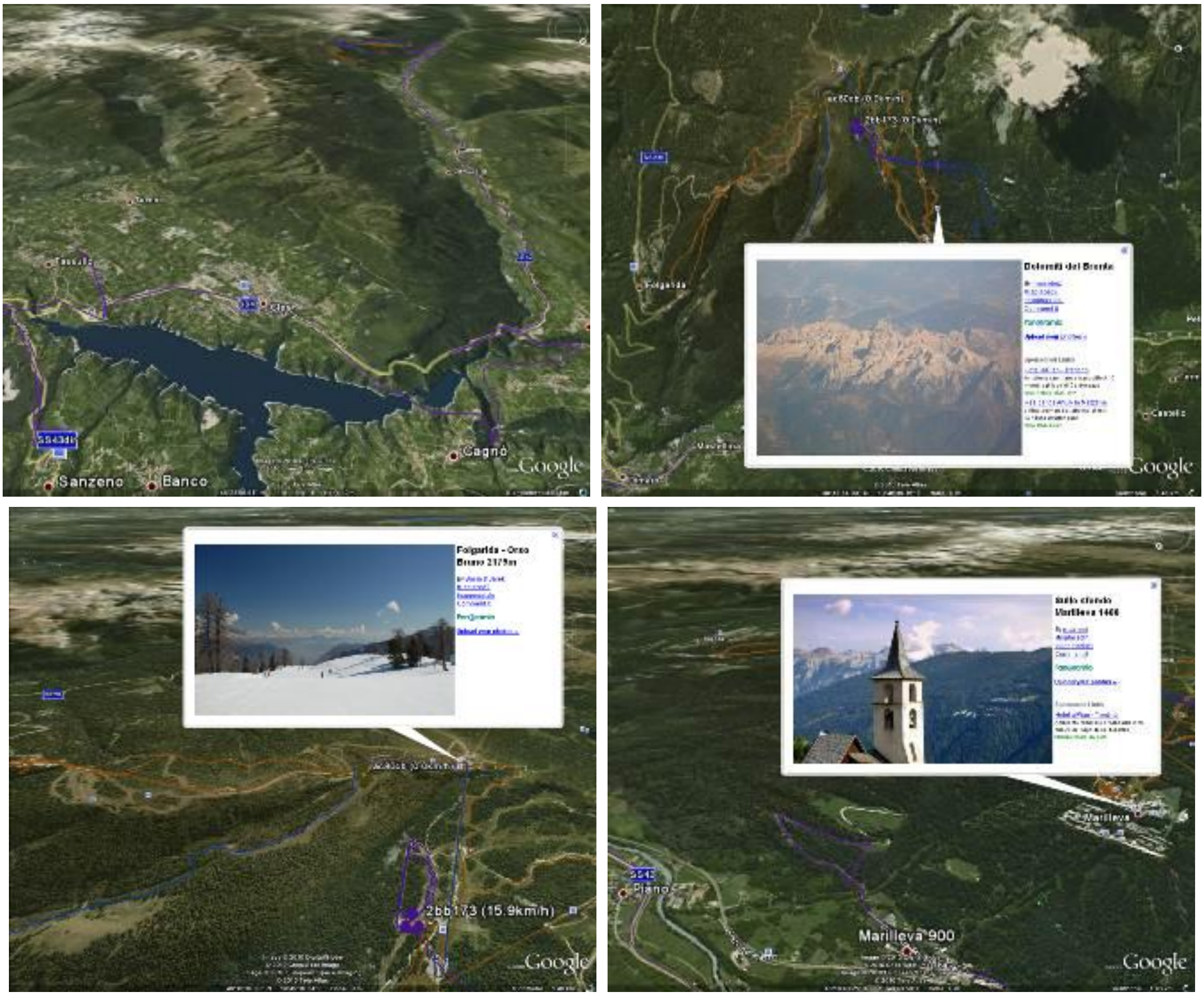


**Fig. 8:** Views on the same “real facts” strongly depend on the viewers, even in such simple cases as an Italian winter holiday site in the western dolomites in 2010.

### 5.2. PPGIS – towards a new participatory practice

Since 2009, a new project (building among others on SGC) explores the enhanced effect of GIS tools on the quality and speed of consensus building (GEOKOM-PEP, 2009, Jekel et al., 2009, Vogler et al., 2010), see Fig. 6.

It is expected that earlier research work will be corroborated indicating that interindividual negotiation processes and consensus finding is significantly enhance when stakeholders use (virtual) maps to visualize their proposals, views and recommendations for solutions.

## 6. Conclusions

This article told the story of the steady development of university courses while gradually increasing the complexity of communication and assessment structures. The guiding philosophy is web based collaborative learning in cases and constructionism.

Seen from the perspective of trainers and learners, the bundle of formerly cognition-oriented targets is enriched: (i) find learning targets yourself, (ii) form teams, (iii) give and get feedback, (iv) reflect and stepwise improve own and others' work.

Concluding from the courses described in this paper, participating students can be observed to pass through consecutive steps as a function of novelty and appeal:

1. learn facts
2. play with facts according to game rules
3. play with rules in an autopoietic way.

Geographic information systems such as virtual globes promise to provides prime options and facilities to ease sustainable decision making.

May the interesting experiences made by game based learning enhanced by virtual globes contribute to developing a sustainable humane future!

## 7. Acknowledgements

Ideas and concepts presented in this overview were developed by initiative of the author using mainly his own resources, but were later aided by existing structures for which he expresses his gratitude: thanks go to FHJ for initial support and encouragement, to Graz University for transposing into more professional structures, to UniGame project members for reflecting ideas and concepts, to eTutors for helping to implement game structures in WebCT, to colleagues at ÖAW for thorough discussion and initiating project proposals, to all students of past courses for their readiness to adapt to unusual circumstances and for participating in role-playing and certainly to my children for insights that life is gaming.

## 8. References

Ahamer, G. (1999). "*Technologiefolgenabschätzung*", „*Systemtheorie und Biologie*“, „*Umwelttechnik*“. Three integrated web based lectures at the University of Applied Technology Fachhochschule Joanneum Graz, „Civil Engineering and Construction Management“ and „Industrial Electronics“, available in WebCT via http://wizard.fh-joanneum.at:8900.

Ahamer, G. (2002). "*Global Change*" – Konzept für den kooperativen Aufbau einer internetgestützten Lehrveranstaltung für Lehrende an der Universität Graz / Studium Umweltsystemwissenschaften. In collaboration with D. Carstensen, Stabsstelle Lehrentwicklung und Evaluation der Universität Graz, as of 15. Februar 2002. Available at http://plato.uni-graz.at:8000/Global_Change.

Ahamer, G. & Carstensen, D. (2002). *Gemeinsame Entwicklung einer internetgestützten Lehrveranstaltung „Global Change“* – während des Sommerkurses entstehendes Kurzkonzept. First draft as of 25.7.02, Graz University.

Ahamer, G.; Ebner, M.; Hasler, A.; Schmickl, T.; Steininger, K. (2002). *Global Change in unserer vernetzten Umwelt* – Handlungskompetenz zur Auffindung von nachhaltigen Konsenslösungen unter Erstellung von Vorhersagen. Konzept für ein interdisziplinäres Praktikum für das Studium „Umweltsystemwissenschaften“ an der Karl-Franzens-Universität Graz, WS03/04, Implementation see at http://www.uni-graz.at/usw/lehre/ahamer.htm and in the web plattform WebCT http://plato.uni-graz.at:8000/SCRIPT/001605.

Ahamer, G. (2003). "*Idea and conceptual design for a web based game*", 9.1.2003; as well as "*Game concept for the Graz contribution to UniGame*", Working paper delivered for an EU project under the Minerva programme (UniGame = Game-Based Learning in Universities and Lifelong Learning, Contract No. 101288-CP-1-2002-1-AT-MINERVA-M) to FH Joanneum Graz, 17.2.2003.

Ahamer, G.; Dziabenko, O.; Schinnerl, I. (2003). *The 'Global Change' Contribution to UniGame - Game Scenario*, 31 pages, 13. 2. 2003. UniGame project under the EU Minerva Programme , Graz.

Ahamer, G. (2004). Negotiate your future: Web based role-play. *Campus-Wide Information Systems*, *21*(1), 35-58.

Ahamer, G. (2004a). *Rules of the new web-supported negotiation game "SurfingGlobalChange"*. 9. Europäischer Kongress der Gesellschaft für Medien in der Wissenschaft, 15.-17. September 2004 (GMW04), Universität Graz.

Barrows, H. (2002). Is it truly possible to have such a thing as dPBL (distributed problem-based learning)? *Distance Education*, *23*(1), p. 119-122.

Barz, A., Carstensen, D., Reissert, R. (1997). *Lehr- und Evaluationsberichte als Instrumente zur*

*Qualitätsförderung. Bestandsaufnahme der aktuellen Praxis*. Gütersloh: CHE-Arbeitspapier Nr. 13.

Baumgartner, P. (2002). *eLearning & eTeaching: Didaktische Modelle*. Vortrag an der FH Joanneum Graz, Institut für Organisation und Lernen (IOL), Abt. Wirtschaftspädagogik und Evaluationsforschung, Universität Innsbruck.

BMBWK (2000). *Die Initiative Neue Medien in der Lehre an Universitäten und Fachhochschulen in Österreich (NML)*. Available at http://serverprojekt.fh-joanneum.at/sp/index.php?n=ini.

Bork, A. (2001). Adult education, lifelong learning, and the future. *Campus-Wide Information Systems, 18*(5), p. 195-203.

Bruns, B., Gajewski, P. (2002). *Multimediales Lernen im Netz – Leitfaden für Entscheider und Planer*. Berlin, München: Springer.

Bucciarelli, L.L. (1998). An ethnographic perspective on engineering design. *Design Studies*, 9(3), 159-168.

Burns, A., (2002). *Civilization III: Digital Game-Based Learning and Macrohistory Simulations*. Australian Foresight Institute / Disinformation, July 2002, see http://www.disinfo.com/pages/article/id2273/pg1.

Carstensen, D., Reissert, R. (1997). *Qualitätsförderung in Hochschulen – Das Verfahren der internen und externen Evaluation*. In: Studium und Lehre an den Hochschulen in Baden-Württemberg, Dokumente vom GEW-Hochschultag '96 an der Universität Heidelberg, Broschüre, GEW BaWü (Hrsg), Stuttgart 5/97.

Carstensen, D. (2002). *Reflexionen und Protokolle über den Sommerkurs „Global Change" an der Universität Graz*. Stabsstelle Lehrentwicklung und Evaluation der Universität Graz, 8.8.03, 22.8.03, 5.9.02.

CML (2002). *Train the Trainer – Ausbildung zur professionellen Gestaltung von Lehrveranstaltungen mit Telelern-Elementen*. Available at http://train-the-trainer.fh-joanneum.at/.

EIA (1997). Richtlinie 97/11/EG des Rates vom 3. März 1997 zur Änderung der Richtlinie 85/337/EWG über die Umweltverträglichkeitsprüfung bei bestimmten öffentlichen und privaten Projekten. Amtsbl. Nr. L 073 vom 14.03.1997.

EIA (2000). Österreichisches Bundesgesetz über die Prüfung der Umweltverträglichkeit (Umweltverträglichkeitsprüfungsgesetz 2000, UVP-G 2000). BGBl. 697/1993 idF BGBl. I 89/2000.

Fischer-Kowalski, M., Pelikan, J., Schandl, H. (1995). *Große Freiheit für kleine Monster*. Wien: Verlag für Gesellschaftskritik.

Florian, M. (2004). *Global Change – sozioökologische Kompetenzen am Beispiel der neuen EU-Mitgliedsstaaten*. Rohkonzept für eine Neuauflage von SurfingGlobalChange in einem 6-stündigen interdisziplinären Praktikum für das Studium Umweltsystemwissenschaften, Uni Graz, Inst. für Geographie.

Foerster, H. v. (2003). *Wahrheit ist eine Erfindung eines Lügners – Gespräche für Skeptiker*. Carl-Auer-Systeme Verlag.

GEOKOM-PEP (2009). Geovisualisierung und Kommunikation in partizipativen Entscheidungsprozessen, http://projects.giscience.at/geokom-pep.

Gierlinger-Czerny, E, Peuerböck, U., Gudera, U. & Berdnik, E. (2002). *Workshop "Selbstgesteuertes Lernen" und Webteaching*, abgehalten am 4. Dezember 2002 in der Vortragsreihe UniImpulse an der Universität Graz.

GS (2010). Global Studies. Master Curriculum at Graz University, see http://www.uni-graz.at/globalstudies/.

Heaton, L. (2002). Designing Work. Situating Design Objects in Cultural Context. *Journal of Design Research*, 2(2).

Horx, M. (2002). *Die acht Sphären der Zukunft*. Vienna: Signum.

IE (2010). Internationale Entwicklung (International Development). Curricula at Vienna University, see http://www.univie.ac.at/ie/.

Jekel, T. (2007), "What you all want is GIS2.0". Collaborative GI based learning environments: spatial planning and education. In: Car, A., Griesebner G. & Strobl, J., GI-Crossroads@GI-Forum. Heidelberg: Wichmann, pp. 84-89.

Jekel, T. & Jekel, A. (2007), Lernen mit GIS 2.0. Kreative Lernwege durch die Integration von digitalen Globen und Lernplattformen. In: Universitäre Lehre neu Denken. Münster: Waxmann (= Medien in der Wissenschaften).

Jekel, T. & Kloyber, L. (2007), Die Einbindung sozialen Raums in GIS als Grundlage partizipativer Planung. In: SIR Berichte & Mitteilungen 33, 123 - 132.

Jekel, T., Pree, J. & Kraxberger, V. (2007), Kollaborative Lernumgebungen mit digitalen Globen - eine explorative Evaluation. In: Jekel/Koller/Strobl, Lernen mit Geoinformation II. Heidelberg: Wichmann, pp. 16 - 126.

Jekel, T., Koller, A. & Strobl, J. (Eds.) (2007), Lernen mit Geoinformation II. Heidelberg: Wichmann.

Jekel, T. (2008), Children Mapping Global Change. Participatory GI-Based Learning. Proceedings, MapIndia Conference, Greater Noida.

Jekel, T. et al. (2009). GEOKOM-PEP, Project proposal.

Kerres, M. (2001). *Multimediale und Telemediale Lernumgebungen – Konzeption und Entwicklung*. 2. Auflage, Oldenburg Verlag.

Kerres, M. (2001a). Medien und Hochschule. Strategien zur Erneuerung der Hochschullehre. In: Ludwig J. Issing, Gerhard Stärk (Hrsg.), *Studieren mit Multimedia und Internet – Ende der traditionellen Hochschule oder Innovationsschub?* (Reihe Medien in der Wissenschaft, Bd. 16) Waxmann: Münster.

Klabbers, J.H.G. (2001). The emerging field of simulation and gaming: Meanings of a retrospect. *Simulation & Gaming*, *32*(4), 471-480.

Lo, S., Koubek, A. & Jandl, M. (1999). *Telelernen an der FH Joanneum: Konzepte & Erfahrungen*. Working paper at the ICL Conference, Villach - Austria.

MacGregor, S. (2002). New Perspectives for Distributed Design Support. *Journal of Design Research*, 2(1).

Mayer, I.S., van Daalen, C.E., Bots, P.W.G. (2004). Perspectives on policy analyses: a framework for understanding and design. *Int. J. Technology, Policy and Management*, 4(2), 169-191.

Mayer, I.S., Bockstael-Blok, W., Valentin, E.C. (2004a). A Building Block Approach to Simulation: An Evaluation Using Containers Adrift. *Simulation & Gaming – An International Journal*, 35(3), 29-52.

Meadows, D.L. (2001). Tools for understanding the limits to growth: Comparing a simulation and a game. *Simulation & Gaming, 32*(4), 522-536.

Myers, D. (1999). Simulation, gaming, and the simulative. *Simulation & Gaming, 30*(4), 482-489.

NML (2002). *Policy Statement für den Einsatz Neuer Medien für die Lehre und das Lernen*. Projektgruppe Neue Medien in der Lehre an der Universität Graz, as of 3. 10. 2002, available via http://neuemedien.uni-graz.at.

Ossimitz, G. (2000). *Entwicklung systemischen Denkens - Theoretische Konzepte und empirische Untersuchungen*. Klagenfurter Beiträge zur Didaktik der Mathematik, Profil Verlag.

Perry, M. and Sanderson, D. (1998). Coordinating joint design work: the role of communication and artefacts, *Design Studies*, Vol. 19, No. 3, pp.273–288.

Prensky, M. (2001), *Digital Game-Based Learning*. New York: McGraw-Hill.

Rauch, F. (2000). *Konzepte in der Umweltbildung*. Kapitel 1 der Habilitationsschrift, IFF Klagenfurt.

Rauch, F. (2002). Education for Sustainability: a Regulative Idea and Trigger for Innovation. In: Scott, W., Gough, S. (Eds.): *Key Issues in Lifelong Learning and Sustainability: A Critical Review*. Routhledge Falmer: London.

Rauch, F. (2002a). *Gesellschaftliche Herausforderungen an das Bildungswesen und die Rolle der Umweltbildung*. IFF Klagenfurt; approximately the German version of: Rauch F., The Potential of Education for Sustainable Development for Reform in Schools. In: Environmental Education Research, 8(1).

Rauch, H. (2003). *Report about the social dynamics of the digital learning game "SurfingGlobalChange" (SGC).* Expert opinion by Institut für Socialanalyse.

Reilly, D.A. (2003). The Power Politics Game: Offensive realism in theory and practice. *Simulation & Gaming, 34*(2), 298-305.

Reissert, R., Carstensen, D. (1998). *Praxis der internen und externen Evaluation. Handbuch zum Verfahren*, HIS-Kurzinformation "Spezial", Hannover, available at http://evanet.his.de/evanet/PDF/Pdf_dok/Handbuch.fuer.evanet1.pdf.

Restrepo, J., Christiaans, H. (2004). Problem Structuring and Information Access in Design. *Journal of Design Research*, 4(2).

Richmond, B. (1993). Systems thinking: Critical thinking skills for the 1990s and beyond. *System Dynamics Review*, *9*(2), 113-133.

Rogers, C.R. (1974), *Lernen in Freiheit*. München: Kösel.

Strobl, J. (2007), Geographic Learning in Social Web Environments. – In: Donert, K. *GIS in Geography in Higher Education*, Teaching Geography in Higher Education, San Diego, ESRI Publications.

WebCT (2009). *Web Course Tools*. Description available at http://www.webct.com.

USW (2010). Environmental Systems Analysis (Umweltsystemwissenschaften), Reports of Interdisciplinary Practicals, see http://www.uni-graz.at/usw1www/usw1www_magazin/usw1www_berichte.htm.

Vygotsky, L. (1978), *Mind in Society: The development of higher psychological processes*. Harvard University Press. Cambridge.

Vygotsky,L.S. (1986), *Thought and Language*. The MIT Press, Cambridge, MA.

Vogler, R., Ahamer, G. & Jekel, T. (2010): GEOKOM-PEP. Pupil led research into the effects of geovisualization. In: Jekel, T., Koller, A., Donert, K. & Vogler, R. (eds.): Learning with Geoinformation V; Heidelberg: Wichmann, pp. 51-60.

## Author Biography

**Gilbert Ahamer:** Born and working in the historic city of Salzburg, he has always been interested in the history of technology and science, despite being trained as a physicist, environmentalist and economist. At the Austrian Academy of Sciences he tries to link the concept of "spaces" with options provided by Information and Communication Technologies ICT.
Participation and GIS as a case study promise to facilitate procedures of human understanding.

# State-Space Modelling of Dynamic Systems Using Hankel Matrix Representation

H. Olkkonen[1], S. Ahtiainen[1], J.T. Olkkonen[2] and P. Pesola[1]

[1] Department of Physics and Mathematics, University of Eastern Finland, 70211 Kuopio, Finland
[2] VTT Technical Research Centre of Finland, B.O. Box 1000, 02044 VTT, Finland
(email: hannu.olkkonen@uef.fi)

***Abstract*:** In this work a dynamic state-space model was constructed using a Hankel matrix formulation. A novel update algorithm for computation of the state transition matrix and its eigenvalues was developed. The method suits for analysis and synthesis of the rapidly changing dynamic systems and signals corrupted with additive random noise. The knowledge of the time varying state transition matrix and its eigenvalues enables accurate and precise numerical operators such as differentiation and integration in the presence of noise.

***Keywords*:** State-space modelling, dynamic systems analysis

## 1. Introduction

Estimation of the state of the dynamic systems and signals has been an object of vital research for many years impacted by the discovery of the Kalman filter (KF), extended Kalman filter (EKF), neural network algorithms and the LMS and RLS algorithms [1-7]. The adaptive algorithms have found to be suitable methods for many kind of linear and nonlinear system modelling. The update of the model parameters is based on the use of the forgetting functions. State-space models, which are based on matrix formulations have gained acceptance in various control system analysis and synthesis. A state-space approach differs significantly from the adaptive methods such as KF, EKF and RLS algorithms in that the system matrices are solved directly form the measured data matrices using least squares (LS) or total least squares (TLS) methods. The noise inherent in data matrices is usually cancelled by the singular value decomposition (SVD) based subspace methods [8-9]. A disadvantage in the SVD based solutions is the treatment of the data matrix blocks, which give the system matrices in a defined time interval. The matrices are then supposed to be time-invariant within the time interval. However, in rapidly changing dynamic systems the matrices may change abruptly and the SVD based methods give only a time averaged estimates.

In this work we present a dynamic state-space model, where the state transition matrix is updated at every time increment. The dynamic system modelling is based on the Hankel data matrix representation. We present a novel update algorithm for computation of the state-transition matrix and its eigenvalues. The method can be adapted for system state-space modelling and filtering the measurement signals in the presence of noise.

## 2. Theoretical considerations

### 2.1 The dynamic state-space model

The dynamic state-space model under consideration is defined as

$$\begin{aligned} X_{n+1} &= F_n\, X_n \\ y_n &= C\, X_n + w_n \end{aligned} \tag{1}$$

where the state vector $X_n \in R^{Nx1}$, the state transition matrix $F_n \in R^{NxN}$ and the vector $C = [1\ 0 \cdots 0] \quad \in R^{1xN}$ .The scalar $w_n \in R^{1x1}$ is a random zero mean observation noise. The signal $y_n \in R^{1x1}$ consists of measurements at time intervals $t = nT\ (n = 0,1,2,...)$, where $T$ is the sampling period. Let us define the data vector $Y_n = \left[ y_n\ y_{n-1}\ y_{n-2} \cdots y_{n-N-1} \right]^T$ where T denotes matrix transpose. The Hankel structured data matrix $H_n \in R^{NxM}$ is defined as

$$\begin{aligned} H_n &= \begin{bmatrix} y_n & y_{n-1} & \cdots & y_{n-M-1} \\ y_{n-1} & y_{n-2} & \cdots & y_{n-M-2} \\ \vdots & \vdots & \vdots & \vdots \\ y_{n-N-1} & y_{n-N-2} & \cdots & y_{n-N-M-2} \end{bmatrix} \\ &= \left[ Y_n \quad Y_{n-1} \quad \cdots \quad Y_{n-M-1} \right] \end{aligned} \tag{2}$$

where the subscript n in $H_n$ refers to the most recent data point $y_n$. The antidiagonal elements of the $H_n$ data matrix are equal. The state-space model (1) can be represented in the following form

$$H_{n+1} = F_n\, H_n + W_{n+1} \tag{3}$$

where $W_{n+1}$ is the Hankel structured noise matrix. The least squares estimate of the state transition matrix $F_n$ comes from

$$F_n = H_{n+1}\, H_n^T (H_n\, H_n^T)^{-1} = H_{n+1}\, H_n^{\#} = R_n C_n^{-1} \tag{4}$$

where the pseudoinverse matrix $H_n^{\#} = H_n^T (H_n\, H_n^T)^{-1} \in R^{MxN}$ . The matrices $R_n = H_{n+1} H_n^T \in R^{NxN}$ and $C_n = H_n H_n^T \in R^{NxN}$ . The $C_n$ matrix is symmetrical, i.e. $C_n^T = C_n$ . It has a stable inverse since it is positive definite and all eigenvalues are nonnegative. The rank of the state transition matrix $F_n$ defines the system order. In many applications the state transition matrix should be evaluated at T intervals. In complex dynamic systems the dimension of the state transition matrix is high and the computation of the pseudoinverse matrix $H_n^{\#}$ would be an overwhelming task. In this work we show that with a special partitioning the $R_n$ and $C_n$ matrices into submatrices the computational load is drastically diminished.

### 2.2 Computation of the state transition matrix $F_n$

The data matrix $H_n$ is partitioned as

$$H_n = \begin{bmatrix} D_n \\ d_n \end{bmatrix} \tag{5}$$

where the matrix $D_n \in R^{(N-1)xM}$ and the vector $d_n \in R^{1xM}$. The data matrix $H_{n+1}$ is partitioned as

$$H_{n+1} = \begin{bmatrix} d_{n+1} \\ D_n \end{bmatrix} \tag{6}$$

where the vector $d_{n+1} \in R^{1xM}$ and the matrix $D_n$ is identical to that in (5). Now we have

$$C_n = \begin{bmatrix} D_n \\ d_n \end{bmatrix} \begin{bmatrix} D_n^T & d_n^T \end{bmatrix} = \begin{bmatrix} D_n D_n^T & D_n d_n^T \\ d_n D_n^T & d_n d_n^T \end{bmatrix} = \begin{bmatrix} A_n & b_n \\ b_n^T & c_n \end{bmatrix} \tag{7}$$

where the matrix $A_n \in R^{(N-1)x(N-1)}$, the vector $b_n \in R^{(N-1)x1}$ and the scalar $c_n \in R^{1x1}$. The analytic solution of the inverse matrix is

$$C_n^{-1} = \begin{bmatrix} A_n & b_n \\ b_n^T & c_n \end{bmatrix}^{-1} = \begin{bmatrix} A_n^{-1} + s_n m_n m_n^T & -s_n m_n \\ -s_n m_n^T & s_n \end{bmatrix} \tag{8}$$

where the vector $m_n = A_n^{-1} b_n \in R^{(N-1)x1}$ and the scalar $s_n = (c_n - b_n^T m_n)^{-1} \in R^{1x1}$. The inverse matrix has the same block dimensions as the $C_n$ matrix. Correspondingly, we have

$$R_n = H_{n+1} H_n^T = \begin{bmatrix} d_{n+1} \\ D_n \end{bmatrix} \begin{bmatrix} D_n^T & d_n^T \end{bmatrix} = \begin{bmatrix} d_{n+1} D_n^T & d_{n+1} d_n^T \\ D_n D_n^T & D_n d_n^T \end{bmatrix} = \begin{bmatrix} e_n & g_n \\ A_n & b_n \end{bmatrix} \tag{9}$$

where the vector $e_n \in R^{1x(N-1)}$ and the scalar $g_n \in R^{1x1}$. The matrix $A_n$ and the vector $b_n$ are identical to in (7). Now we obtain the state transition matrix as

$$F_n = R_n C_n^{-1} = \begin{bmatrix} e_n & g_n \\ A_n & b_n \end{bmatrix} \begin{bmatrix} A_n^{-1} + s_n m_n m_n^T & -s_n m_n \\ -s_n m_n^T & s_n \end{bmatrix}$$

which finally yields

$$F_n = \begin{bmatrix} h_n & k_n \\ I & \varnothing \end{bmatrix} \tag{10}$$

where the scalar $k_n = g_n s_n - s_n e_n m_n$ and the vector $h_n = e_n A_n^{-1} - k_n m_n^T$. The identity matrix $I \in R^{(N-1)x(N-1)}$ and the zero vector $\varnothing \in R^{(N-1)x1}$.

### 2.3 Updating the state transition matrix $F_n$

The updated matrix $C_{n+1}$ is partitioned into the four sub-blocks

$$C_{n+1} = \begin{bmatrix} d_{n+1} \\ D_n \end{bmatrix} \begin{bmatrix} d_{n+1}^T & D_n^T \end{bmatrix} = \begin{bmatrix} d_{n+1} d_{n+1}^T & d_{n+1} D_n^T \\ D_n d_{n+1}^T & D_n D_n^T \end{bmatrix} = \begin{bmatrix} p_{n+1} & e_n \\ e_n^T & A_n \end{bmatrix} \tag{11}$$

where $p_{n+1}$ is a scalar. The essential observation is that $C_{n+1}$ contains a submatrix $A_n$, which is the same as in partitioning the $C_n$ matrix (7). The vector $e_n$ equals to that in (9). The analytic matrix inversion gives

$$C_{n+1}^{-1} = \begin{bmatrix} p_{n+1} & e_n \\ e_n^T & A_n \end{bmatrix}^{-1} = \begin{bmatrix} r_{n+1} & -r_{n+1} q_n \\ -r_{n+1} q_n^T & A_n^{-1} + r_{n+1} q_n^T q_n \end{bmatrix} \tag{12}$$

where the vector $q_n = e_n A_n^{-1}$ and the scalar $r_{n+1} = (p_{n+1} - q_n e_n^T)^{-1}$. We may note that the vector $q_n$ is previously computed as a first term in vector $h_n$ in (10). Finally, the computed $C_{n+1}^{-1}$ matrix is represented in the repartitioned form (7)

$$C_{n+1}^{-1} = \begin{bmatrix} A_{n+1} & b_{n+1} \\ b_{n+1}^T & c_{n+1} \end{bmatrix}^{-1} = \begin{bmatrix} A_{n+1}^{-1} + s_{n+1} m_{n+1} m_{n+1}^T & -s_{n+1} m_{n+1} \\ -s_{n+1} m_{n+1}^T & s_{n+1} \end{bmatrix} = \begin{bmatrix} T_{n+1} & z_{n+1} \\ z_{n+1}^T & w_{n+1} \end{bmatrix} \tag{13}$$

where the vector $m_{n+1} = A_{n+1}^{-1} b_{n+1}$ and the scalar $s_{n+1} = (c_{n+1} - b_{n+1}^T m_n)^{-1}$. The matrix $T_{n+1} \in R^{(N-1)x(N-1)}$, the vector $z_{n+1} \in R^{(N-1)x1}$ and the scalar $w_{n+1} \in R^{1x1}$ are picked up from the computed inverse matrix $C_{n+1}^{-1}$ (12). By comparing the block matrices we obtain the solution for the vector $m_{n+1}$ and the inverse block matrix $A_{n+1}^{-1}$ as

$$\begin{aligned} m_{n+1} &= -w_{n+1}^{-1} z_{n+1} \\ A_{n+1}^{-1} &= T_{n+1} + m_{n+1} z_{n+1}^T \end{aligned} \tag{14}$$

Now we get the updated $R_{n+1}$ matrix

$$R_{n+1} = \begin{bmatrix} e_{n+1} & g_{n+1} \\ A_{n+1} & b_{n+1} \end{bmatrix} \tag{15}$$

where the vector $e_{n+1} \in R^{1x(N-1)}$ and the scalar $g_{n+1} \in R^{1x1}$. Finally, the uptake of the state transition matrix $F_{n+1}$ comes from

$$F_{n+1} = R_{n+1}C_{n+1}^{-1} = \begin{bmatrix} e_{n+1} & g_{n+1} \\ A_{n+1} & b_{n+1} \end{bmatrix} \begin{bmatrix} A_{n+1}^{-1} + s_{n+1}m_{n+1}m_{n+1}^{T} & -s_{n+1}m_{n+1} \\ -s_{n+1}m_{n+1}^{T} & s_{n+1} \end{bmatrix} = \begin{bmatrix} h_{n+1} & k_{n+1} \\ I & \varnothing \end{bmatrix} \qquad (16)$$

where the scalar $k_{n+1} = g_{n+1}s_{n+1} - s_{n+1}e_{n+1}m_{n+1}$ and the vector $h_{n+1} = e_{n+1}A_{n+1}^{-1} - k_{n+1}m_{n+1}^{T}$. An interesting feature is that the vector $b_{n+1}$ needs not to be known in the uptake process (16).

### 2.4 Fast uptake of the $R_n$ matrix

The uptake process (16) needs the computation of the vector $e_{n+1} = d_{n+2}D_{n+1}^{T}$ and the scalar $g_{n+1} = d_{n+2}d_{n+1}^{T}$, if partitioning (5) is used. This would require one vector-matrix and one vector-vector multiplication. Fast uptake of the $R_n$ matrix is obtained if we adapt the partitioning (2) for the data matrix

$$H_n = \begin{bmatrix} Y_n & Y_{n-1} & \cdots & Y_{n-M-1} \end{bmatrix} \qquad (17)$$

Now we have

$$R_n = [Y_{n+1}Y_n^T \quad Y_nY_{n-1}^T \quad \cdots \quad Y_{n-M}Y_{n-M-1}^T] \qquad (18)$$

and the uptake of the $R_n$ matrix is yielded as

$$R_{n+1} = R_n + Y_{n+2}Y_{n+1}^T - Y_{n-M}Y_{n-M-1}^T \qquad (19)$$

The uptake of the $R_n$ matrix needs only two vector-vector multiplications and the vector $e_{n+1}$ and the scalar $g_{n+1}$ are simply picked up from the $R_{n+1}$ matrix (19).

### 2.5 Computational complexity

The uptake of the state transition matrix requires the computation of the matrix inverse $C_{n+1}^{-1}$ (12), which needs one vector-vector multiplication. The uptake of the vector $m_{n+1}$ and the inverse block matrix $A_{n+1}^{-1}$ (14) needs one vector-vector multiplication. Finally, the uptake of the scalar $k_{n+1}$ and the vector $h_{n+1}$ in (16) requires one vector-matrix and three vector-vector multiplications. Thus the computational complexity of the algorithm is $O(n^2) + 5O(n)$.

## 3. Applications of the method

### 3.1 Computation of the eigenvalues of the state transition matrix

The Hankel data matrix representation (3,5) of the dynamic state-space model leads to a companion matrix structure of the state transition matrix $F_n$ (10) , which involves only the vector $h_n$ and the scalar $k_n$. An important advantage of the companion matrix structure is that the eigenvalues $\lambda_1, \lambda_2, ..., \lambda_N$ of the state transition matrix can be directly computed as the roots of the polynomial having coefficients [1 $-h_n$ $-k_n$ ]. The eigenvalues of the state transition matrix give important knowledge of the order and the stability of the system and its dynamic behaviour. The eigenvalues also aid in selection of the model order. The occurrence of very small eigenvalues indicates that the system order is smaller that the model order, which leads to overmodelling. When the model order equals the system order, the scalar coefficient $k_n$ attains a value $k_n = -1$.

### 3.2 Signal prediction and state-space filtering

The knowledge of the state transition matrix $F_n$ enables the prediction of the measurement signal $y_n$ as

$$H_{n+1} = F_nH_n \Rightarrow \hat{y}_{n+1} = C\,F_nH_n \qquad (20)$$

where $C = [1\ 0 \cdots 0] \in R^{1xN}$. Using the Hankel data matrix representation we may define the prediction data matrix as

$$\hat{H}_n = \begin{bmatrix} \hat{y}_n & \hat{y}_{n-1} & \cdots & \hat{y}_{n-M-1} \\ \hat{y}_{n-1} & \hat{y}_{n-2} & \cdots & \hat{y}_{n-M-2} \\ \vdots & \vdots & \vdots & \vdots \\ \hat{y}_{n-N-1} & \hat{y}_{n-N-2} & \cdots & \hat{y}_{n-N-M-2} \end{bmatrix} \qquad (21)$$

$$\Rightarrow \hat{H}_{n+1} = F_n\,\hat{H}_n$$

The state-space filtered signal $\hat{y}_n$ can be obtained as a mean of the antidiagonal elements. In the following we describe several matrix operators based on the state transition matrix. In all computations the filtered data matrix (21) is applied.

### 3.3 Numerical signal processing

The knowledge of the state transition matrix $F_n$ enables the numerical signal processing of the state-space filtered signal. In the following we develop matrix operators based on the state transition matrix for numerical interpolation, differentiation and integration of the measurement signal.

The state transition matrix can be presented in the eigenvalue decomposited form $F_n = U_nD_nU_n^{-1}$, where $D_n \in R^{NxN} = diag(\lambda_1\ \lambda_2 \ldots \lambda_N)$ and $U_n \in R^{NxN}$. Based on (21) we have a result

$$\begin{aligned} \hat{H}_{n+1} &= F_n\hat{H}_n = U_nD_nU_n^{-1}\hat{H}_n \Rightarrow \\ \hat{H}_{n+\Delta} &= U_nD_n^{\Delta}U_n^{-1}\hat{H}_{n+\Delta} = F_n^{\Delta}\hat{H}_n \end{aligned} \qquad (22)$$

where the time-shift $\Delta \in [0, T]$. Now we may define the interpolating time-shift operator $S_{n,\Delta} \in R^{NxN}$ as

$$\hat{H}_{n+\Delta} = S_{n,\Delta}\hat{H}_n \Rightarrow S_{n,\Delta} = F_n^{\Delta} \qquad (23)$$

Next, we may define the differentiation operator $D_n \in R^{NxN}$ as

$$\frac{d}{dt}\hat{H}_n = D_n\hat{H}_n \Rightarrow \hat{H}_{n+1} = e^{D_n}\hat{H}_n \qquad (24)$$

Due to (20) we have

$$F_n = e^{D_n} \Rightarrow D_n = \log m(F_n) \qquad (25)$$

where $\log m(\cdot)$ denotes matrix logarithm.

Further, by defining the integral operator $I_n \in R^{NxN}$ as

$$\int \hat{H}_n dt = I_n\hat{H}_n \qquad (26)$$

Since the differentiation and integral operator are inverse operators

$$I_n \hat{H}_n = D_n^{-1} \hat{H}_n = \left[ \log m(F_n) \right]^{-1} \hat{H}_n \Rightarrow \\ I_n = \left[ \log m(F_n) \right]^{-1} \tag{27}$$

and the definite integral is yielded as

$$\int_{(n-d)T}^{nT} \hat{H}_n dt \\ = \left( \left[ \log m(F_n) \right]^{-1} - \left[ \log m(F_{n-d}) \right]^{-1} \right) \hat{H}_n \tag{28}$$

The interpolating, differentiation and integral operators are commutative, i.e. $S_n D_n = D_n S_n$ and $S_n I_n = I_n S_n$. The computation of the second, third etc. derivatives and integrals of the signals are also possible using the matrix operators, e.g. the second derivative operator is obtained as $D_n^2 = \left[ \log m(F_n) \right]^2$. It should be pointed out that applied to the state-space filtered signals the numerical operators are analytic, i.e. they produce results with machine precision.

## 4. Discussion

The distinct difference between the present algorithm and the SVD based methods is that the present algorithm updates the state transition matrix $F_n$ at every time interval, while the SVD based algorithms [8-9] compute the state transition matrix in data blocks. Our algorithm is more feasible in the analysis of the fastly changing dynamic systems and especially for real-time applications, where the eigenvalues of the state transition matrix give actual information on the system functioning.

A key idea in this work is the repartitioning scheme (13), which yields the uptake of the vector $m_{n+1}$ and the inverse block matrix $A_{n+1}^{-1}$ (14) and then the uptake of the state transition matrix $F_{n+1}$ (16). The companion matrix structure of the matrix $F_n$ enables the computation of the eigenvalues of the state transition matrix via the roots of the polynomial [1 $-h_n$ $-k_n$ ]. This procedure is much faster than the direct eigenvalue decomposition of the $F_n$ matrix. The knowledge of the eigenvalues yields a plenty of numerical signal processing tools, such as interpolation, differentiation and integration operators (21,22,26), which compete with the conventional B-spline signal processing algorithms [10-12].

## References

[1] F. Daum, Nonlinear filters: Beyond the Kalman Filter, IEEE A&E Systems Magazine, vol. 20, pp. 57-69, Aug. 2005.

[2] A. Moghaddamjoo and R. Lynn Kirlin, “Robust adaptive Kalman filtering with unknown inputs,” IEEE Trans. Acoustics, Speech and Signal Process. vol. 37, No. 8, pp. 1166-1175, Aug. 1989.

[3] J. L. Maryak, J.C. Spall and B.D. Heydon, “Use of the Kalman filter for interference in state-space models with unknown noise distributions,” IEEE Trans. Autom. Control, vol, 49, No. 1, pp. 87-90, Sep. 2005.

[4] R. Diversi, R. Guidorzi and U. Soverini, “Kalman filtering in extended noise environments,” IEEE Trans. Autom. Control, vol. 50, No. 9, pp. 1396-1402, Sep. 2005

[5] D.-J. Jwo and S.-H. Wang, Adaptive fuzzy strong tracking extended Kalman Filtering for GPS navigation, IEEE Sensors Journal, vol. 7, no. 5, pp. 778-789, May 2007.

[6] S. Attallah, The wavelet transform-domain LMS adaptive filter with partial subband-coefficient updating, IEEE Trans. Circuits and Systems II, vol. 53, no. 1, pp. 8-12, Jan. 2006

[7] H. Olkkonen, P. Pesola, A. Valjakka and L. Tuomisto, “Gain optimized cosine transform domain LMS algorithm for adaptive filtering of EEG,” *Comput. Biol. Med*, vol, 29, pp. 129-136, 1999.

[8] S. Park, T.K. Sarkar and Y. Hua, A singular value decomposition-based method for solving a deterministic adaptive problem, Digital Signal processing 9, 57-63, 1999.

[9] T.J. Willink, “Efficient adaptive SVD algorithm for MIMO applications,” IEEE Trans. Signal Process., vol. 56, no. 2, pp.615-622, Feb. 2008.

[10] M. Unser, A. Aldroubi and M. Eden, “B-spline signal processing. I. Theory,” *IEEE Trans. Signal Process.*, vol. 41, no. 2, pp. 821-833, Feb. 1993.

[11] M. Unser, A. Aldroubi and M. Eden, “B-spline signal processing. II. Efficiency design and applications,” *IEEE Trans. Signal Process.*, vol. 41, No. 2, pp. 834-848, Feb. 1993.

[12] J.T. Olkkonen and H. Olkkonen, “Fractional time-shift B-spline filter,” IEEE Signal Process. Letters, vol. 14, No. 10, pp. 688-691, Oct. 2007.

# Incident Response Process Guidelines - For Information Security Management

**[1]S. P. DATTA, [2]PRANAB BANERJEE**

[1]Prof., Eastern Institute Of Management, Kalyani University, Kolkata, India-700071

[2]Prof., Dept. of Electronics & Telecommunication Engineering, Kolkata, India-700032

E-mail: sp_datta2000@yahoo.co.in, pkbeceju65@gmail.com

## ABSTRACT

Attacks on information systems and networks have become more numerous, sophisticated, and severe in recent years. New types of security-related incidents emerge more frequently. While preventing such attacks would be the ideal course of action for organizations, not all information system security incidents can be prevented. Every organization that depends on information systems and networks to carry out its mission should identify and assess the risks to its systems and its information and reduce those risks to an acceptable level [1], [2]. An important component of this risk management process is the trending analysis of past computer security incidents and identifying effective ways to deal with them. An incident response capability is therefore necessary for rapidly detecting incidents, minimizing loss or destruction, mitigating the weakness that were exploited, and restoring the information system. A well-defined incident response capability helps the organization detect incidents rapidly, minimize loss and destruction, identify weaknesses, and restore information technology (IT) operations rapidly.

Because performing incident response effectively is a complex undertaking, establishing a successful incident response capability requires substantial planning and resources. Continuous monitoring of threats through intrusion detection and prevention systems (IDPS) is essential. Establishing clear procedures for assessing the current and potential business impact of incidents is critical, as it is implanting effective methods of collecting, analyzing, and reporting data. Building relationships and establishing suitable means of communication with other internal groups (human resources, legal, etc.) and with external groups like (FISMA, OMB 79, CERT/CC, etc.) are also vital. National Institute of Standards and Technology (NIST) Special Publication (SP 800-61), *Computer Security Incident Handling Guide*, details a four-phase incident response process.

This article seeks to assist organizations in mitigating the risks from computer security incidents by providing practical guidelines on responding to security incidents effectively and efficiently. It includes guidelines on establishing an effective incident response program. Primarily focus has been put in detecting, analyzing, prioritizing and handling incidents. Organizations' agencies CERTS (Computer Security Response Teams) are encouraged to tailor the suggested guidelines to meet their specific security and mission requirements.

## 1. INTRODUCTION

Overall incident response process life-cycle is segregated into four phases – preparation, detection and analysis, containment/eradication/recovery, and post incident activity[3]. This primarily encompasses the following items:

Organizing a computer security incident response capability;
Handling incidents from initial preparation through the post-incident lessons learned phase;
Handling specific types of incidents

- Denial of Service (DOS)
- Malicious codes

- Unauthorized access
- Inappropriate usage
- Multiple components.

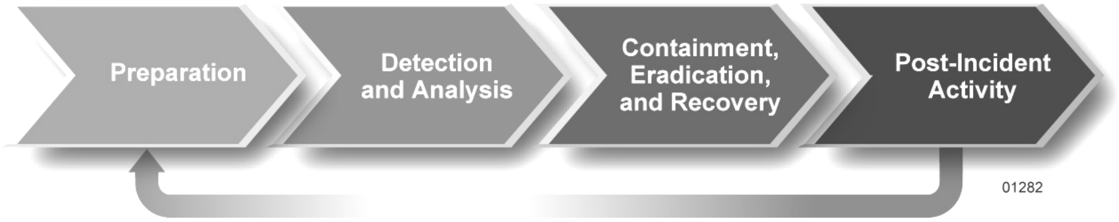


**Figure 1. Incident Response Life Cycle**

Figure 1 illustrates the incident response life cycle. The initial phase of incident response process involves establishing and training an incident response team, and acquiring the necessary tools and resources. During preparation, the organization also attempts to limit the number of incidents that will occur by selecting and implementing a set of controls based on the result of risk assessments. However, residual risk will inevitably persist after controls are implemented. Detection of security breaches is thus necessary to alert the organization whenever the incident occur. In keeping with the severity of the incident, the organization can act to mitigate the impact of the incident by containing it and ultimately recovering from it. After the incident is adequately handled, the organization issues a report that details the cause and cost of the incident and the steps the organization should take to prevent future incidents.

## 1.1. PREPARATION

Incident preparation involves not only establishing an incident response capability so that the organization is ready to respond to incidents but also preventing incidents by ensuring that systems, networks, and applications are afforded sufficient security. Incident prevention is now considered a fundamental component of incident response programs, also known as incident management programs, although the incident response team is not typically responsible for it. The incident response team's expertise should be used to establish recommendations for securing systems and preventing incidents, as much as possible. This phase provides basic advice on preparing to handle incidents and on preventing incidents.

### 1.1.1. PREPARING FOR INCIDENT RESPONSE

Organizing an effective incident response capability involves the participation of many people within the organization. Making the right planning and implementation decisions is key to establishing a successful incident response program. One of the first planning tasks should be to develop an organization-specific definition of the term "incident" so that the scope of the term is clear. Additional tasks that should be performed during the preparation phase include the following:

**Create an Incident Response Policy.** The policy should define what events are considered incidents, establish the organizational structure for incident response, define roles and responsibilities, and list the organization's incident reporting requirements.

**Develop Incident Response and Reporting Procedures.** Based on the incident response policy, standard operating procedures (SOPs) are a delineation of the specific technical processes, techniques, checklists, and forms used by the incident response team. SOPs should be comprehensive and detailed to ensure that the organization's priorities are properly reflected in response operations. In addition, following standardized response procedures is also an effective way to minimize errors. Prior to implementation, the organization should test incident response SOPs in order to validate their accuracy and usefulness. Once validated, the SOPs must be widely disseminated throughout the organization. Incidents occur in unpredictable ways; therefore, it is impractical to develop comprehensive procedures with step-by-step instructions for handling every incident. The best that the organization can do is to remain prepared to handle any type of incident, and more specifically, to handle common types of incidents.

**Establish Guidelines for Communicating with External Parties.** During the incident response process, the organization may need to communicate with outside parties, including other incident response teams, law enforcement, the media, vendors, and external victims. Because such communications often need to occur quickly, organizations should have predetermined communication guidelines so that only the appropriate information is shared with the right parties. However, if sensitive information is inappropriately released, it can lead to greater disruption and financial loss than the incident itself. Creating and maintaining a list of internal and external points of contacts (POC), along with backups for each contact, should assist in making communications among parties easier and faster.

**Define Incident Response Team Services.** Although the main focus of an incident response team is performing incident response, additional services an incident response team can provide to the organization include security advisory distribution, vulnerability assessment, intrusion detection, and education and awareness.

**Adopt a Team Structure and Staffing Model.** The organization should adopt the team structure and staffing model best suited to its needs. When contemplating the best team structure and staffing model, an organization need to consider several factors, such as size of the organization, the geographic diversity of major computing resources, the need for 24/7 availability, cost, and staff expertise.

**Staff and Train the Incident Response Team.** Members of the incident response team should have excellent technical and problem-solving skills because they are critical to the team's success. Excellent teamwork, organizational culture, communication ability, and speaking skills are important as well. Most incident response teams have a team manager and a deputy team manager who assumes authority in the absence of the team manager. In addition, some teams also have a technical lead who assumes oversight of and final responsibility for the quality of the technical work performed by the entire incident response team. Also, larger teams often assign an incident lead as the primary POC for handling a specific incident.

Organizations find difficulties to maintain situational awareness for handling large-scale incidents because of their complexity. Many people within the organization may play a role in the incident response, and the organization may need to communicate promptly and efficiently with various external groups. Collecting, organizing, and analyzing all the pieces of information so that the right decisions can be made and executed are not easy tasks. The key to maintaining situational awareness is to prepare the organization thoroughly to handle large-scale incidents. Two specific actions that support this matter are as follows:

**Establish and Maintain Accurate Notification Mechanisms.** Organizations should establish, document, maintain, and exercise on-hour and off-hour contact and notification mechanisms for various individuals and groups within the organization (e.g., chief information officer [CIO], head of information security, IT support, business continuity planning) and outside the organization (e.g., incident response organizations, counterparts at other organizations).

**Set Written Guidelines for Prioritizing Incidents.** Incident response teams should handle each incident with the appropriate priority, based on the criticality of the affected resources and the current and potential technical effect of the incident. For example, data destruction on a user workstation might result in a minor loss of productivity, whereas root compromise of a public Web server might result in a major loss of revenue, productivity, access to services, and reputation, as well as the release of confidential data (credit card numbers, social security numbers, etc.). Because incident responders normally work under stressful conditions ripe for human error, it is important to clearly define and articulate the incident handling priority process. The incident handling priority process should include a description of how the incident response team should react under various circumstances, as well as a service-level agreement (SLA) that documents appropriate actions and maximum response times. This prioritization should facilitate faster and more consistent decision making.

### 1.1.2. PREPARING TO COLLECT INCIDENT DATA

Organizations should be prepared to collect a set of objective and subjective data for each incident. Over time, the incident data collected by the organization can be used for many ends. For example, data on the total number of hours the incident response team has dedicated to incident response activities and its cost over a particular period of time, may be used to justify additional funding of the incident response team. A study of incident characteristics may reveal systemic security weaknesses and threats, changes in incident trends, or other data that can be used in support of the risk assessment process. Another good use of the data is measuring the success of the incident response team. If incident data is collected and stored properly, it should provide several measures of the success of the incident response team. Furthermore, organizations that are required to report incident information will need to collect the necessary data to meet their requirements.

In the process of preparing to collect incident data, organizations should focus on collecting data that is actionable. Absolute numbers are not informative—understanding how they represent threats to and vulnerabilities of the business processes of the organization is what matters. Organizations should decide what incident data to collect based on reporting requirements and on the expected return on investment from the data (e.g., identifying a new threat and mitigating the related vulnerabilities before they can be exploited).

### 1.1.3. PREVENTING INCIDENTS

Preventing problems is normally less costly and more effective than reacting to them after they occur. Thus, incident prevention is an important complement to an incident response capability. If security controls are insufficient, high volumes of incidents may occur, overwhelming the resources and capacity for response, which would result in delayed or incomplete recovery, possibly more extensive damage, and longer periods of service unavailability. Incident handling can be performed more effectively if organizations complement their incident response capability with adequate resources to actively maintain the security of networks, systems, and applications. This process is intended to reduce the frequency of incidents, thereby allowing the incident response team to focus on handling serious incidents. Examples of practices that help to prevent incidents are as follows:

- Having a patch management program to assist system administrators in identifying, acquiring, testing, and deploying patches that eliminate known vulnerabilities in system software and application software;

- Hardening all hosts appropriately to eliminate vulnerabilities and configuration weaknesses – adopting the principle of least privilege and elimination of default settings. Warning banners should be displayed whenever a user attempts to gain access to a secure resource. Host should have auditing enabled to log security-related events;

- Configuring the network perimeter to deny all activity that is not expressly permitted;
- Deploying necessary software (licensed) throughout the organization to detect and stop malicious code. Malicious code protection should be deployed at the host level, the application server level, and application client level;

- Making users aware of policies and procedures on the appropriate use of networks, systems, and applications. Improving user awareness regarding incident should reduce the frequency of incident, particularly those involving malicious codes and violations of acceptable use policies.

## 1.2. DETECTION AND ANALYSIS

Detection and analysis are, for many organizations, the most challenging aspects of the incident response process, in other words, accurately detecting and assessing possible incidents - determining whether an incident has occurred and, if so, the type, extent, and magnitude of the problem. Incidents can be detected through many different means, with varying levels of detail and fidelity. Automated detection capabilities include network-based and host-based intrusion detection and prevention systems (IDPSs), antivirus software, and log analyzers. Incidents may also be detected through manual means, such as user reports. Some incidents have overt signs that can be easily detected, whereas others are virtually undetectable without automation.

In a typical organization, the thousands or millions of possible signs of incidents that occur any given day are recorded mainly by computer security software. Signs of an incident fall into one of two categories: indications and precursors. A precursor is a sign that an incident may occur in the future. An indication is a sign that an incident may have occurred or may be occurring now. Some types of indications that exist are as follows:

- The network intrusion detection sensor alerts when a buffer overflow attempt occurs against an FTP server;
- The antivirus software alerts when it detects that a host is infected with a worm;
- The Web-server crashes;
- User complain of slow access to host on the internet;
- The system administrator sees a filename with unusual characters;
- The user calls the help desk to report a threatening e-mail message;
- The host records an auditing configuration change in its log;
- The application logs multiple failed login attempts from an unfamiliar remote system;
- The e-mail administrator sees a large number of bounced e-mails with suspicious contents;
- The network administrator notices an unusual deviation from typical network traffic.

One should not think of incident detection as being strictly reactive. In some cases, the organization may detect activities that are likely to precede an incident. For example, a network IDPS may record unusual port scan activity targeted at a group of hosts, which occur shortly before a DoS attack is launched against one of the same host. The intrusion detection alerts regarding the scanning activity serve as a precursor of the subsequent DoS incident. Other examples of precursor are:

- Web-server log entries that show the usage of a Web vulnerability scanner;
- An announcement of a new exploit that targets a vulnerability of the organization’s mail server;
- A threat from a hacktivist group that the group will attack the organization.

Not every attack can be detected through precursors. If precursors are detected, the organization may have an opportunity to prevent the incident (by altering its security posture through automated or manual means).

Automation is needed to perform an initial analysis of the detected data and select events of interest for human review. Event correlation software and centralized logging can be of great value in automating the analysis process. However, the effectiveness of the process depends on the quality of the data that goes into it. Organizations should establish logging standards and procedures to ensure that adequate information is collected by logs and security software and that the data is reviewed regularly. Proper and efficient reviews of incident-related data require people with extensive, specialized technical knowledge and experience.

When a potential incident is identified, the incident response team should work quickly to analyze and validate it, documenting each step taken. The team should rapidly perform an initial analysis to determine the incident’s scope, attack methods, and targeted vulnerabilities. Performing the initial analysis and validation is challenging. The recommendations for making incident analysis easier and more effective are as below:

- Profile networks and systems;
- Understand normal behavior;
- Use centralized logging and establish a log retention policy;
- Perform event correlation;
- Keep all host clocks synchronized;
- Maintain and use a knowledge base of information;
- Use Internet search engines for research;
- Run Packet Sniffers to collect additional data;
- Consider filtering the data;
- Create a Diagnosis Matrix for less experienced staff as per Table 1;
- Seek assistance from others.

Organizations need to quantify the effect of its own incidents. To assign a severity rating for an incident, organizations should first determine the effect ratings and criticality ratings for the incident, based on Tables 2 and 3.

**Table 1. Excerpt of a Sample Diagnosis Matrix**

| Symptom | Denial of Service | Malicious Code | Unauthorized Access | Inappropriate Usage |
|---|---|---|---|---|
| Files, critical, access attempts | Low | Medium | High | Low |
| Files, inappropriate content | Low | Medium | Low | High |
| Host crashes | Medium | Medium | Medium | Low |
| Port scans, incoming, unusual | High | Low | Medium | Low |
| Port scans, outgoing, unusual | Low | High | Medium | Low |
| Utilization, bandwidth, high | High | Medium | Low | Medium |
| Utilization, e-mail high | Medium | High | Medium | Medium |

**Table 2. Effect Rating Definitions**

| Value | Rating | Definition |
|---|---|---|
| 0.00 | None | No effect on a single agency, multiple agencies, critical infrastructure |
| 0.10 | Minimal | Negligible effect on a single agency |
| 0.25 | Low | Moderate effect on a single agency |
| 0.50 | Medium | Severe effect on single agency, negligible effect on multiple agencies or critical infrastructure |
| 0.75 | High | Moderate effect on multiple agencies or critical infrastructure |
| 1.00 | Critical | Severe effect on multiple agencies or critical infrastructure |

**Table 3. Criticality Rating Definitions**

| Value | Rating | Definition |
|---|---|---|
| 0.10 | Minimal | Non-critical system, systems, or infrastructure |
| 0.25 | Low | System or systems that support a single agency’s mission (DNS servers, domain controllers), but are not mission critical |
| 0.50 | Medium | System or systems that are mission critical to a single agency |
| 0.75 | High | System or systems that support multiple agencies or sectors of the critical infrastructures (root DNS servers) |
| 1.00 | Critical | System or systems that are mission critical to multiple agencies or critical infrastructure |

This analysis should provide enough information for the team to prioritize subsequent activities, including the containment of the incident. When in doubt, incident handlers should assume the worst until additional analyses indicate otherwise. In addition to prioritization guidelines, organizations should also establish an escalation process for those instances when the incident response team fails to respond to an incident within the designated time.

The incident response team should maintain records about the status of incidents, along with other pertinent information. Using an application or database for this purpose is necessary to ensure that incidents are handled and resolved in a timely manner. The incident response team should safeguard this data and other data related to incidents because it often contains sensitive information concerning recent security breaches, exploited vulnerabilities, and users that may have performed inappropriate actions.

### 1.3. CONTAINMENT, ERADICATION AND RECOVERY

It is important to contain an incident before it spreads to avoid overwhelming resources and increasing damage caused by the incident. Most incidents require containment, so it is important to consider it early in the course of handling each incident. An essential part of containment is decision making, such as shutting down a system, disconnecting it from the network, or disabling certain system functions. Such decisions are much easier to make if strategies and procedures for containing the incident have been predetermined. Organizations should define acceptable risks in dealing with incidents and develop strategies accordingly.

Containment strategies vary based on the type of incident. For example, the overall strategy for containing an e-mail-borne virus infection is quite different from that of a network-based distributed denial of service attack. Organizations should create separate containment strategies for each major type of incident. The criteria for choosing the appropriate strategy should be documented clearly to facilitate quick and effective decision making. Examples of criteria include potential damage to and theft of resources, the need to preserve evidence, the effectiveness of the strategy, the time and resources needed to implement the strategy, and the duration of the solution.

In certain cases, some organizations delay the containment of an incident so that they can monitor the attacker's activity, usually to gather additional evidence. If an organization knows that a system has been compromised and allows the compromise to continue, it may be liable if the attacker uses the compromised system to attack other systems.

After an incident has been contained, eradication may be necessary to eliminate components of the incident, such as deleting malicious code and disabling breached user accounts. For some incidents, eradication is either unnecessary or is performed during recovery. In recovery, administrators restore systems to normal operation and (if applicable) harden systems to prevent similar incidents. Recovery may involve such actions as:

- Restoring systems from clean backups;
- Rebuilding systems from scratch;
- Replacing compromised files with clean versions;
- Installing patches;
- Changing passwords; and
- Tightening network perimeter security.

It is also often desirable to employ higher levels of system logging or network monitoring as part of the recovery process. Once a resource is successfully attacked, it is often attacked again, or other resources within the organization are attacked in a similar manner.

## 1.4. POST-INCIDENT ACTIVITY

After a major incident has been handled, the organization should hold a lessons-learned meeting to review the effectiveness of the incident handling process and identify necessary improvements to existing security controls and practices. Lessons-learned meetings should also be held periodically for lesser incidents. Questions to be answered in the lessons learned meeting include:

- Exactly what happened, and at what time?
- How well did staff and management perform in dealing with the incident? Were the documentation procedures followed? Were they adequate?
- What information was needed sooner?
- Were any actions taken that might have inhibited recovery?
- What would the staff and management do differently the next time a similar incident occurs?
- What corrective actions can prevent similar incidents in future?
- What additional tools or resources are needed to detect, analyze, and mitigate future incidents?

The information accumulated from all lessons-learned meetings, as well as the data collected while handling each incident, should be used to identify systemic security weaknesses and deficiencies in policies and procedures. Follow-up reports generated for each resolved incident can be important for evidentiary purposes, used as a reference in handling future incidents, and used in training new incident response team members. An incident database, with detailed information on each incident that occurs, can be another valuable source of information for incident handlers.

## 2. CONCLUSION

The major steps to be performed in the initial handling of an incident (*categorized*) are:

- Determine whether an incident has occurred
  Analyze the precursors and indicators,
  Look for correlating information,
  Perform research (e.g., search engines, knowledge base),
  As soon as the handler believes an incident has occurred, start documenting the investigation and gathering evidence;

- Classify the incident using the categories presented (e.g., denial of service, malicious code, unauthorized access, inappropriate usage, multiple component);
- Follow the appropriate incident category checklist.

The items address only the detection and analysis of an incident; after that has been completed, incident responders should use checklists that are geared toward a particular type of incident. The steps followed for handling incidents (*uncategorized*) that do not fit into any of the aforesaid five categories are:

- Prioritize handling the incident based on the business impact

Identify which resources have been affected and forecast which resources will be affected,

Estimate the current and potential technical effect of the incident,

Find the appropriate cell in the prioritization matrix, based on the technical effect and affected resources;

- Report the incident to the appropriate internal personnel and external organizations;
- Acquire, preserve, secure, and document evidence;
- Contain the incident;
- Eradicate the incident

Identify and mitigate all vulnerabilities that were exploited,

Remove malicious code, inappropriate materials, etc.;

- Recover from the incident.

**REFERENCES:**

1. National Institute of Standards and Technology Special Publication 800-30, *Risk Management Guide for Information Technology Systems*, July 2002.

2. Federal Information Processing Standard 199, *Standards for Security Categorization of Federal Information and Information Systems*, February 2004.

3. National Institute of Standards and Technology Special Publication 800-61, *Computer Security Incident Handling Guide*, January 2003.

4. ISO/IEC International Standard ISO/IEC 17799, *Information Technology – Code of Practice for Information Security Management*, February 2001.

5. National Security Agency (UK), The NSA Security Manual.

6. Information System Security Association (USA), *Generally Accepted Information Security Principles, Version 3.0.*

7. Julia Allen, *The CERT Guide to System and Network Security Practice*, Addison Wesley, Boston.

8. Micky Krause and Harold Tripton, *Information Security Management Handbook*, Auerbach, Boca Raton, Fl.

# A Step To Embedded Database

## A Techno Change

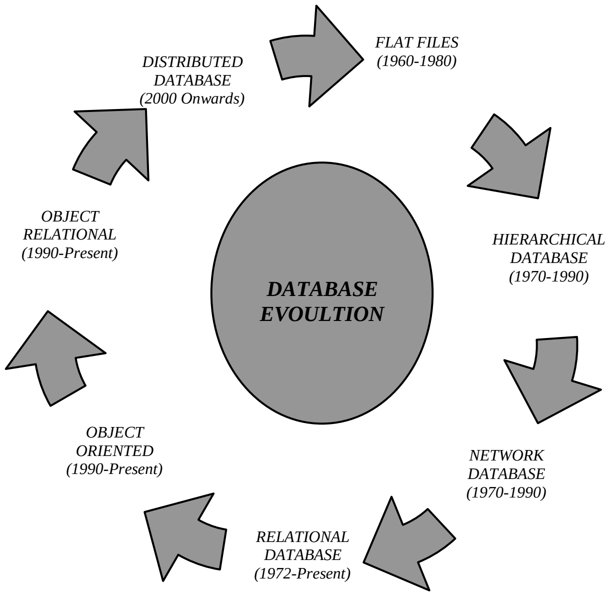
Manik Sharma
Assistant Professor & Head, PG Deptt. of computer Science and Applications
Sewa Devi SD College Tarn Taran
Manik_sharma25@yahoo.com

***Abstract***— Recent advances in device tools and connectivity have tiled the way for next generation applications that are data-driven, where data can reside anywhere, can be accessed at any time, from any client. Embedded systems are computers (microprocessors) that are enclosed (embedded) in customized hardware. An **embedded database** system is a database management system which is closely coupled with application software that requires access to stored statistics or data, such that the database system is "hidden" from the application's end-user and requires little or no ongoing maintenance. More than 20 years one could argue that since the beginning of software, embedded databases have been in existence. The operations of the embedded database are invoked by the application. The embedded database is embedded within an application either as in-line code or linked libraries unlike the traditional general purpose enterprise relational databases such as Oracle, DB2, and SQL Server etc which normally run as the separate applications that are independent of the system application. The key operational advantage of the embedded database is that using the embedded database the users and administrators are not burdened with time-consuming installations or maintenance as the database is packaged with the application and is generally self maintaining. The embedded databases can be relational, hierarchical, network model, XML based, object oriented etc. The embedded DBMS are typically used in the mobile phones, PDA's, set-top boxes, automotives etc.

***Keywords-*** *Embedded Database, DBMS, Software, Oracle, mobile phones*

## 1. Introduction

As we know that the database is defined as collection of interrelated data. Initially small databases were first developed or funded by the U.S. government for agency or professional use. In the 1960s, some databases became commercially available, but their use was funneled through a few so-called research centers that collected information inquiries and handled them in batches [1]. Recent advances in device technology and connectivity have paved the way for next generation applications that are data-driven, where data can reside anywhere, can be accessed at any time, from any client. Embedded systems are computers (microprocessors) that are enclosed (embedded) in customized hardware. Examples of embedded systems are portable medical equipment, cellular phones, or consumer electronics items. An **embedded database** system is a database management system which is closely tied with application software that requires access to stored data, such that the database system is "hidden" from the application's end-user and requires little or no ongoing maintenance [2]. More than 20 years one could argue that since the beginning of software, embedded databases have been in existence.



In other words we can say that embedded database is embedded in some another software applications. The operations of the embedded database are invoked by the application. The embedded database is embedded within an application either as in-line code or linked libraries unlike the traditional general purpose enterprise relational databases such as Oracle, DB2, and SQL Server etc which normally run as the separate applications that are independent of the system application.

The key operational advantage of the embedded database is that using the embedded database the users and administrators are not burdened with time-consuming installations or maintenance as the database is packaged with the application and is generally self maintaining. The embedded databases can be relational, hierarchical, network model, XML based, object oriented etc.

The embedded DBMS are typically used in the mobile phones, PDA's, set-top boxes, automotives etc.

**Worth of Embedded Database:** Modern embedded devices are now responsible for storing more data than ever before. Some devices get an edge on the competition by synchronizing data without interrupting normal use. Important data must not be lost to corruption caused by a power failure. For these devices, performance and reliability are critical.

## 2. EMBEDDED DBMS CHARACTERISTICS

The data access and management requirements of the applications described above are significantly different from that of traditional server DBMS. These new applications must be able to run on multiple tiers ranging from devices to servers to web and would benefit from various existing database mechanisms. However, these database mechanisms (like query, indexing, persistence) must be unlocked from the traditional monolithic DBMS and made available as embeddable components (e.g. DLLs) that can be embedded within applications, thereby, enabling them to meet the requirements described above. Such Mobile and Embedded DBMS have the following characteristics:

1. **Embeddable in applications** – Mobile and Embedded DBMS form an integral part of the application or the application infrastructure, often requiring no administration. Database functionality is delivered as part of the application (or app infrastructure). While the database must be embeddable as a DLL in applications, it must also be possible to deploy it as a stand-alone DBMS with support for multiple transactions and applications.

2. **Small footprint** – For many applications, especially those that are downloadable, it is important to minimize DBMS footprint. Since the database system is part of the application, the size of the DBMS affects the overall application footprint. In addition to the small footprint, it is also desirable to have short code paths for efficient application execution. Most of these applications do not require the full functionality of commercial DBMSs; they require simple query and execute in constrained environments.

3. **Run on mobile devices** – The DBMS that run on mobile devices tend to be specialized versions of mobile and embedded DBMS. In addition to handling the memory, disk and processor limitations of these devices, the DBMS must also run on specialized operating systems. The DBMS must be able to store and forward data to the back-end databases as synchronization with backend systems is critical for them.

4. **Componentized DBMS** – Often, to support the small footprint requirement, it is important to include only the functionality that is required by the applications. For example, many simple applications just require ISAM like record-oriented access. For these applications, there is no need to include the query processor, thereby increasing the footprint. Similarly, many mobile and mid-tier applications require only a small set of relational operators while others require XML access and not relational access. So, it should be possible to pick and choose the desired components.

5. **Automatic DBMS** – The embedded DBMS is invisible to the application user. There can be no DBA to manage the database and operations like backups, recovery, indexing, tuning etc. cannot be initiated by a DBA. If the database crashes, the recovery must start instantaneously. The database must be self managed or managed by the application. Also, embedded DBMS must auto install with the application it should not be installed explicitly (user action) or independently. Similarly when the application is shutdown, the DBMS must transparently shutdown.

6. **In-Memory DBMS** – These are specialized DBMS serving applications that require high performance on data that is small enough to be contained in main memory. In-memory DBMS require specialized query processing and indexing techniques that are optimized for main memory usage. Such DBMS also can support data that may never get persisted.

7. **Codeless database** – Portable database should be free from any threat. The executable code can become the reason for malfunctioning or destruction of data in the form virus. By eliminating any code storage in the database, we can make our database consistent and safe [4].

8. **Portable databases** – There are many applications which require very simple deployment – installing the application should install the database associated with

it. This requires the database to be highly portable. Typically, single file databases (e.g. like Microsoft Access databases) are ideally suited for this purpose. Again, there should be no need to install the DBMS separately – installing the application installs the DBMS and then copying the database file completes the application migration. With the help of portable database [5] application we can reduce the anomalies in database migration.

9. **Synchronize with back-end data sources** – In the case of mobile and cached scenarios, it must be possible to synchronize the data with the back-end data sources. In typical mid-tier (application server) caches, the data is fetched from the back-end databases into the cache, operated on, and synchronized with the back-end database.

10. **Remote management** – While mobile and embedded DBMS must be self managed, it is important to allow them to be managed remotely also, especially those on mobile devices. In enterprises (e.g. FedEX, UPS), mobile devices must be configured and managed in a manner compliant with the company

## 3. Embedded software in India

Typically software for embedded systems need to have a very small footprint (i.e. be able to run in a small amount of memory) and often have to work in real-time. Companies here in India offer specialized operating systems and languages, which make this possible. These companies ensure the designing, developing, and testing of software for embedded systems and components meet specific customer requirements. They use diverse, real-time operating systems, devices and platforms and associated embedded tools and technologies. The various Embedded Software domains[5] in India are:

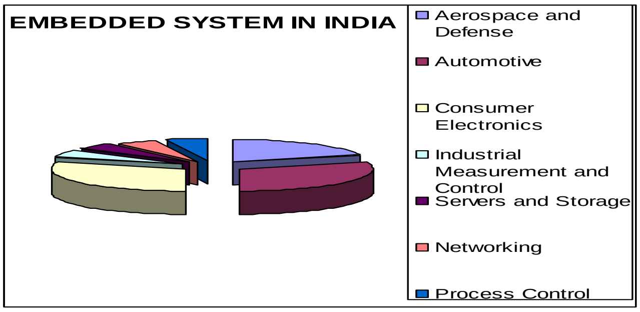


Figure2: Embedded System in India

The various embedded database system in commercial market are ElevateDB, Interbase, Oracle Berkley DB, SolidDB etc.

## 4. Conclusion

Embedded databases differ from typical databases such as DB2, Oracle, and SQL Server in that it is entirely embedded into the application or hardware device in such a way that the user has very little knowledge, if any, of its existence. Users and administrators are now free from the huge tension of installing and maintaining the database because the database is closely bundled with the application and should be self maintaining. Embedded databases are potable in nature and meant to run on many different platforms with various programming interfaces. It also help in reducing engineering and quality assurance cost, eliminate some support cost in setup and implementation. The nature of embedding databases' instruction sets being linked specifically within and for a specific application gives them a small footprint. Because embedded database reduces the instruction set hence it allows them to achieve performance that is hard to beat. The above discussion end with that future is of embedded database. Besides, embedded industry will flourish in the area of automotive, industrial, consumer electronics in coming year.

## REFERENCES

[1] Microsoft Encarta Library 2005.
[2] Graves, Steve. "COTS Databases For Embedded Systems", *Embedded Computing Design* magazine, January, 2007. Retrieved on August 13, 2008.
*[4] TinyDB: http://telegraph.cs.berkeley.edu/tinydb/.*
*[5] http://msdn.microsoft.com/en-us/library/ff647179.aspx (online)*
*[6]* Ramachandra Budihal, Emerging trends in embedded systems and applications (online)
*http://www.eetimes.com/discussion/other/4204667/Emerging-trends-in-embedded-systems-and-applications*

# Elucidation of Protein Interaction via Google and Gene Ontology

B.V.Subba Rao[1], Dr.K.V.Sambasiva Rao[2]

[1]Associate Professor, Department of IT, P.V.P Siddhartha Institute of Technology, Vijayawada-7,India.
Email ID: bvsrau@gmail.com, Ph:9440109139.
[2]Principal, M.V.R. College of Engineering and Technology, Vijayawada, Krishna Dt., A.P, India
Email ID: principal@mvrcoe.ac.in, Ph:9440115556.

***Abstract***: In the track of the growing quantity of biomedical text, there is a need for regular extraction of information to support biomedical researchers. Due to condensed biomedical information databases, the extraction cannot be done straightforward using dictionaries, so several approaches using associated rules and machine erudition have previously been proposed. Our work is motivated by the earlier approaches, but is novel in the sense that it combines Google and Gene Ontology for annotating protein connections. We got promising empirical results - 57.5% terms as valid GO annotations, and 16.9% protein names in the answers provided by our system ProG. The total error-rate was 25.6% consisting mainly of overly general answers and syntactic errors, but also including semantic errors, other biological entities and false information sources.

***Keywords:*** Biomedical Literature, Data Mining, Gene Ontology, Google API.

## 1. Introduction

With the growing importance of precise and up-to-date databases about proteins and genes for research, there is a need for efficient ways of updating these databases by extracting information from biomedical research text [8], e.g. those indexed in MEDLINE. Examples of information resources containing such information are LocusLink, UniGene and Swiss-Prot for protein info and the Gene Ontology for semantic labels. Due to the huge and rapidly growing amounts of biomedical literature, the extraction process needs to be more automatic than previously. Current extraction approaches have provided promising results, but they are not sufficiently accurate and scalable. Methodologically all the suggested approaches belong to the information extraction field [3], and in the biomedical domain they range from simple automatic methods to more sophisticated, but slightly more manual, methods. Good examples are: Learning relationships between proteins/genes based on co-occurrences in MEDLINE abstracts [9] manually developed information extraction rules, information extraction (e.g. protein names) classifiers trained on manually annotated training corpora [12], and classifiers trained on automatically annotated training corpora.

### A. Research Hypothesis

Internet Search Engines such as Google, Yahoo MSN, Bing, Alta Vista and Ask Me Search Engines are the world’s largest readily available information sources, also in the biomedical domain. Based on promising results from recent work on using Google for semantic annotation of biomedical literature, we are encouraged to investigate if Google can be used to find protein interactions that match the Gene Ontology (GO). This leads to the hypothesis: Can Internet Search engines such as Google be used to annotate protein interactions in the Gene Ontology framework.

The rest of this paper is organized as follows. Section 2 describes the materials used, section 3 presents our method, section 4 presents empirical results, section 5 describes related work and section 6 describes conclusion and future work.

## 2. Materials

See fig. 1 for an overview of the system. As input for our experiments we used the following:

10 proteins that is already well-known to our biology experts. 37 verb-templates suggested by Martin.

### A. Proteins

The following proteins were used as input to the system. Proteins user are ‘EGF’, ’TNF’, ’CCK’, ’gastrin’, ‘CCKBR’, ‘CREB’ and ‘CREM’.

In addition, each protein is also described by several other names or synonyms in the literature. E.g. gastrin is also known as ‘g14’, ‘g17’, ‘g34’, ‘GAS’, ‘gast’, ‘gastrin precursor’, ‘gastrin 14’, etc. So our biologists compiled a list of roughly 10 synonyms for each protein, giving us about 100 terms total to annotate.

### B. Interaction Verbs

We selected our interaction verb templates from table 1 in. They had a list of 44 verbs, but we chose to use only 37 of these verbs. The reason for this is that we are focusing on simple statements like “gastrin activates”, with the object of the verb following directly after the verb template. The following table shows the original list of verbs, with the removed ones in parenthesis.

Verb templates used are acetylates, activates, binds, blocks, bonds, degrades, hydrolyses, increases, interacts with, mediates, phosphorylates, reacts with, releases, stimulates, transforms, triggers, upregulates.

## 3. Our Approach

We have taken a modular approach where every sub module can easily be replaced by other similar modules in order to improve the general performance of the system. There are five modules in the system. The first one sets up the search queries, the second runs the queries against Google, the third one tokenizes the results, the fourth parses the tokenized text, and the fifth and last module extracts all the results and presents them to the human evaluators. See figure 1.

### A. Data Selection

N (=100) protein names are combined with M (=37) verb templates, giving a total of N x M (3700) queries to run against Google.

### B. Google

The queries are fed to the PyGoogle module which allows 1000 queries to be run against the Google search engine every day with a personal password key. In order to maximize the use of this quota, the results of every query are cached locally, so that each given query will be executed only once. If a search returns more than ten results, the resultset can be expanded by ten at a time, at the cost of one of the 1000 quota-queries every time. We decided to use up to 30 results for each query in this experiment.

### C. Tokenization

The text is tokenized to split it into meaningful tokens, or "words". We use a simple WhiteSpaceTokenizer from NLTK, where every special character (like ( ) " ' - , and .) is treated as a separate token.

### D. Parsing

Each returned hit from Google contains a "snippet" with the given query phrase and approximately ten words on each side of it. We use some simple regular grammars to match the phrase and the words following it. If the next word is a noun it is returned. Otherwise, adjectives are skipped until a noun is encountered, or a "miss" is returned.

### E. Expert Evaluation

The results were merged so that all synonyms were treated as if the main protein name had been used in the original query. Then the results were put into groups (one group for each protein-verb pair) and sorted alphabetically within that group. These results were then presented to the biologists, who evaluated the usefulness of our results from Google.

## 4. Empirical Results

Fig. 2 and 3 show the results. The first one shows that more than half of the extracted terms were terms that could be used to annotate the given protein around one fifth of the results contained an identifiable protein name that could be stored as a protein-protein interaction. Only one quarter of the terms were deemed not useful. The different kinds of "not useful"-errors can be read out of fig. 3.

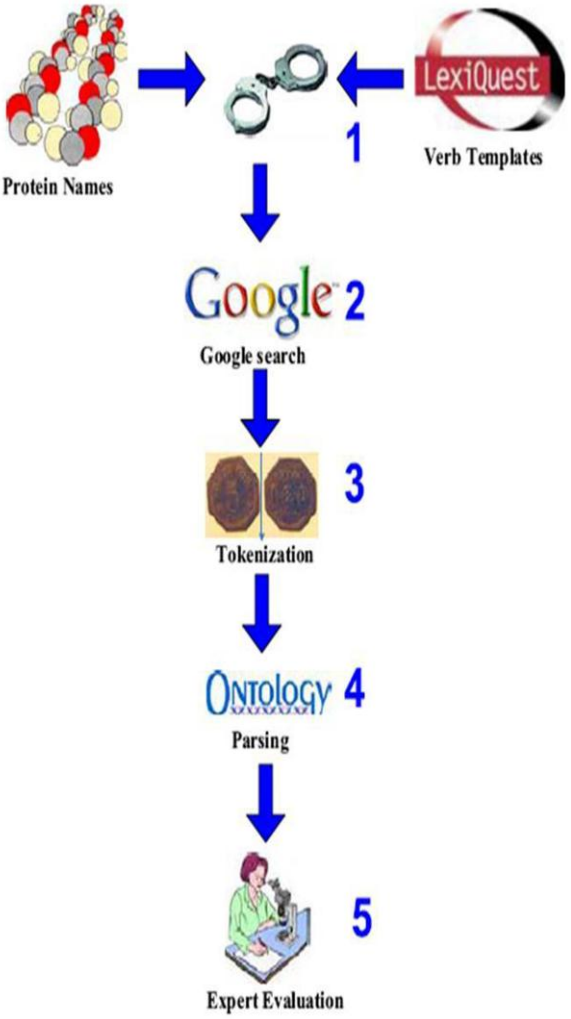


Fig. 1. Overview of our Approach (named ProG) according to the Gene Ontology (GO).

## 5. Related Work

Our specific approach was on using Google and Gene Ontology for annotating protein interactions. We haven't been able to find other work that does this, but the closest are Dingare et al., that uses results from Google search as a feature for a maximum entropy classifier used to detect protein and gene names [5, 6] and our previous work on semantic annotation of proteins (i.e. tagging of individual proteins, not their GO relation). Google has also been used for semantic tagging outside of the biomedical field, e.g. in Cimiano and Staab's PANKOW system [2] and in [4, 7, 10, 11].

A comprehensive overview of past methods for protein-related information extraction is provided in.

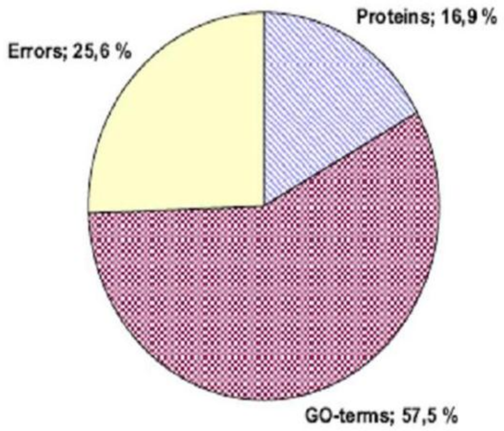


Fig. 2. Main Results

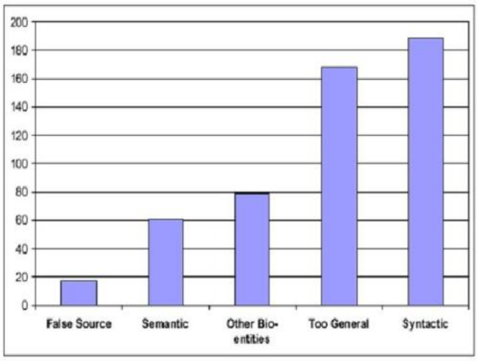


Fig. 3. Breakdown of Errors

## 6. Conclusion And Future Work

This paper presents a novel approach - ProG - using Google to find semantic (GO-) annotations for specific proteins. We got empirically promising results - 57.5% semantic annotation classes, and 16.9% protein names in the answers provided by ProG. This means that 74.4% of the results are useful. This encourages further work, possibly in combination with other approaches (e.g. rule based information extraction methods), in order to improve the overall accuracy. In the similar task of protein name identification, recently presented precision scores ranges from 70 to 75% [1]. Hopefully, more advanced methods will greatly reduce the number of errors (useless information), which is currently at 25.6%. Disambiguation is another issue that needs to be further investigated, because sometimes different search-results are really just one single identity, because of synonyms and acronyms for example. Other opportunities for future work include:

– Improve tokenization. Just splitting on whitespace and punctuation characters is *not* good enough. In biomedical texts non-alphabetic characters such as brackets and dashes need to be handled better.

– Search for other verb templates using Google. E.g. Which templates give the best results, and what about negations ("does not activate ...")

– Investigate whether the Google ranking is correlated with the accuracy of the proposed semantic tag. Are highly ranked pages better sources than lower ranked ones?

– Test our approach on larger datasets, e.g. using *all* the returned results from Google.

– Combine this approach with more advanced natural language parsing techniques in order to improve the accuracy.

– In order to find multiword tokens, one could extend the search query *("X activates")* to also include neighboring words of X, and then see how this affects the number of hits returned by Google. If there is no reduction in the number of hits, this means that the words are "always" printed together and are likely constituents in a multiword token. If you have only one actual hit to begin with, the certainty of the previous statement is of course very weak, but with increasing number of hits, the confidence is also growing. – In this experiment very crude Part Of Speech (POS) tagging is done, so our results can be seen as a baseline for this kind of experiment. In the future we want to improve the results, for example by utilizing better grammars, and more advanced natural language understanding techniques.

## References

[1] Razvan Bunescu, Ruifang Ge, Rohit J. Kate, Edward M. Marcotte, Raymond J. Mooney, Arun Kumar Ramani, and YukWah Wong. Comparative Experiments on Learning Information Extractors for Proteins and their Interactions. *Journal Artificial Intelligence in Medicine: Special Issue on Summarization and Information Extraction from Medical Documents (Forthcoming)*, 2004.

[2] Philipp Cimiano and Steffen Staab. Learning by Googling. *SIGKDD Explorations Newsletter*, 6(2):24–34, December 2004.

[3] J. Cowie and W. Lehnert. Information Extraction. *Communications of the ACM*, 39(1):80–91, January 1996.

[4] Stephen Dill, Nadav Eiron, David Gibson, Daniel Gruhl, R. Guha, Anant Jhingran, Tapas Kanungo, Sridhar Rajagopalan, Andrew Tomkins, John A. Tomlin, and Jason Y. Zien. SemTag and seeker: bootstrapping the semantic web via automated semantic annotation. In *Proceedings of the Twelfth International World Wide Web Conference, WWW2003*, pages 178–186. ACM, 2003.

[5] Shipra Dingare, Jenny Finkel, Christopher Manning, Malvina Nissim, and Beatrice Alex. Exploring the Boundaries: Gene and Protein Identification in Biomedical Text. In *Proceedings of the BioCreative Workshop*, March 2004.

[6] Shipra Dingare, Jenny Finkel, Christopher Manning, Malvina Nissim, Beatrice Alex, and Claire Grover. Exploring the Boundaries: Gene and Protein Identification in Biomedical Text. Submitted to BMC Bioinformatics, 2004.

[7] Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. Unsupervised Named- Entity Extraction from the Web: An

Experimental Study. Submitted to Artificial Intelligence, 2004.

[8] Jun ichi Tsuji and Limsoon Wong. Natural Language Processing and Information Extraction in Biology. In *Proceedings of the Pacific Symposium on Biocomputing 2001*, pages 372–373, 2001.

[9] Tor-Kristian Jenssen, Astrid Lægreid, Jan Komorowski, and Eivind Hovig. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, 28(1):21–28, May 2001.

[10] Vinay Kakade and Madhura Sharangpani. Improving the Precision of Web Search for Medical Domain using Automatic Query Expansion. Online, 2004.

[11] Udo Kruschwitz. Automatically Acquired Domain Knowledge for ad hoc Search: Evaluation Results. In *Proceedings of the 2003 Intl. Conf. on Natural Language Processing and Knowledge Engineering (NLP-KE'03)*. IEEE, 2003.

[12] B.V.Subba Rao, Dr.K.V.Sambasiva Rao, Semantic Explanation of Biomedical Literature using Google. *In proceedings International conference on Websciences-2009, pages 452-457*.

**Author's Biographies**

**B.V.Subba Rao**, presently working as Associate Professor in P.V.P Siddhartha Institute of Technology Vijayawada, affiliated to Jawaharlal Nehru Technological University. He received his M.Tech degree with distinction in Computer Science and Engineering from Acharya Nagarjuna,University. He received Gold medal in his post graduate studies. He is pursuing Ph.D in Computer Science and Engineering at Acharya Nagarjuna University, Guntur. He has guided 30 post Graduated and 40 graduate projects. He has published 4 papers (National / Conference Proceedings) and has Academic participation in 24 International / National Seminars / workshops and Conferences. He is a member of Computer Society of India (CSI), Association for Computing Machinery (ACM), and Indian Society for Technical Education (ISTE). His current research interests are in the areas of Artificial Intelligence, Natural Language Processing, Information Retrieval systems and Bioinformatics.

**Dr.K.V.Sambasiva Rao**, presently working as a Principal of MVR College of Engineering and Technology, Paritala. He pursued his M.E from BITS, Pilani and Doctorate from IIT Delhi. He has a total of 21 years of rich experience comprising teaching, research and industry. He has published 4 books, 18 papers in international and national journals. He has conducted numerous national conferences, workshops with the support of AICTE, DST and other government bodies. He has given more than 50 seminar talks at various technical institutions. He has guided 25 Masters level projects and is research director for 11 Ph.D candidates. His biography was included in MARQUI'S INTERNATIONAL, New Jersey, USA "Who is who in the World" in the year 1999 and was awarded "Outstanding achievement award" by International Biography Centre, Cambridge, UK. He is the life member of 3 professional bodies.

# Reap Information through a Procedure Framework in an Electronic Process Steer

G.Sivanageswara Rao[1], Dr.K.Krishna Murthy[2] , B.V.Subba Rao[3]

[1]Reader, PB Siddhartha PG Centre, Vijayawada. E-mail:sivanags@india.com
[2]Professor and Director, Dept. of PG Studies, PB Siddhartha PG Centre, Vijayawada.
[3]Associate Professor, Dept. of IT, PVP Siddhartha Institute of Technology, Vijayawada.

***Abstract:*** A key leverage for small software consultancy companies is the collective knowledge possessed by their consultants. There have been some studies in the literature on how to harvest and transfer this knowledge, but most studies are aimed at large multinational corporations. In this paper we describe an ongoing research project, aimed at improving knowledge sharing in a small software consultancy company through the use of a method framework in an electronic process guide coupled with an experience repository.

## 1. Introduction

Small software consultancy companies have to leverage their position in the market to stay ahead of their competitors. One way to achieve this is by providing their customers with tailored solutions to their problems. They can do this by drawing upon the collective knowledge of their consultants. When the company is small and the consultants are spread over the sites of many customers, it becomes difficult to gain access to and draw upon the collective experience of all the co-workers. Consequently the solutions provided by the consultants might not be of sufficient quality to make their customers return to the company when they need consultants for a new project.

One solution to the problem with a dispersed workforce is experience repositories. A lot of research has gone into this field, however most of this research has been focused on large companies and little data exists on the application of this in small companies [1, 2]. In [3] the authors examine challenges facing small businesses when implementing knowledge management efforts. Small businesses are particularly vulnerable to knowledge erosion, yet they seldom have the time and resources needed to implement the knowledge management programs described for larger companies. However, the authors suggest that small businesses can benefit just as much from well thought out knowledge management efforts.

According to [1], which describes the successful use of an experience repository in a small software company, detailed data on its use and structure can be used to better understand how experience supports activities in the company. This can in turn lead to improvements in experience management concepts, techniques and tools. In this research report we describe our work in a small software consultancy company that wishes to manage their knowledge through a method framework implemented in an Electronic Process Guide (EPG) coupled with an Experience Repository (ER).

## 2. Latest Trends

Most of an organization's corporate knowledge is contained in documents or in the minds of its human resources. To make effective use of this corporate knowledge, organizations must be able to access, harvest, organize and redistribute it. Some of the latest trends are Contextual Knowledge Harvesting, Content based knowledge harvesting, Natural language based knowledge harvesting, XTM and Knowledge harvesting using Seman Text.

## 3. Context

The company we investigated currently has 17 employees. Their main activities are hiring out consultants as developers, developing complete solutions for customers and hiring out consultants as advisors for selecting technology, strategy or process. Typically, no more than four to five consultants are at any time working for the same customer.

The managers of the company wish to leverage the company in the market by providing solutions to the problems of their customers. The solutions should make them stand out and increase the probability that the customers later returns with new projects. In order to do this, they wish to foster an environment were all ideas and knowledge are shared freely among the employees, and where the employees can draw upon the experience of each other to provide good services to their customers. This work is difficult since a lot of the employees at any given time are out at the site of customers where they don't have direct access to their colleagues. To remedy this situation they wish to collect the experience of their employees in an Experience Repository (ER). This will allow their employees to have easy access to the experience of their coworkers.

## 4. Method

Due to the cooperative nature of this research project, we decided to adopt action research as our approach. The most prevalent description of action research is found in [4]. The approach requires the establishment of a client-system infrastructure or research environment. In our case this was already taken care of through the researchers' and company's involvement in a mutual research program. The approach further specifies five identifiable phases, which are iterated: diagnosing, action planning, action taking, evaluating and specifying learning. This paper sums up our work and findings from the initial phases and what effect this has had on the development of the new tool. The plans for the next phases are outlined up in section 6: Future work.

For the initial diagnosing phase, we decided to use semi structured interviews. We scheduled interviews with 12 of the employees. The interviews were carried out using an interview guide. Basically we wanted answers to three questions: What was the current approach to knowledge sharing, what should the new tool contain, and what kind of functionality should it provide? All of the interviews were taped using a Dictaphone and were subsequently transcribed. The material was then coded and analyzed using the constant comparison method and the NVivo tool [5].

The problem with the adopted approach is that our results will be difficult to generalize due to our single case. Rather they will contribute to the understanding of the concepts of Experience Repositories. If the results from our study should coincide with the research literature some generalization might be possible.

## 5. Interview Results

The company seemed to have a good environment for informal sharing of experience in that people knew one another and knew whom to contact if they were stuck. There did not seem to be much formal gathering of experience. If experience from a project was collected, it was mostly done in an ad hoc manner, and it was not easily available. The gathering of experiences today was mostly done through private initiative and saved for personal use. Lately a few employees had begun using post mortem analysis [6] at the end of their projects, but they did not have a place to structure and access this information. The fact that a lot of work was done at the site of customers was also seen as a hindrance to collecting project experience. It seemed to be easier to get help with technical problems than problems related to process. More structure and information related to process was seen as desirable. When asked about what information they wanted the new tool to contain, the employees provided us with a myriad of elements. A few, however, was mentioned more often than the others: document templates, patterns, a good process, help with customer relations and practical experience. Document templates were seen as potential help to increase productivity. Both inexperienced and experienced project managers saw a benefit from having a set of standardized templates in order to save time on documentation. Patterns were also mentioned as something that should be readily available. Good ideas and smart solutions that other people had thought of were worth repeating. However, the employees stressed the need for trust. It was important for them to know that a pattern could actually deliver what it promised. A good development process and the need for help with questions related to process was often mentioned during the interviews. This need was considered especially important for the start-up of new projects. Inexperienced project managers expressed a need for a process that would help and guide them through the initial phases. Experienced managers expressed the need for a process that would help them keep on track throughout the project. A well-defined process was also seen as something they could market to their customers to gain an edge over their competitors. The employees often mentioned the need for guidelines and advice on how to improve customer relations. There was a broad agreement that more customer involvement would enhance the quality of the end product. The employees agreed that there had not been a lot of focus on this in the past and that guidelines for this would be most welcome in the new method framework. When it came to choosing a process, a template or a pattern, the employees would like to know what kind of experience others had made when using these items. They saw a great potential in linking the experience of the company's developers to templates, patterns and processes, in order to be able to assess them for their own projects based on their colleagues' experience.

## 6. Initial Work and Challenges Ahead

After the initial interviews we moved on to the action-planning phase of our research. This phase consisted of meetings with the company where we presented the result of our interviews. The interviews indicated that there was a demand for a tool that would

help the employees to share and structure their experience, especially experience surrounding the development process. It also indicated that the culture of the company supported free sharing of information and experience, and that the employees saw the benefits of using such a tool as the management was suggesting. With the support from the employees established, we arranged a discussion on the functionality and the content of the new tool. It was decided that the company should develop an empty method framework tailored to the development process of the company. This framework would be implemented in a dynamic EPG, which would then be coupled to an ER. The employees would use this tool to enter their experience related to roles, artifacts and activities. The goal is to create a process guide based on the collective experience of all the employees in the company, which can then be used to increase the quality and consistency of their work. Both the decision to couple the ER to the process of an EPG and making the tool highly interactive to enable fast feedback is supported by  which describes good practices regarding ER and [2] which describes a successful implementation of an EPG/ER After the meeting where this was discussed, we moved on to the action-taking part of our research. The company put one consultant on the project of working out a method framework. The framework was based on the Rational Unified Process (RUP), and was tailored to the company's process. During this process the input of both employees and scientists was sought in order to make the framework as similar to the current practice as possible. One of our main challenges in the time ahead will be to keep the ER alive. An ER that is not used by the developers is of no value to the company. Experience from other ER initiatives has shown that there are three factors that influence the use of an ER: □□The ER must contain a minimum amount of experiences that can be searched. The amount of experience available is critical. If there is little experience available in the ER, the developers will neither use it nor contribute their own experiences to it.

The experience that is found must be considered to be relevant for the developers in their day-to-day work. It must help them to do a better job and it must be up to date. One of the most de-motivating things that can occur when using an ER is to find an experience with and interesting title but with outdated contents.

It must be possible to establish a community of practice based on the ER contents. This means that not only must the experience be relevant – it must be possible to discuss, and augment existing experiences, that is; the ER must work as a forum where people can exchange ideas. All of these mechanisms are used to keep up the interest for the ER among the developers. On the other hand, the interest can only be kept if the content is good. In order to meet these challenges we will use several strategies. The most important mechanisms to achieve our goals are to: Keep the ER open. As a consequence of this, everybody can add his or her own experiences. There will only be one restriction – all input must be traceable to the person that contributed it.

Build discussion treads. These are important both to keep the experiences up-todate and to keep the community of practice alive.

## 7. Future Work

When the framework is finished and implemented in the EPG/ER tool it will be presented to the employees. After this, the employees will enter into a period of filling up the framework with relevant experience. The next challenge for the scientists will be to come up with good methods for extracting most of the experience of the employees in a way that is not too intrusive to the regular work of the company, yet still captures the most crucial knowledge.

After an initial trial period the tool will be approved for use in projects. The role of the scientists then switches to an observational role. We plan on following the use of the EPG/ER for two years (the remaining period of our research project). By collecting information along the way and comparing it with the research literature, we hope to be able to ascertain how successful the knowledge initiative have been for the company and how it might apply to companies in similar contexts.

## References

[1] Louise Scott, Ross Jeffery: *The Anatomy of an Experience Repository*, Proc. International Symposium on Empirical Software Engineering, 2003

[2] Felicia Kurniawati, Ross Jeffery: *The Long-term Effects of an EPG/ER in a Small Software Organisation*, Proc. Australian Software Engineering Conference, 2004

[3] Wickert Anja, Richard Herschel: *Knowledge management issues for smaller businesses*, Journal of Knowledge Management, Vol 5, no. 4, pp. 329-337, 2001

[4] Susman G., Evered R.: *An assessment of the scientific merits of action research*, Administrative Science Quarterly, 23(4), pp. 582 – 603, 1978

[5] Web:http://www.qsrinternational.com last visited 06.09.04

# Evaluating The Effectiveness of Global eXtreme Programming Framework through its Artifacts

Ridi Ferdiana[1], Lukito Edi Nugroho[1], Paulus Insap Santoso[1], Ahmad Ashari[2]

[1]Department of Electrical Engineering and Information Technology, Gadjah Mada University
[2]Department of Computer Science and Electronics, Gadjah Mada University
Yogyakarta - Indonesia
ridi@te.gadjahmada.edu, lukito@mti.ugm.ac.id, insap@te.gadjahmada.edu, ashari@ugm.ac.id

***Abstract***: Every framework needs an evaluation. The evaluation determines the capability of the framework to support the phases, activity, roles, and product of the software engineering lifecycle. It is also to make sure that the proposed framework can be adopted in real project and performed better than its predecessor. However, the standard of framework evaluation is somewhat limited. Development team chooses the framework based on the previous real project experience and others case studies. This paper will introduce a novel approach to evaluate the agile software engineering framework through its artifacts. An example of a reference model developed for the Global eXtreme Programming framework (GXP). GXP is used a case study to illustrate how the approach may be applied.

***Keywords***: Software engineering Framework, Evaluation, Artifacts, Agile, Global eXtreme Programming, Global Software Development

## 1. Introduction

Software engineering evaluation usually happens in organization. The term organization is meant to apply to a software development group in a company who wants to build software. There are several software engineering frameworks that exist. Therefore, the organization usually does the evaluation to choose which framework that appropriates for the project.

Evaluations are context-dependent, which mean evaluations result can't become to be the best in all circumstances. It's possible that an evaluation in one organization being identified as superior, but similar evaluation in another organization would come to a different conclusion [2]. For example, suppose two organizations compare the multi-site development using the Global Software Development (GSD) framework [7]. One organization can say that GSD provide better communication values. Therefore, GSD is enough for them. However, other organization might say that GSD provides better in the term's artifacts. Hence, the difference in the results of the evaluation might be due to properties that assessed not the method that used.

This paper specifically proposes the evaluation method for the organization that implement agile for their multi-site software development. The paper chooses GXP framework as a framework that combines agile process, XP method, and multi-site development model [1].

One of the valuable information that stored in the multi-site development is artifacts. Artifacts are documentation that lives in a software development project. In multi-site software development, artifacts hold ultimate sources for the project information because of several reasons such as follows.

1. Rather than source codes that only understand by the developers, artifacts can be learnt by any stakeholders in projects.
2. Artifacts can be updated together by any people who join the project in multi-site development.
3. Artifacts can works as offline reference when the peers in the site team can't work together during the geographical or time zone difference.

Based on those facts this paper will use artifacts as an evaluation object.

The results of the evaluation provide a comparison context. Evaluation context in the research is a comparative evaluation. Therefore, it assumes that there are several alternative ways to do the same thing and identify which of the alternatives is the best in specific circumstances. In addition, comparison will be made against an ideal framework. This paper will compare the GXP framework with GSD Framework.

The rest of paper is organized as follows. First, we discuss the existing solution in the software engineering framework evaluation. Secondly, we describe our research approach to synthesize and do the evaluation. The research then reports the result by a discussion of the implication of those results, limitation of the work and future research directions.

## 2. Current Research Solution

In order to get a valid evaluation results, this research will follow prevailing evaluation method called DESMET [2]. DESMET method is intended to help an evaluator to plan and execute evaluation exercise, which is concerned with the evaluations methods and tools. There are several evaluation types in DESMET, which are establishing measurable effects of using objects (Quantitative) or establishing object appropriateness (Qualitative). There are also circumstances which the hybrid (Qualitative and Quantitative) method is used for selected objects. DESMET evaluation is organized through three different ways, which are formal experiment, case study, and survey. Each way can be executed through quantitative, qualitative, or hybrid approach. Figure 1 displays the DESMET evaluation process.

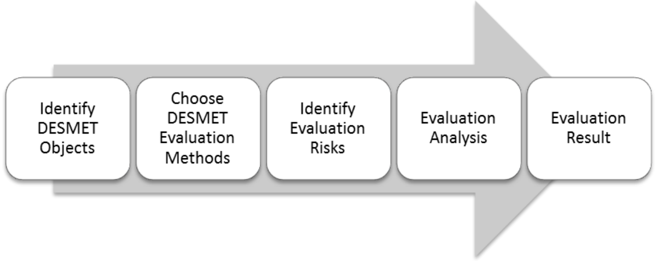


**Figure 1.** Framework syntax block diagram

### 2.1 Identify DESMET objects and evaluation method

In the research context, objects are the frameworks that are evaluated. The objects are evaluated through an evaluation method. DESMET provides several evaluation methods. Table 1 provides the evaluation method that can be selected by the evaluator.

**Table 1.** DESMET Evaluation method [3]

| Evaluation method | Conditions favoring method |
|---|---|
| **Quantitative experiments** | Benefits clearly.<br>Staff available for taking part in experiment.<br>Method/tool related to single task<br>Benefits directly measurable from task output.<br>Relative small learning time.<br>Desire to make context independent. |
| **Quantitative case studies** | Benefits quantifiable on a single project.<br>Benefit quantifiable prior to product refinement.<br>Stable development procedures.<br>Staff with measurement experience.<br>Timescales for evaluation equal with normal projects. |
| **Quantitative surveys** | Benefits not quantifiable on a single project.<br>Existing database of project achievements.<br>Projects with experience of using method/tool. |
| **Feature analysis – screening** | Large number of methods/tools to assess.<br>Short timescales for evaluation exercise. |
| **Feature analysis case study** | Benefits difficult to quantify.<br>Benefits observable on a single project.<br>Stable development procedures.<br>Tool/method user population limited.<br>Timescales for evaluation equal with normal projects. |
| **Feature analysis experiment** | Benefits difficult to quantify.<br>Benefits directly observable form task output.<br>Relative small learning time.<br>Tool/method user population very varied. |
| **Feature analysis survey** | Benefits difficult to quantify.<br>Tool/method user population very varied.<br>Benefits not observable on single project.<br>Projects with experience of using method/tool. |
| **Qualitative effects analysis** | Available of expert opinion assessments of methods/tools.<br>Lack of stable development procedures.<br>Requirement to mix and match method/tool.<br>Interest in evaluation of generic methods (process)/tool. |
| **Benchmarking** | Method/tool not human-intensive.<br>Output of method/tool able to be ranked. |

Since DESMET provides several ways to evaluate the objects, it needs to identify and select which evaluation method is the most appropriate. DESMET provides specific criteria that use to determine research circumstances, which are:

1. The evaluation context.
2. The nature of the expected impact of using the objects.
3. The nature of the objects to be evaluated.
4. The scope of impacts of the objects.
5. The maturity of the objects.
6. The learning curve associated with the objects.
7. The measurement capability of the organization undertaking the evaluation

The evaluation context of the research is to monitor changes as part of the process improvement program which can involve the evaluation of the proposed framework. The nature of the impact of the research is qualitative impact e.g. better visibility of progress, rapid development and better cost efficiency. The impacts are developed through Qualitative Effects Analysis or Feature Analysis that described in DESMET.

The scope impact of the evaluation has two majors dimension, which are product granularity and extent of impact. Product granularity identifies that the framework is applied to whole development life cycle. The extent of impact of the framework is likely to be felt over the GSD process work flows which are requirement engineering, project planning, architecture design, and product development.

DESMET encourages that the maturity of the framework indicates the extent to which there is likely to be information about it readily available. Since the framework is not used on commercial projects, DESMET states that there would be not sufficient information about the object to warrant a survey (quantitative or qualitative).

Learning curve discusses the time it would take by the organization to learn the objects. Learning time is defined into two aspects, which is time required to understand the principles and time to become proficient in its use. This learning curve related directly with the maturity of the organization. DESMET assumed that the evaluation capability is divided on 4-point ordinal scales, which are:

1. Level 1: Severely limited evaluation capability. The organization does not have well-defined standards.
2. Level 2: Qualitative evaluation capability. The organization has well-defined standards for software development, and adherence to those standards is monitored.
3. Level 3: Quantitative and qualitative evaluation capability. The organization has well-defined standards for software development, and adherence to those standards is monitored.
4. Level 4: Full evaluation capability. The organization has well-defined standards for software development, and adherence to those standards is monitored.

### 2.2 Identify Evaluation Risks

DESMET implicitly reports several risks that might occur when choosing its evaluation method. Relative risk, cost risk, and human risk are the main risks that happen during its evaluation [4]. DESMET identified two risks elements, which are false positive (The results may imply that a method or tool is beneficial when it is not) and false negative (The results may imply that a method or tool is not beneficial when it is beneficial). The extent of the risk varies for the different methods. The research should ensure that the evaluation method which it has been advised to use will give sufficient confidence that the results of the evaluation are correct. In this setting, "sufficient" must be assessed in relation to the potential cost of an incorrect decision. An incorrect decision would be regarded as serious if a large investment decision was affected, or the framework was to be used to produce safety-critical applications.

Relative risk is the first risk that identified overall risk that exists in selected method. DESMET associated the relative risk into five classes, which are very low, low, medium, high, and very high. Selected method in the research which is featuring analysis case studies is identified as the method with low relative risk. The risk will be lowered when replication and randomization are done with additional experiment. However, the feature analysis case studies identified still subjective by the organization that used the framework.

Cost risk is the risk that should be proportional to the consequential benefits if the framework were judged beneficial and introduced, less the expected investment and introduction costs. As an initial crude guideline, the relative cost of the various evaluations methods are shown as high, medium, low, very low. Feature analysis experiment has high relative cost rather than the others. The cost of experiments is due to a number of different staffs undertaking what may be non-revenue-producing work. The feature analysis experiments have a cost basis like a staff cost and direct cost. Staff cost is cost, which is expanded to define the experiment, devise a scoring method, familiarize the evaluator(s) with the framework, complete the feature analysis questionnaire, collate a result, and produce evaluation reports. Direct costs arise from hardware need, software (including the proposed tools from the framework), and support training for the method.

The attitude and motivation of participants in an evaluation exercise can misrepresent its results. The distortion effects come about because the behavior of staff participating in an evaluation exercise is at adjustment with their working attitude. It can be over-optimistic assessment or pessimistic outcomes. Those behaviors also called as sociological effects [6].

DESMET provides some recommendations that to reduce the risks that happen because of relative risks, cost risks, and human risks which are.

1. Doing the experiment separately from the real project with the separately budget.
2. Successful completion of the experiment project is a formal success of the experiment success.
3. The experiment should be treated as a real project. Therefore, the result will be as real as the real projects.
4. The roles and responsibilities is defined and understood by the people although they are in blind mode.

Those recommendations will be blended through further evaluation analysis that described in the following section.

### 2.3 Evaluation analysis

In its simplest form, feature analysis provides yes/no response to the existence of a particular property. For analogy example, considers when an organization bought the notebook for their productivity workforce, they might list all properties that you believed to be requirements of the notebook and then allocate a "tick" or "cross" for each property for each notebook candidate, according to whether it obsessed that property. The organization would then count the number of ticks that each candidate had received. Those with the highest counts would offer a short list of candidates on which to carry out a more detailed evaluation to decide their relative value for cash or some other conditions for finally deciding which one to choose.

Implementing the features in the evaluation process will gain some advantages of feature analysis. Simple pre-requisites, can be executed to any required level of detail, and can be applied to any type of method; process and tools are key advantages of feature analysis. In addition, feature analysis also has major advantage that it is not restricted to technical evaluations only, but also managerial and business acquisition evaluations.

Feature analysis also has several disadvantages, which are subjectivity, inconsistencies, collating score, and generating too many features. Feature analysis is been based on judging methods against some "evaluation criteria" which are identified subjectively and context dependent. There is also a problem of inconsistency in scoring between different assessors. If different assessors evaluate different tools, they may have different degrees of familiarity with an understanding of the method/tool. The various score has to be collated and compared to decide the relative order of merit of the methods or tools. Another problem of feature analysis is that hundreds of features may be identified. This makes performing an assessment of a specific method/tool very time consuming and makes it difficult to analyses all the scores. Knowing the disadvantages and keep it in balance will provide sufficient understanding of the assessed framework.

### 2.4 Evaluation Result

The objective of the evaluation is usually to provide input into a decision about whether to adopt a framework for use by organization. Feature analysis experiment will compare the proposed formalized framework with legacy distributed software development framework. The framework evaluation will normally be intended to carry a specific purpose in qualitative and quantitative degrees. The result of the evaluation needs to provide information in the following areas:

1. Suitability of purposes. It discusses the appropriateness of the framework
2. Economic issue. It discusses the investment and the return of investment when adopting the framework.
3. Drawbacks. It discusses any aspects that make the framework less attractive.

4. Advantages. It discusses any aspects that make the framework more attractive.

In addition, the evaluation method should also help clarify the important features of the method or tool in the context of the organization environment. This can be done easily by identify the framework with the others.

## 3. Evaluation Methodology

Although DESMET provides a comprehensive guide to evaluate the method and tools, it is only providing a generic approach to do an evaluation. The research has specific need to evaluate the multi-site agile development. Therefore, it needs a modification in the usability of DESMET.

The research chooses case studies approach that already described generic in DESMET. The idea of case study approach is a means of evaluating a framework as part of the normal software development activities undertaken by an organization through real projects. The research discusses the implementation of case studies into three main phases, which are the evaluation preparation, evaluation implementation and evaluation result.

### 3.1 Evaluation Preparation

Evaluation preparation deals in several steps of feature analysis case studies. They are identifying the case study context, define and validate the hypothesis, select the host projects, identify the method of comparison, minimize the effect of confounding factors, and plan the case study.

**Identify the case study context.** For evaluations, the research needs both projects to evaluate “which is better” as well as how and why. The how and why aspect can provide valuable awareness into why technology results in better results. The framework which is wanted to evaluate is called treatments. In order to determine whether a treatment is beneficial, it needs to compare it with an alternative treatment or with the currently used framework. The framework that presently uses is mentioned as the control treatment. The control treatment provides a baseline of information to enable comparison to be made. The case study would not only be interested in deciding whether the proposed framework is better than the current framework. It’s also needed to define fully to provide the guidance so that the experiment could be replicated.

The research selects two candidate frameworks. The first framework is original GSD framework that used in GSD handbook. Original GSD framework will be used as a controlled project. Controlled project is a real world project which is used GSD process and Unified Process method as a framework to build distributed software development. Both combinations are gently called as legacy GSD. The second framework is GXP framework. GXP framework will be used as an experiment project. Experiment project is a simulation project that has the same people skill set as the controlled project, same technical complexity of the project, and same situation in the environment. The similarity of the project will lead an acceptance for the framework to be evaluated throughout several assumptions of circumstances.

Candidate framework also identified the experiment contexts. The experiment context sets the objective and limitations within which is experimented must operate. The evaluation identifies the experiment sponsor, the available resources, time scales, and the importance of the experiments for the organization. Table 2 provides fact tables for case study context.

**Define and validate the** hypothesis. This step is restating an evaluation goal in a testable manner. The hypothesis in this research is based on the identification of a particular problem in the distributed software development process which is proposed framework is intended to solve.

Defining the hypothesis is started by specifying the goal of a case study, the framework that wishes to evaluate, the aspect of the framework that is interested in investigating, and an effect in the variable that related directly with the evaluation. The goal of a case study is to identify benefits and weakness of the proposed framework compared with the current framework. GXP framework is a framework that wishes to evaluate, and legacy GSD is a framework that works as the evaluation baseline.

**Table 2.** Case studies contexts Example

| The context of case study planning | Controlled Project | Experiment Project |
|---|---|---|
| **Case study sponsor** | Manufacturing organization | IT Research organization |
| **Organization resources for the case studies** | 3 person client, 7 development members | 2 person client, 5 development members |
| **Case studies timescales** | 6 months | 6 months |
| **Case study importance for the organization** | Internal web development for main business process | Project management web application for internal organization |

The feature analysis case study uses two main variables, which are responses variables and states variables. Response variables are variables, which is expected change be different pursuant to applying the treatment (i.e. faster than before, efficient than previous, etc.). State variables are factors that characterize the experiments and can influence your evaluation results (i.e. application area, system type, organization model, etc.).

Response variables related directly with case study results. It typically developer productivity and product quality, which are expected to change or to be different pursuant to applying the treatment. The research identifies several response variables based on agile artifacts as follows.

1. Burndown trend pattern. It defines how the project velocity based time and remaining of works. It reflects the productivity of the team.
2. Defect rate. It stores numbers of errors on iteration. It reflects the software quality regarding by the defects.
3. Check-in operation. It stores the historical integrations by the developer to the system. It reflects the continuous improvement of the team.
4. Investment costs. It summarized altogether the development process investments. It reflects the efficiency of investment for the framework regarding the tools, communication, and method investments.
5. Communication pattern. It defines the communication trend between member and client. It counts the

numbers of communications through messages (Email, Instant Messaging) and others online communication.

State variables define the characteristics project that typical project has in the organization. The state variables are described as follows.

1. Application area of the projects. It defines the kind of the project that developed in case study.
2. Framework that used. It defines the software engineering framework that used in case study.
3. Business type. It defines the business type that runs by the organization.
4. Scale of projects. It defines the product size range.
5. Complexity level of problems. It defines the complexity of the software regarding of the business process domain.
6. Quality and experience of the staff. It defines the technical skill and experience skill of the team members and clients.
7. Physical and workforce environment. It describes the daily works and physical environments of the projects.

**Select the host project.** The host project which is chosen represents the type of projects that usually performs by the organization. It is to ensure that the results from the evaluation are applicable to more than just the trial project. Host project is defined through organization profiles. Since the organization profile that the research chosen is typically an organization that uses an online solution through the web, the host project will use a web project as a host project. This organization profile is connected with state variables to define the detail characteristics of the project. Table 3 provides the state variables for the controlled and experiment projects.

**Table 3.** State variables Example

| State variables | Controlled Project | Experiment Project |
|---|---|---|
| **Application area** | Line of Business App. | Line of Business App. |
| **Framework that used** | GSD process + UP method | GXP framework |
| **Business Type** | Manufacturing | Software |
| **Scale of projects** | 10000-18000 LOC | 10000-18000 LOC |
| **Complexity level of problems** | Low | Low |
| **Quality and experience of the staff** | 4 years experiences average | 2 years experiences average |
| **Physical and workforce environment** | High pressure with overtime | Casual without overtime |

Application area is defined as a kind of software that being developed, which are business application. The kind of software is adopted and extended from Productivity rates for common project types (McConnell, 2006). Scale of projects and complexity levels of problems also derived from pre-projects estimation based on use case metrics and user stories that aligned with productivity rates for common project types.

**Minimize the effect of confounding.** The casual problem in the feature analysis case study is a subjectivity of the result. The degree of subjectivity is been based on judging method against some “evaluation criteria” which are identified subjectively based on context dependent in case studies. There is also a problem of inconsistency in scoring between evaluators, the different evaluator will give different score interpretations in a different way. The various score has to be collated and compared to decide the relative order of distinction frameworks. Furthermore, certain features may attract higher average scores than others because an assessor possibly will appreciate them better and be more able to be familiar with the framework. Another problem of feature analysis is that hundreds of features may be acknowledged, and it will become time consuming to evaluate.

Those confidence problems are constraints the evaluation to decide the level of confidences such as.

1. The case study is held by one organization with several assessors, although it will context dependent and subjective. It will give the same baseline of the confidence result.
2. The case study will use the software development life cycle (SDLC) process as a base path to achieve the result. Using SDLC will give real work experience regarding the framework.
3. The case study will limit the response variable that already stated in above.

In order to give an objective result, the research makes several assumptions that described in the next section. The research makes several assumptions regarding the risk that will be handled in the evaluation process. These assumption purposes are to limit the divergences of the researches object. There are three assumptions that assumed in the research which are.

1. People assumptions. The case study is done by the equal people who have the same level of experience at least two years as the developer, technical skills in the web framework they build, and proficiency skill in terms of teamwork.
2. Product assumptions. The case study is done by the equal project complexity, including the same framework that used to build the product. It means that the projects have a same project types (web application), same project purposes (business application), and same amount of user stories.
3. Process assumptions. The case study is done by the equal base process model. The experiment uses a global software development situation as a background process that adopted in different method and tools.

Those assumptions are the boundary and limitation for the case study is done in terms of objectivity.

### 3.2 Evaluation Plan

Evaluation plan must identify all issues to be addressed so that the evaluation runs smoothly. This including the evaluator, the data gathering procedures, and the measures needed for the analysis.

**Evaluators** are the entire person or team which responsible in running exercise for an evaluation. The team is responsible for.

1. Preparing the evaluation plan.
2. Identifying the candidate framework
3. Identifying each distinct group of user population
4. Eliciting the requirements of each user
5. Identifying the features to be assessed
6. Organizing the assessment whereby each of the frameworks.
7. Collation and analyze scores.
8. Preparation of the evaluation report.

The evaluation involves the clear separation between evaluation staff and experimental subjects. Evaluation staff should work as independent assessor, which evaluates the experiment outside the subject's environment. Therefore, evaluation staff should not technically join the project in order to concentrate with the evaluation process. On the other side, experimental subjects should not know that their works are evaluated as a case study. This approach is to make sure the subject to work and do the evaluation naturally with fewer human risk side effects like novelty or expectation effect as stated in section 2.2.

**Data's gathering procedures**. Data's gathering procedures are started by composing the evaluation form. Evaluation form should represent explicitly or implicitly what states in the evaluation required properties. Evaluation form should also describe the experiment plan and baseline rating. After the evaluation form is done, the evaluator should works as follows.

1. Identifying the subject by looking at the potential user who holds specific roles in the project.
2. Identifying the tasks to be performed by the experimental subjects using the framework.
3. Organizing any exercise or support for experimental subjects.
4. Running the required measurements according to the case study plan
5. Evaluating the forms and preparing the evaluation reports.

Case study minimizes the effect of individual assessor differences since it will use the real-life experience of project development. However, the problem is whether the result of the experiment will scale up in the different project. This problem will be covered by using a baseline rating in the semantics of measurement section.

**The measures needed by the analysis**. The measurement is done based on the case study execution. The case study will gain both quantitative and qualitative information. The quantitative information provides direct calculation and comparison between proposed and existing framework in terms of numbers. The qualitative information provides additional information how the subject of the experiment uses the properties that discussed in the evaluation. The measurement is done through these steps.

1. Evaluator joins in both projects as a quality control or assessor team member.
2. Evaluator examines the execution of the projects by following SDLC and project execution.
3. Evaluator submits the reports by filling the assessment's sheet for each project.

### 3.3 Evaluation Execution

This section provides detail information the execution of the evaluation based on the measurement steps which are described in section 3.2. Those steps are joining the team, following the SDLC execution, and submitting the reports. The SDLC steps that proposed in the research work as figure 2.

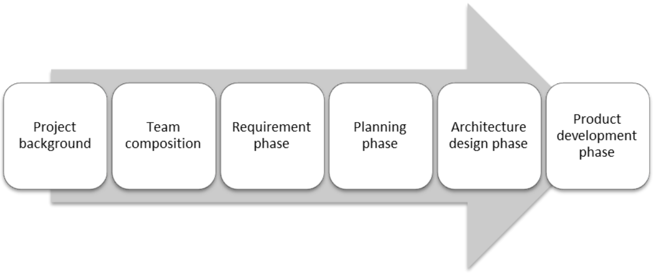


**Figure 2.** Evaluation execution block diagram

Project background discusses the attributes of the evaluation case studies. The research chooses project type, environment, project length, team members, framework adoption, web framework engine, platform that used for the project, and amount of software features. Those attributes should be at least same in several things in order to create a feasible comparison between case studies.

Team composition discusses the team model that implemented in the case studies. Team structure leads the mechanism of collaboration and communication between sites. In this step, the research should identify the team structure, team roles, and team communication workflow.

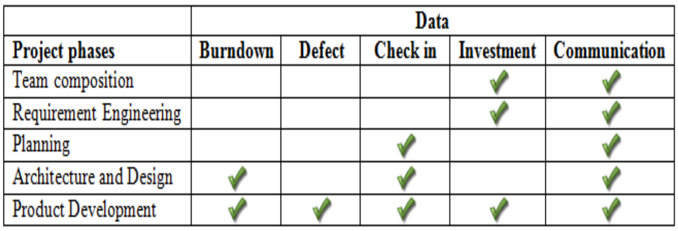
Requirement phase discusses to gather the market intent, the vision of the product, and the business purposes that are covered by the product. This phase identifies team action to fulfill this phase. It includes the artifact and the process that done by the team.

The target of planning phase is creating bundles of the features that will be developed by the team. These bundles can then be prioritized through an agreement between the client and the team. The research identifies what the team does in the planning phase.

This phase focuses in creating a design of the software, including the user interface design and software architecture design. The research identifies what the team does in the design and architecture phase.

Product development phase is the roughest phase in the software development cycle. This running application and the source code are the main output of this phase. Table 4 describes the portion of response variable that gathered in evaluation SDLC phase.

**Table 4.** Data acquired in case studies phases



| | Data | | | | |
|---|---|---|---|---|---|
| Project phases | Burndown | Defect | Check in | Investment | Communication |
| Team composition | | | | ✓ | ✓ |
| Requirement Engineering | | | | ✓ | ✓ |
| Planning | | | ✓ | | ✓ |
| Architecture and Design | ✓ | | ✓ | | ✓ |
| Product Development | ✓ | ✓ | ✓ | ✓ | ✓ |

## 4. Evaluation result sample

There are five response variables as a result in the researches, which are communication pattern, investment cost, check-in operation, defect rate, and burndown rate.

Communication patterns discuss the kind of patterns that happen in the overall project communication. The communication patterns discuss multimodal communication that happens in the both projects such as.

1. Email communication. It measures the amount of the email that happens between the client and development team. For the evaluation purposes, only unique email that counted in the evaluation. For an example if the manager sends same email to all the team and there are seven separate copies that the email is still counted as one email only.
2. Phone. It measures how many hours that spends for personal phone call (between peer) or conference phone calls.
3. Video conferences. It measures how many effective hours that spend for video conferences certainly it is excluding setup and configuration testing.
4. Instant messaging. It measures how many effective hours that spend for instant messaging, including private chat or conference chat sessions.
5. On site meeting. It measures how many hours that spends for onsite meeting, including the travel hours and others.

Table 5 provides the samples of communication measurement for case studies. The organization can calculate the detail of communication and see a balance between direct communication and indirect communication. The better communication provides the better result in project collaboration and cooperation.

**Table 5.** Communication measurements for case studies

| Communication Type | Unit | Controlled | Experiment |
|---|---|---|---|
| **Email** | item | 127 | 227 |
| **Phone** | hour | 180 | 40 |
| **Video Conference** | hour | 13 | 0 |
| **Instant Messaging** | hour | 2 | 113 |
| **On site meeting** | hour | 192 | 27 |

Investment cost discusses the framework investment that needed by the project. Investment cost includes several fix costs and variable cost such as.

1. Integrated development environment tools (IDE tool). The IDE tool is a combination of compiler, debugger, and application designer. Visual Studio, Net beans, and Eclipse are the sample of the IDE.
2. Developer component. Developer component is the third party component that purchased for developer productivity.
3. The CASE tool. The CASE tool is the software engineering software that helps the team to develop diagram, chart, and others engineering representation. Visual paradigm, enterprise architects, or Visio are the sample of the CASE tool.
4. The project management tool. The project management tool is software to manage and track project progress. Microsoft project is the sample of project management software.
5. Travel cost. Travel cost calculates the total of expands that used for the travelling budget.
6. Communication cost. Communication cost estimates the communication expands like the internet, phone, or short messages investment.
7. Hardware cost. Hardware cost estimates the cost of development hardware cost. Development hardware is a set of hardware that needed for development only. Development server, notebook to build the codes is the samples of the hardware cost.

These cost is obtained through an informal audit. The evaluator estimated the cost by seeing several proof of receipt, existing item, and investigated the prices of the development assets.

As mentioned, Burndown patterns are a set of pattern that displays the progress of the project in terms of remaining works versus the iteration of time. The interesting part is that the analysis of the burndown chart can expose various indicators on how the team is undertaking the plan and what can they do to improve further. The burndown pattern answer this following question

1. How good the team planning?
2. How well is this team executing against the planned stories in iteration?
3. Is this team self-organized and are they working in unison as a "team"?

Both two projects are side by side compared within their burndown chart. The data burndown chart is gathered from the manual calculation based on the project planning and requirements documents. The research counts the requested features, the works constraint, and others technical works that needed to be done. Figure 3 shows the burndown sample between two projects. The slight slope provides better project productivity.

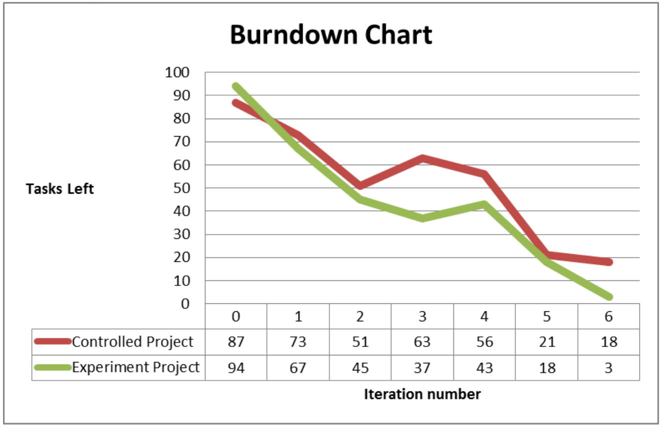


**Figure 3.** Burndown Chart

Defect rate discusses how many defects that exist in the both projects. Defect can be categorized as follows.

1. Bug or program error that happens when the software is running.
2. The misleading feature or wrong features that built by the team.

3. The integration or algorithm error. The errors that make the software work, but it gives a different result than expected.

Based on these categories, the evaluation counts the error from the project. The source of the information comes from feedback log, error notification through email, and compiler error counter. Figure 4 shows the sample of defect rate. The less defect rate shows the better project management to handle the defect.

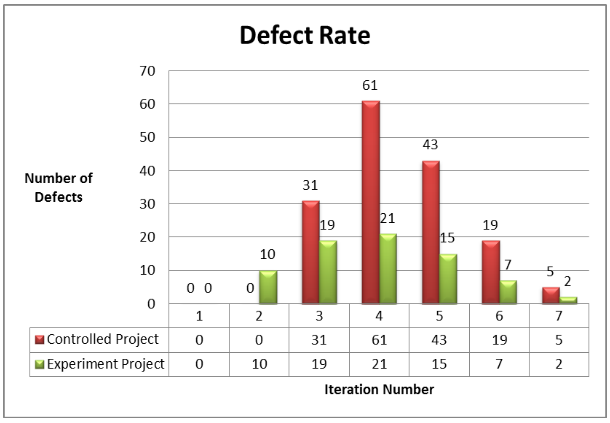


**Figure 4.** Defect rate chart

Check-in operation discusses how many changes that happen in the released codes and how the team makes a better improvement of the codes. The revision of the system, the fresh new build, and the new version of the software is clear identification of the check-in operation.

Check-in operation is measured through several approaches. The evaluator uses a milestone counter, nightly build, and revision logs that note by the team which are.

1. The successful build of the system and uploaded into the development system.
2. The minor revision of the project such as adds news features; fix several bugs and new facelift of the user interface.
3. The milestone of the project, since every milestone of the project shows several improvement features of the project.

Figure 5 shows the check-in operation for the both projects. The higher check in rates of operation shows the continued improvement of the project.

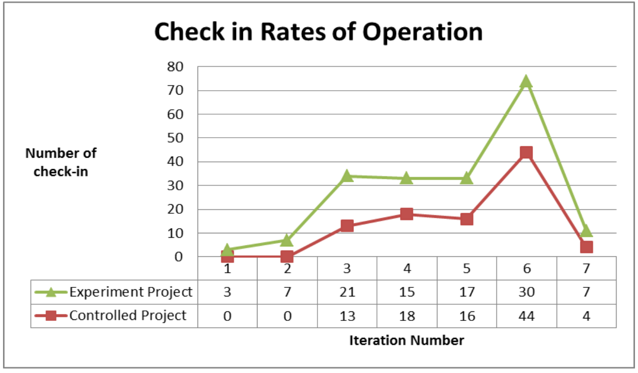


Figure 5. Check in rates of operation

The evaluation results provide a summary for the organization to choose between two or more frameworks. In a simple manner, organization can compare the result between them and choose which one that provides better productivity and efficiency.

The challenge that usually happens to be evaluating the framework without sacrifice the organization productivity. Some of the organization has no interest to create the experiment project, and others have no interest to adopt the new framework to their real project because its risks. In this kind of situation, the research advises the team to evaluate the framework based on the real and low risk project. Evaluating the real and low risk project can capture the picture of the team when using the framework.

## 5. Conclusion and Future Work

In software engineering, evaluating a software methodology become the important thing for the organization. Organization that chooses the correct methodology will gain the benefit as long as its productivity.

This paper limits an evaluation approach to evaluate a software engineering framework. This paper chooses DESMET as a baseline of the research framework and adds several steps that proposed in this research such as follows.

1. Adopting DESMET as an evaluation preparation step. In this step state variables and case study context is chosen.
2. Following SDLC in the evaluation execution step.
3. Identifying the results based on the response variables that already defined in the preparation step. Response variables are a set of artifacts that implicitly described the performance of the project.

As a further work, the framework evaluation should be detailed with the risk that identified when adopting this evaluation and improving the efficiency the execution evaluation without leaving a risk for the case study.

## References

[1] Ferdiana, R. Nugroho, E.L, Santoso, I.P, and Ashari, A. 2010. Global eXtreme Programming, a Software Engineering Framework for Distributed Agile Software Development. IJCSET 1, 3 (October 2010), ISSN 2044-6004.

[2] Kitchenham, B. A. 1996a. Evaluating software engineering methods and tool—part 1: The evaluation context and evaluation methods. SIGSOFT Softw. Eng. Notes 21, 1 (Jan. 1996), 11-14.

[3] Kitchenham, B. A. 1996b. Evaluating software engineering methods and tool—part 2: Selecting an appropriate evaluation method—technical criteria. SIGSOFT Softw. Eng. Notes 21, 2 (Mar. 1996), 11-15.

[4] Kitchenham, B. A. 1996c. Evaluating software engineering methods and tool—part 3: selecting an appropriate evaluation method—practical issues. SIGSOFT Softw. Eng. Notes 21, 4 (Jul. 1996), 9-12.

[5] McConnell, S. 2006 Software Estimation: Demystifying the Black Art. Microsoft Press.

[6] Sadler, C. and Kitchenham, B. 1996. Evaluating software engineering methods and tool—part 4: the influence of human factors. SIGSOFT Softw. Eng. Notes 21, 5 (Sep. 1996), 11-13.

[7] Sangwan, R., Bass, M., Mullick, N., Paulish, D. J., and Kazmeier, J. 2007. Global Software Development Handbook (Auerbach Series on Applied Software Engineering Series). Auerbach Publications.

## Author Biographies

**Ridi Ferdiana.** Mr. Ridi Ferdiana was born in 1983. He is a doctoral student at Gadjah Mada University, Yogyakarta since 2008. He earned his master degree from the same university in 2006. In his professional area, he holds several professional certifications such as MCP, MCTS, MCPD, MCITP, MVP and MCT. In his daily research activities he really enjoys to learn about software engineering, business platform collaboration, and programming optimization.

**Lukito Edi Nugroho.** Born in 1966, Dr. Lukito Edi Nugroho is an Associate Professor in the Department of Electrical Engineering and Information Technology, Gadjah Mada University. He obtained his M.Sc and PhD degrees from James Cook University in 1995 and Monash University in 2002, respectively. His areas of interest include software engineering, distributed and mobile computing, and application of ICT in education.

**Paulus Insap Santosa.** Insap was born in Klaten, 8 January 1961. He obtained his undergraduate degree from Universitas Gadjah Mada in 1984, master degree from University of Colorado at Boulder in 1991, and doctorate degree from National University of Singapore in 2006. His research interest including Human Computer Interaction and Technology in Education.

**Ahmad Ashari** Place and date of birth: Surabaya, May 2$^{nd}$ 1963. Get Bachelor’s degree 1988 in Electronics and Instrumentation, Physics department Gadjah Mada University, Yogyakarta. Master degree 1992 in Computer Science, University of Indonesia, Jakarta Doctor Degrees 2001 in Informatics, Vienna University of Technology. Major Field of study is distributed system, Internet, Web Services, and Semantic Web.

# Unsupervised Segmentation of Multispectral Textured Images using GA-GMRF model

Mridula J[†] and Dipti Patra[*]
IPCV Lab, Dept. of Electrical Engineering
National Institute of Technology, Rourkela-769008, Orissa, India
[†]e-mail: mridulamahesh3@gmail.com
[*]e-mail: dpatra@nitrkl.ac.in

***Abstract*****—** This paper proposes an hybrid genetic algorithm (GA) and Gaussian Markov random field model (GMRF) based method for segmentation of multi-spectral textured images. It also evaluates the popular unsupervised image segmentation approaches, Genetic algorithm based clustering and simple Gaussian Markov random field model with the hybrid GA-GMRF method for high spatial resolution textured imagery. Each method is described and the compatibility of each method with the textured image is examined. It is observed that GA based clustering is more suitable for medium resolution imagery and for images without textures. GMRF model which gives desirable results for textured images requires several iteration steps to approximate near optimal solutions. The hybrid GA-GMRF method, in which the powerful global exploration of GA is used to initialize the ICM algorithm, has found more promising and gives improved results in terms of both accuracy and time complexity than the two other methods for multi-spectral textured images.

***Keywords*** **-** Unsupervised image segmentation, Genetic algorithm, Markov random fields, texture, high spatial resolution multi-spectral images

## 1. Introduction

Image classification is a task that classifies pixels of an image using different labels so that the image is partitioned into non overlapping labeled regions. Land cover classification is one of the fundamental operations in the arena of multi-spectral image analysis. This is for the reason that classification results are the basis for many environmental and socio economic applications at global, regional and even local level. However deriving land cover information from multi-spectral imagery is a difficult task because of the complexity of the landscape and the spatial as well as spectral resolution of the imagery being used [1]-[2]. Coarse spatial resolution data are preferable at continental or global scale. At regional level, medium spatial resolution imagery is often used and for classification at local level, high spatial resolution data are helpful. Multispectral data at moderate and coarse spatial resolution can be differentiated based on the spectral reflectance patterns. At high spatial resolution the role of texture assumes more significance. Although a large number of segmentation techniques are available in the literature for classification, there is no standard criterion on, which method is more suitable or more effective. The segmentation problem is addressed as supervised and unsupervised segmentation. When the number of classes, image labels and model parameters are unknown, it is completely unsupervised. If the number of classes is known then it is partially unsupervised.

When the image is of low or medium resolution and the number of classes is known *a priori*, among the several unsupervised classification techniques, K- means algorithm is one of the most widely used ones. But it has a limitation that it gets stuck at sub optimal solutions depending on the choice of initial cluster centers [3], [4]. Genetic algorithm (GA) is a stochastic search technique based on the mechanics of natural selection and genetics originated from the imitations of natural evolutions on the earth. This has been successfully employed in the classification of artificial as well as real image data sets in [3]. They work with the strings that encode candidate solutions called chromosomes and collection of such chromosomes is known as population. An objective and fitness function that represents the degree of goodness of the string is associated with each string. This fitness function is used to guide the stochastic selection of the chromosomes which are then utilized to generate new candidate solutions through crossover and mutation. Crossover allows solutions (chromosomes) to exchange information and produce new chromosomes. Mutation is used to randomly change the value of genes and increase the diversity in the population.

With high spatial resolution imagery, the objects make up the thematic classes as the spectral resolution of the sensor nears the object size on the ground and thus brings in the texture effects. This limits the potential of spectral information since same spectral reflectance value can correspond to different objects. Hence the contextual classifiers that utilize both spectral and spatial information are particularly worthful for fine resolution [5]-[6]. Zoltan Kato and Ting-Chuen Pong [7] proposed an MRF model based image segmentation method which aims at combining colour texture features which relies on Bayesian estimation via combinatorial optimization. The algorithm is highly parallel. Brandt C.K Tso and Paul M. Mather [8] presented an MRF model using Genetic algorithms for multi source remote sensing imagery. As Markov random field (MRF) model utilizes both spectral and spatial information to model the local structure of an image, it is undoubtedly, a potent mathematical tool for contextual modeling of spatial data [2], [9], [10]. For modeling textures, spatial interaction models particularly conditional Markov models are more useful. GMRF are a special case of Markov random fields used to model textured images [9], [11], [12]. GMRF model representing colour texture that takes into account both within band and between bands interaction has been proposed and successfully implemented by Panjwani and Healey [11]. In MRF based segmentation, the most popular criterion for optimality has been maximizing a posteriori probability (MAP) distribution criterion. Simulated annealing (SA) and iterated conditional modes (ICM) algorithm are two unremarkably used methods for pixel labeling among the existing MAP criterion algorithms. SA can converge to global optimum, but suffers from intensive computation. On the other hand, the results obtained from ICM heavily depend on initialization and hence there is a probability of trapping into local maxima. Hence it suffers from inaccurate estimations. Tseng and Lai [6] have successfully employed hybrid GA-MRF based segmentation, where GA is used to provide better initialization for the ICM algorithm.

In this paper we examine the hybrid GA-GMRF model based segmentation method that takes into account both spatial interaction within each of the bands and interaction between different bands and present a comparison of this method with methods including Genetic algorithm based clustering [3] and GMRF model based segmentation using ICM algorithm [6] for high spatial resolution imagery. The advantages and disadvantages of each method are evaluated by simulations and classification results. Section II reports Genetic algorithm based segmentation and section III, MRF model based segmentation with inter and intra band spatial interaction. Section III describes the hybrid segmentation technique based on GA and MRF with inter and intra band spatial interaction. Results are presented in section IV and section V makes conclusion.

## 2. Genetic Algorithm Based Clustering

The operations performed in GA based clustering as given by (3) which is considered for comparison is reproduced here for the ease of reference. The number of clusters is known *a priori.*

### 2.1 Chromosome representation

A chromosome may be encoded with binary, integer or real numbers. We have taken real numbers as the cluster centroid value will be a real number.

### 2.2 Population initialization

Each string in the population encodes the centers of k clusters. These centers are initialized by random selection from the data set. As an example one chromosome of the initial population (parent generation) is given below:

(51.0, 220.0) (67.0, 54.0) (78.0, 134.0) (98.0, 76.0)

In the example it is assumed that the image has 4 clusters and is with 2 bands. Number of clusters and the number of bands depends on the multi-spectral image being considered.

### 2.3 Fitness Computation

A function of the ratio of the sum of within cluster separation to between cluster separations which is known as Davies Bouldin's index is used to compute the fitness of a chromosome [3]. The equations are as follows,

The within cluster scatter of cluster k is given by,

$$S_k = \left( \frac{1}{C_k} \sum_{x \in N_k} \|x - z_k\|^2 \right)^{1/2} \qquad (1)$$

Where $S_k$ is the average Euclidean distance of vectors in class k, $z_k$ is the centroid of the class k and is computed as $z_k = \frac{1}{n_k} \sum_{x \in c_i} x$ and $n_i$ is the number of points in cluster $C_{i.}$

The distance between cluster $C_i$ and $C_j$ is given by,

$$d_{ij,t} = \left\{ \sum_{s=1}^{p} |z_{is} - z_{js}|^t \right\}^{1/t} \qquad (2)$$

Where $d_{ij,t}$ is called the Minkowski distance of order t between the centroids. Here we have considered the distance of order 2.

$$\text{Then } R_{i,t} = \max_{j, j \neq i} \left\{ \frac{s_i + s_j}{d_{ij,t}} \right\} \qquad (3)$$

is computed to define DB index as,

$$DB = \frac{1}{k} \sum_{i=1}^{k} R_{i,t} \qquad (4)$$

The clustering with the minimum DB index gives the properly clustered image.

### 2.4 Selection

A proportion of existing population is selected to breed a new generation during each successive generation. Roulette wheel selection is employed for selection of individuals.

### 2.5 Crossover

Single point cross over procedure is carried out by stochastic means with probability µc.
Example:

P1: (51.0, 220.0) (67.0, 54.0) | (78.0, 134.0) (98.0, 76.0)
P2: (65.0, 212.0) (86.0, 25.0) | (133.0, 49.0) (19.0, 26.0)

P1 and P2 represent parent1 and parent2. The line shown is the point where crossover takes place. The genes after that position are exchanged to produce children.

Child1: (51.0, 220.0) (67.0, 54.0) | (133.0, 49.0) (19.0, 26.0)
Child2: (65.0, 212.0) (86.0, 25.0) | (78.0, 134.0) (98.0, 76.0)

### 2.6 Mutation

The valid genes in the chromosomes are mutated with a probability μ$_c$

### 2.7 Termination

The execution is terminated with maximum number of iterations. An elite chromosome preserved in a location outside the population with maximum fitness contains the centers of the final cluster

## 3. Gaussian Markov Random Field Model

MRF models constitute a powerful tool in image analysis process due to their ability to integrate contextual information associated with the image data [10]. The MRF approach shows the global model of the contextual information by using only local relations among neighbouring pixels. A large category of global contextual models are equivalent to local MRFs. This can be proved by Hammersley-Clifford theorem and thus model complexity is reduced to a great extent. GMRF are a special case of Markov random fields used to model textured images [9].

Let X(i,j) = [ $x_1$(i,j) $x_2$(i,j) … $x_p$(i,j) ] represent multi-spectral Gaussian random vector of a pixel at location (i,j) in a textured region R. ‘p’ represents the number of bands in the multi-spectral image. Let μ$_1$, μ$_2$,.., μ$_p$ represent mean color intensities in each band. The conditional probability distribution function for MRF, which is assumed to be Gaussian is given by,

$$P\left(X(i,j)\,|\,R|\right)=\frac{1}{\left((2\pi)^{P}\,|\sum\,|\right)^{1/2}}\cdot$$
$$\exp\left\{\frac{-1}{2}\left[e_1(i,j)e_2(i,j)\ldots e_p(i,j)\right]\ \sum{}^{-1}\left[e_1(i,j)e_2(i,j)\ldots e_p(i,j)\right]\right\} \tag{5}$$

Where [e$_1$(i,j) e$_2$(i,j) …. e$_p$ (i,j) ] is a zero mean Gaussian noise vector. The spatial interactions of the multi-spectral pixels is given by,

$$e_y(i,j)=\left(x_y(i,j)-\mu_y\right)-\sum_{z=1}^{p}\sum_{(m,n)\in N}\alpha_{yz}(m,n)\left(x(i+m,j+n)\right) \tag{6}$$

Where μ$_y$ is the mean of the variable x$_y$(i,j), α$_{yz}$ are model parameters and y takes the values from 1 to p. Subscript N represents different neighbourhood systems. ∑ is the correlation matrix and is given by,

$$\sum=\begin{bmatrix} v_{11} & v_{12} & . & . & v_{1p}\\ v_{21} & v_{22} & . & . & v_{2p}\\ . & & & & .\\ . & & & & .\\ v_{p1} & v_{p2} & . & . & v_{pp}\end{bmatrix} \tag{7}$$

Where v$_{kl}$ is the expected value of e$_k$ e$_l$ and is represented by,

$$v_{kl}=E\left[e_k e_l\right]=\frac{1}{M_R}\sum_{(i,j)\in R}e_k(i,j)e_j(i,j) \tag{8}$$

### 3.1 Parameter estimation

The spectral vector of a pixel in a multi-spectral image is influenced by its neighbouring pixel vectors. As given by Eq. (5) pixel in each band depends on neighbours of the same band as well as the neighbouring pixels of different bands. Hence the parameters are estimated using pseudo likelihood method adopted in [6], as the product of conditional probability densities of all pixels in region R is not a true likelihood. The product is called the pseudo likelihood function and is given by,

$$\prod_{(i,j)\in R}\frac{1}{\left((2\pi)^{p}\left|\Sigma_R\right|\right)^{1/2}}\exp\left\{\frac{-1}{2}\left[e_1(i,j)e_2(i,j)\ldots e_p(i,j)\right]\Sigma_R^{-1}\left[e_1(i,j)e_2(i,j)\ldots e_p(i,j)\right]^{t}\right\} \tag{9}$$

The parameters can be estimated by maximizing the pseudo likelihood function as given in [6]. As an example let,

$q_1(i,j)=x_1(i,j)-\mu_1$ , $q_2(i,j)=x_2(i,j)-\mu_2$ ,...,
$q_p(i,j)=x_p(i,j)-\mu_p$ *and* $q_{1,mn}(i,j)=x_1(i+m,j+n)-\mu_1$ ,
$q_{2,mn}(i,j)=x_2(i+m,j+n)-\mu_2$ ,... , $q_{p,mn}(i,j)=x_p(i+m,j+n)-\mu_p$

Then α parameters for band 1 can be solved using the equation below,

$$\sum_{(i,j)\in R}\begin{bmatrix} q_{1,10}^2 & q_{1,10}q_{1,01} & \cdots & q_{1,10}q_{p,10} & \cdots & \cdots\\ q_{1,01}q_{1,10} & q_{1,01}^2 & \cdots & q_{1,10}q_{p,10} & \cdots & \cdots\\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots\\ q_{1,mn}q_{1,10} & q_{1,mn}q_{1,01} & \cdots & q_{1,mn}q_{p,10} & \cdots & \cdots\\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots\\ q_{2,10}q_{1,10} & q_{2,10}q_{1,01} & \cdots & q_{2,10}q_{p,10} & \cdots & \cdots\\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots\\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots\\ q_{p,10}q_{1,10} & q_{p,10}q_{1,01} & \cdots & q_{p,10}^2 & \cdots & \cdots\\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots\\ q_{p,mn}q_{1,10} & q_{p,mn}q_{1,01} & \cdots & \cdots & \cdots & q_{p,mn}^2\end{bmatrix}\begin{bmatrix}\alpha_{11}(1,0)\\ \alpha_{11}(0,1)\\ \cdots\\ \alpha_{11}(m,n)\\ \cdots\\ \alpha_{12}(1,0)\\ \cdots\\ \cdots\\ \alpha_{1p}(1,0)\\ \cdots\\ \alpha_{1p}(m,n)\end{bmatrix}=\sum_{(i,j)\in R}\begin{bmatrix}q_1q_{1,10}\\ q_1q_{1,01}\\ \cdots\\ q_1q_{1,mn}\\ \cdots\\ q_1q_{2,10}\\ \cdots\\ \cdots\\ q_1q_{p,10}\\ \cdots\\ q_1q_{p,mn}\end{bmatrix} \tag{10}$$

## 4. GA–GMRF Hybrid Method

To overcome the drawback of ICM algorithm, which depends on the initialization to produce accurate results, genetic algorithm is employed to give better initialization for ICM algorithm. Using this hybrid method, the fast convergence of ICM and global exploration of GA are achieved simultaneously.

The overall procedure is as follows:

1. A coarse segmentation is performed to get the initial regions using K-means algorithm to reduce the amount of time required in MRF based iterative process and to obtain the mean values.

2. The ICM algorithm initialization using genetic algorithm is performed as follows,

   The spectral vector X(i,j) = [$x_1$(i,j) $x_2$(i,j) …$x_p$(i,j)] representing the spectral components of a pixel is encoded as an individual. Population of individuals is created, each individual is evaluated and the better individuals are selected to raise the next generation. The procedure is continued till the maximum number of generations is accomplished. For population initialization the result of the K-means algorithm is taken as an individual and the remaining individuals are generated randomly. The function defined in Eq. (5) is used as the fitness function to evaluate each individual. Then the other operations of GA ( selection, crossover, mutation) are performed and the individual with the highest fitness is used to give the initial label for ICM algorithm. 100

3. Then the ICM algorithm is performed to produce the final segmented image. The number of generations performed: 100, Population size: 100 and the number of iterations taken by ICM algorithm: 10.

## 5. Simulation Results

The segmentation results using the GA clustering, ICM and hybrid GA-ICM models for high resolution textured image are presented. Fig. (1) shows a synthetic textured image containing four textures. The size of the image is 200 × 200 pixels. Fig. (2), Fig (3) and Fig. (4) show the segmented images using GA, GMRF-ICM and GA-GMRF-ICM respectively. Pseudo colors are used to indicate the segmented results. We found that segmented results for colored texture image using GA clustering are very noisy because of the reason that it considers only the spectral information without taking into account spatial information. Although MRF model based approach gives better results, because of the initialization problem it suffers from misclassification problem. Hybrid GA-MRF-ICM based approach fully utilizes the features of GA to provide better initialization for ICM algorithm and thus effectively reduces the misclassification problem which is observed in Fig. (4). The convergence plot of the ICM algorithm is shown in Fig. (5).

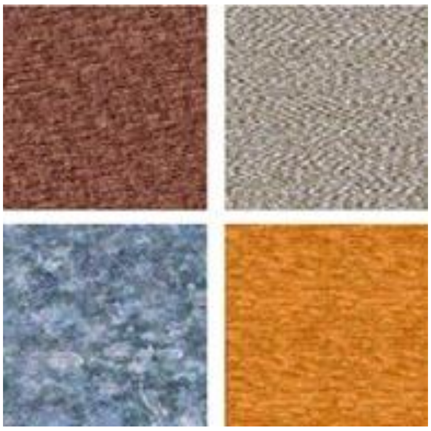


**Figure 1:** Original colored texture image

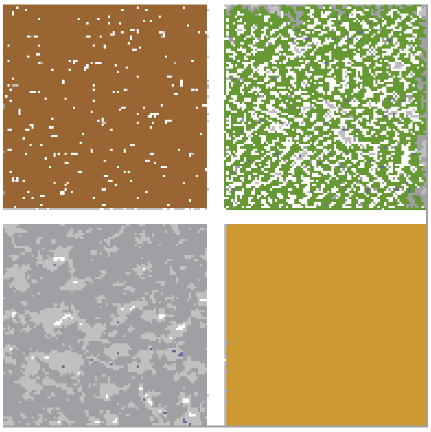


**Figure 2:** Segmented image using GA clustering

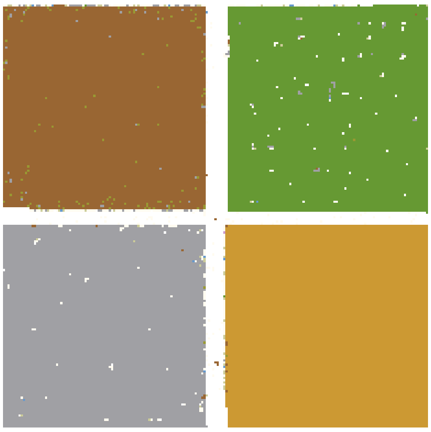


**Figure 3:** Segmented image using ICM algorithm

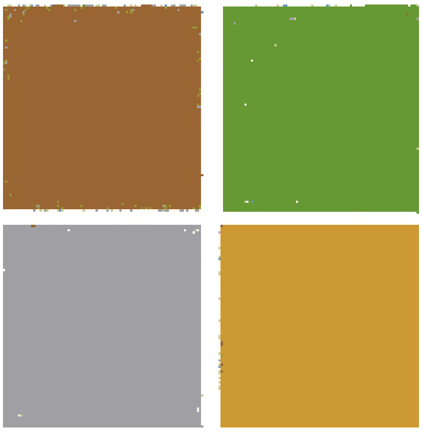


**Figure 4:** Segmented image using Hybrid GA-ICM algorithm

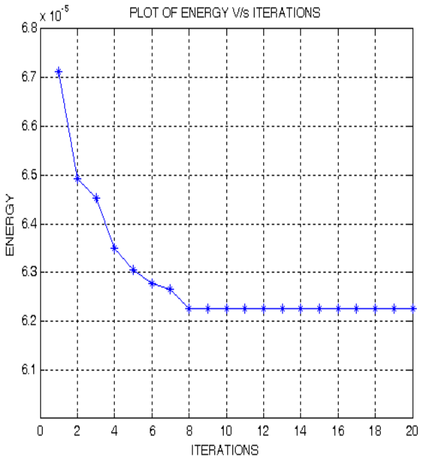


**Figure 5:** Convergence plot of ICM algorithm

The results of the segmentation can be judged by visual interpretation and the error rate. Because the used images are synthetic images error rate can be easily calculated by the following equation,

$$\frac{Number\ of\ misclassified\ pixels}{Total\ number\ of\ pixels\ in\ the\ image}\, x100$$

It is found that, the GA-MRF method has the least error rate compared to other two methods and the error rate of each method is tabulated in the Table 1.

GA based clustering is tested also for medium resolution SPOT image in the multispectral mode having two bands. It is found that GA clustering produces near optimal solutions when the image is non textured as in case of SPOT image. Fig. (6) shows SPOT image of part of the city of Kolkata in the near infrared band. The segmented results are shown in Fig. (7). It can be noted that the GA clustering has yielded optimal results.

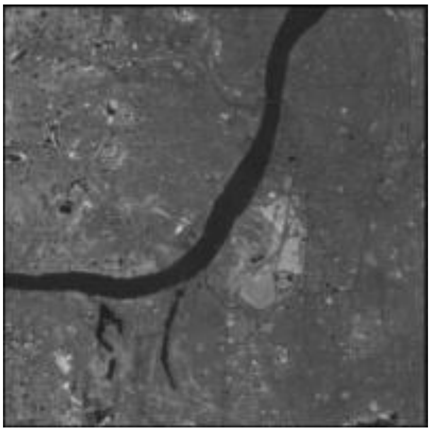


Figure 6: Spot image of Kolkata in near IR band

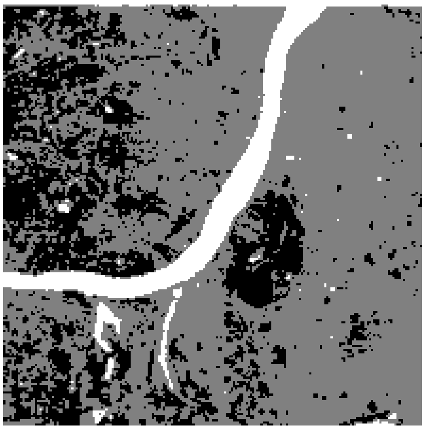


Figure 7: Segmented SPOT image using GA clustering

TABLE 1: ERROR RATE OF SEGMENTATION RESULTS

| | **Genetic algorithm** | **GMRF based segmentation** | **GA-GMRF based Segmentation** |
|---|---|---|---|
| **Error Rate (%)** | More than 40% | 11.59 | 1.2818 |

## 6. Discussion and Conclusion:

Comparing the performances obtained with every single algorithm we can conclude that,

- Genetic algoritm based segmentation approach depends only on the spectral information of the pixel without taking into account the spatial information and hence the results are very noisy. The method is optimal only if it is applied to coarse or medium resolution imagery with no noise and cannot be applied for high resolution textured images. Besides the time complexity is also more.

- GMRF model based segmentation using ICM algorithm has the advantage of the computaional cost. Among the segmentation methods considered in this paper, this is the one that has got the lowest computational time. But the disadvantage of this schemata is that the desirable results are not obtained as the ICM algorithm depends heavily on the initial segmentation.

- Hybrid GA-GMRF method generates the best results among all the three methodologies considered in terms of accuracy and time complexity for multispectral textured images. The reason is due to the fact that firstly the model takes into account not only spatial interaction within each of the color bands but also the interaction between the different bands. Secondly it uses GA for the initialization of the ICM algorithm. Hence it has the advantage of combining the fast convergence of ICM and global exploration of GA.

## References

[1] Lu and Q.Weng, A Survey of image classification methods and techniques for improving classification performance, International Journal of Remote Sensing, 28(2007), pp 823 – 870.

[2] C.H.Chen, Signal and Image Processing for Remote Sensing, Taylor and Francis, Boca Raton, 2007.

[3] Sanghamitra Bandyopadhyay and Ujjwal Maulik, Genetic clustering for automatic evolution of clusters and application to image classification, Pattern Recognition, 35 (2002), pp. 1197–1208.

[4] Zhicun Tan and Ruihua Lu, Application of Improved Genetic K-Means Clustering Algorithm in Image Segmentation, IEEE Conf. Education Technology and Computer Science, 2 (2009), pp. 625-628.

[5] Anne Puissant, Jacky Hirsch and Christiane Weber, The utility of texture analysis to improve per-pixel classification for high to very high spatial resolution imagery, International Journal of Remote Sensing, 26 (2005), pp.733-745.

[6] Din-Chang Tseng and Chih-Ching Lai, A genetic algorithm for MRF-based segmentation of multi-spectral textured images, Pattern Recognition Letters, 20 (1999), pp. 1499-1510.

[7] Zoltan Kato and Ting-Chuen Pong, A Markov random field image segmentation model for color textured images, Image and Vision Computing, 24 (2006), pp. 1103-1114.

[8] Brandt C. K. Tso and Paul M. Mather, Classification of multisource remote sensing imagery using genetic algorithm and Markov random fields, IEEE transactions on Geoscience and remote sensing, 37 (1999), pp. 1255-1260.

[9] J.Besag, Spatial interaction and statistical anlysis of lattice systems, J. Royal Statistical Society B, 36 (1974), pp192-236.

[10] Brandt C. K. Tso and Paul M. Mather, Classification methods for remotely sensed data, 2nd Edition, CRC press, 2009.

[11] Dileep Kumar Panjwani and Glenn Healey, Markov random field models for unsupervised segmentation of textured color images, IEEE Trans. Pattern Analysis and Machine Intelligence, 17 (1995), pp. 939-954.

[12] Y. Li and P. Gong, An efficient texture image segmentation algorithm based on the GMRF model for classification of remotely sensed imagery, International Journal of Remote Sensing, 26 (2005), 22, pp. 5149-5159.

**Mridula J:** was born in 1980. She received B.E degree in computer science engineering from Visveswaraiah Technological University in the year 2002. She is presently perusing M.Tech degree in electrical engineering department under specialization electronics system and communication at National Institute of Technology, Rourkela, India. Her research interests include image processing, computer vision and pattern recognition.

**Dr. Dipti Patra:** was born in 1968. She received B.Sc engineering in electrical engineering, M.E degree in electronic systems and communication and PhD in image processing from National Institute of Technology, Rourkela, India in 1989, 1993 and 2006 respectively.
She is currently an associate professor in the department of electrical engineering, National Institute of Technology, Rourkela, India. Her research interests include signal and image processing, computer vision and pattern recognition.

# A Fuzzy Type-2 Model of IT Governance Concerns

Sang-Hyun Lee[1], Kyung-Il Moon[2]

[1]Department of Computer Engineering, Mokpo National University, jeonnam, korea
[2]Department of Computer Engineering, Honam University,
Seobong-dong, Gwangsan-gu, Gwangju 506-714, Korea
leesang64@gmail.com
kimoon@honam.ac.kr

***Abstract***: IT governance is a topic that has been increasingly discussed since the mid nineties. IT Governance or ICT Governance, is a subset discipline of Corporate Governance focused on information technology (IT) systems and their performance and risk management. However, a shared view on the basic concepts of IT governance is lacking and practitioners do not use present IT governance frameworks to support their decision-making. Now, the definitions of IT governance are broad which in turn implicate difficult and inaccurate assessments. Eventually, IT governance is the preparation for, making of and implementation of IT-related decisions regarding goals, processes, people and technology on a tactical or strategic level. This paper presents a fuzzy type-2 model to understand the relationship among IT Governance concerns, and to assess IT governance complexity. It can be used for a good understanding how the concerns of IT governance behave, how they interact and form the behavior of the whole system. This model is employed to compare how IT governance is defined in practitioners and Cobit.

***Keywords***: Complexity system, Fuzzy type-2 logic, IT Governance concerns**.**

## 1. Introduction

There are reasonable frameworks and definitions of IT governance, but practitioners within the field do not agree with them and not strictly follow them in their quest for IT governance improvement. This has been stated previously, c.f. (Cumps 2006, Dahlberg 2006), but the different concerns of IT governance between literature, practitioners, and best practice frameworks have not been fully investigated. It has been the belief of the authors that IT governance would be defined differently in literature and by practitioners. Therefore, the IT governance concerns are needed to compare how literature and practitioners define the field. First, the theoretical concerns show that “Strategic,” “Monitor,” and “People” have been frequently mentioned within many articles. IT governance mainly comprises strategic concerns according to literature. Regarding the decision-making (DM) phases, monitoring of IT related decisions is emphasized. In literature, IT control frameworks and legislations stipulating the need for internal control are often referred. Technology is not the mayor concerns to decide on, and literature rather stresses the importance of establishing roles and responsibilities, and an accountability framework that supports the organization’s strive to achieve its business goals. A survey with practitioners is described more thoroughly in Simonsson (2006). The survey was made using a commercial, web-based tool for online surveys. 18 participants responded to the survey. Among these, 72 % primarily had the role of consultants in IT governance change projects, but a few CIOs, security and risk managers, and internal auditors also participated. All respondents claimed previous involvement in at least one IT governance change project, above 80 percent in two such projects or more. According to the practitioners responding the survey, IT governance DM is mainly a strategy issue while tactical decisions are less important. Emphasis is put on understanding the situation at hand prior to making a decision, and solving practical issues regarding how each decision is carried out, such as assigning DM authority, coordinating resources, and aligning IT decision-making with external factors. Monitoring the implementation of decisions already made receives somewhat less attention from the practitioners, according to the survey. Practitioners do however agree that IT decisions are mainly about IT goal setting; strategy development, alignment of IT and business goals, etc. Another important topic is the establishment of a corporate decision-making structure with clear assignment of roles and responsibilities, while IT processes and technology issues are less stressed.

Cobit is a well-known framework for IT governance improvement, risk mitigation and IT value delivery (Ridley 2004, Holm Larsen 2006, Debraceny 2006). A survey with Cobit is described more thoroughly in (Simonsson and Johnson, 2006). Strategy, Monitoring and Processes are received the highest marks. Compared to the concerns identified in literature, it is clearly visible that Cobit is focused on decisions regarding the processes while people receive less attention. Further, Cobit spends more effort in discussing the understand phase and less on the decide phase. Strategic concerns are most often dealt with, while tactical concerns are only briefly discussed. Compared to the practitioners’ concerns, Cobit emphasizes processes but lacks hands-on support for decisions regarding people and goal settings. Also, Cobit focuses on decision monitoring to a larger extent than what practitioners do, while the opposite is valid for understand and decide.

Most authors agree on IT governance as a top management concern of controlling IT’s strategic impact, and the value delivered to the business c.f. (Weill 2004, ITGI 2005, De Haes 2005, Ribbers 2002). But whether the core of IT governance is a set of structures, processes and relational mechanisms (De Haes 2005), bundled performance metrics to aid IT process monitoring (ITGI 2005) or cascaded Balanced Scorecards (Kaplan 1996, Van Grembergen 2004) is not agreed upon. There is also a gap between what is stated in literature and the opinions of practitioners: The theories developed in literature are not frequently used by consultants or CIOs (Cumps 2006, Dahlberg 2006). Control Objectives for Information and related Technology, Cobit, is the most renowned framework for support of IT governance concerns (ITGI 2005, Guldentops 2004), but it does not really address the concerns considered important in literature and by practitioners (Simonsson and Johnson, 2006).

The difference between literature, practitioner and Cobit seems to lay in those very interconnections (and interactions) between the concerns of IT governance, and all that they can result in. It is not enough to understand the nature of the "more than one or many concerns" themselves, it's also necessary to understand the exact nature of the interconnections and how they affect the behavior of the whole IT governance. When there are such interconnections and they are "Not simple" and "Difficult," a complexity system can be used. The only consensus on what makes something complex is that the definition of complexity is evolving. IT governance is also a complexity system. But how people apply such terms can vary widely, making it difficult for the rest of us to zero in on the essence of complexity systems. It should come as a relief to know that the experts don't always agree either. In this respect, Lee, Cho and Moon (2010) address really the concerns considered important in literature and by practitioners. They suggest a fuzzy collective behavior model based on IT governance concerns.

The purpose of this paper is to suggest another IT Governance concerns model based on fuzzy type-2 logic, and to represent how the concerns should be really addressed by practitioners and Cobit. This model can be regarded as more realistic version of Lee, Cho and Moon's study. When we go about designing IT Governance structure as a control system, this model will be guiding its organization in view of practitioners. Returning to complex interactions of IT Governance concerns, Practitioners feel a need to attempt relating the system/process to be controlled, the tasks involved in controlling it, the control system, and the context of use. Section II defines IT Governance concerns considered important in literature as a complexity system, and by practitioners neural-network learning as a parameter estimation problem and describes the basic formulation and properties of type-2 fuzzy logic. Section III presents a design process of IT Governance concerns model based on fuzzy type-2 logic. In section IV, it is considered whether and how fuzzy type-2 logic applies to IT governance assessment. Section V reports simulation results comparing the proposed approach with Cobit.

## 2. IT Governance Concerns and Fuzzy Concept

Complex systems typically have some characteristic properties, but the extent to which a particular system exhibits any given property can vary. In this respect, IT governance system includes the fuzzy concepts to a great extent. What makes IT governance complex is the number of decisions that have to be made regarding its design, the number of people or organizations that have to be involved in those decisions, and the fact that they' re probably inconsistent. IT governance has inconsistent objectives, so decisions have to be negotiated. An important consideration is whether a system, such as an enterprise, is planned for by some unified process or independently developed and later merged. It is a matter of control. If the system is a set of independent systems, and you have no control over those decisions, then, maybe it' s a different kind of system. But again, we don' t see a clear-cut distinction between whether an enterprise system or a system of systems involves more or less of this independent decision making. Finally, complexity increases when the number of systems (developed as standalone entities or with interoperability in mind) and disparate stakeholders increase. The complex part is the interaction of people. Even though the concerns of IT governance may be more or less different, their interconnections and interactions can produce the desirable results, explaining why it is hardly a complex system. This section presents IT governance concerns using some complexity profiles, which are related to a fuzzy theory for characterizing the collective behavior of IT governance concerns.

### 2.1 Domain, Scope and DM Complexity

To understand the collective or cooperative behavior of IT governance system, it must be developed concepts that describe the collective behavior in a more general way. It is much easier to think about the problem of understanding collective behavior using the concept of a complexity profile. The complexity profile focuses attention on the scale at which a certain behavior of IT governance concerns is visible to an observer, or the extent of the impact it can have on its environment. The complexity profile counts the number of independent concerns that are visible at a particular scale and includes all of the concerns that have impact at larger scales. The central point is that when the independence of IT governance concerns is reduced, scale of behavior is increased. To make a large collective behavior, the individual concerns that make up this behavior must be correlated and not independent.

First, the domain complexity of IT governance concerns denotes a nonlinear function of what the decisions should consider. It comprises four complexity variables: Goals, processes, people and technology. Goals include strategy-related decisions, development and refinement of IT policies and guidelines, and control objectives used for performance assessments. Processes include the implementation and management of IT processes, e.g. acquisition, service level management, and incident management. People include the relational architecture within the organization, and the roles and responsibilities of different stakeholders. Finally, IT governance is of course about managing the technology itself. The complexity variable technology represents the physical assets that the decisions consider, such as the actual hardware, software and facilities. The practitioners prioritized the complexity variables as they are presented below.

- People variable: It denotes the relational structure within the organization, and the roles and responsibilities of different stakeholders.
- Goal variable: It denotes the development and refinement of an IT strategy, policies, guidelines, and control objectives to monitor whether the goals are achieved.
- Technology variable: It denotes the physical IT-related assets.
- Process variable: It denotes the implementation and management of IT processes and related activities and procedures.

Second, the scope complexity denotes a nonlinear function of different impacts implied by each decision. There is a long term aspect and a short time aspect of every decision that is made. Consequently, there is also a connection between the timeline of the decision and the level at which it is made. Top management make long time plans and set strategic goals, while lower management are authorized to make decisions affecting the near time. Further, strategically important decision requires more preparation than a tactic decision. The scope dimension is used to differentiate between different levels of decision-making. Firstly, there are detailed, rapidly carried out, IT-focused tactic decisions. Examples of tactic decisions include whether to upgrade a certain workstation today or tomorrow, how to configure a user interface that is only used internally, or the manning of a single IT project. There also exists top management, low detailed, business oriented strategic decisions with long timeline. A strategic decision might consider whether it is most appropriate to develop an application in-house or to purchase it off the shelf, or how the performance of IT processes should be reported to top management. The practitioners prioritized the dimensional units as they are presented below.

- Strategic variable: It is related to top-level management decisions, with few details and primarily a business impact. The decision features a business oriented focus with long timeline.
- Tactic variable: It is related to low-level management decisions, with many details and an impact primarily on IT. The decision has typically an operations focus and a short timeline.

Third, the decision making complexity denotes a nonlinear function of different steps required to make decisions within the different domains. This complexity deals with the relation between IT, and the models of the reality used for decision making. Before making any decision regarding e.g. the outsourcing of a helpdesk function, the organization must be clearly understood. Facts have to be thought over and investigated, and transformed into a model. The model might be a simple cognitive map, present nowhere else but in the head of the decision-maker, or a more formalized, abstract model put on print. This process of analysis and understanding is denoted the understanding phase. Once the model is created, the actual decision can be made according to corporate IT principles, in a timely manner, by the right individuals, etc. In the IT governance definition, this is represented by the decision phase, which also includes planning of how to make the decision. Finally, a decision is of little use unless its implementation is followed up and monitored. This can be accomplished by implementing control objects for each process in order to assess real-world performance. The decision-makers compare the state of the reality with the values obtained from the models. Note that these steps are not necessarily formal, but nevertheless exist in one way or another upon making decisions. The practitioners prioritized the complexity variables as they are presented below.

- Understand variable: It denotes the collection of information needed to make a correct decision.
- Decide variable: It is related to how and by whom the decision is made. Decisions are made according to corporate IT principles, at the correct level in an adequate forum, e.g. by a steering committee.
- Monitor variable: It denotes how the implications of a decision are monitored.

### 2.2 Formulation of type-2 fuzzy logic

A distinct advantage of type-2 fuzzy logic is that it is very powerful in handling uncertainties. IT Governance concerns model based on type-2 fuzzy logic can explicitly consider the domain, scope and decision-making complexity variables. By utilizing membership functions in type-2 fuzzy logic capable of handling uncertainty, the model can generate some collective behaviors of IT Governance concerns with reasonable accuracy. Compared with type-1 fuzzy logic, type-2 fuzzy logic has different definitions for membership functions. It also has its own set of operators. With these operators and the extension principle, the properties of type-2 fuzzy logic can be derived from type-1 fuzzy logic. The definition of type-2 fuzzy sets is given by

$$A = \{((x,u), \mu_A(x,u)) \mid \forall x \in X, \forall u \in J_x \subseteq [0,1]\}\cdot$$

Here $0 \le \mu_A(x,u) \le 1$ . A type-2 fuzzy set has additional dimension, $u$, associated with the membership value $\mu_A(x)$. That is, it has a membership function that would yield multi-valued $\mu_A(x)$ for $x=x'$. In particular, $u$ can be viewed as a type-1 fuzzy set, with the membership function $J_{x'}$ in three-dimensional space. $J_x$, a vertical slice of $\mu_A(x, u)$, is called the secondary membership function, denoted by

$$\mu_A(x = x', u) \equiv \mu_A(x') = \int_{u \in J_{x'}} f_{x'}(u)/u\,,$$

where $0 \le f_{x'}(u) \le 1$ and $f_{x'}(u)$ is the amplitude of a secondary membership function called a secondary grade. A integration symbol means that the type-2 fuzzy set has a membership $u$ associated with grade $f_{x'}(u)$ for $x=x'$. Note that, as is customary in the fuzzy logic notation, the integration symbol is not an integration operator but a symbol that represents the collection of all points of $u$ in $J_{x'}$.

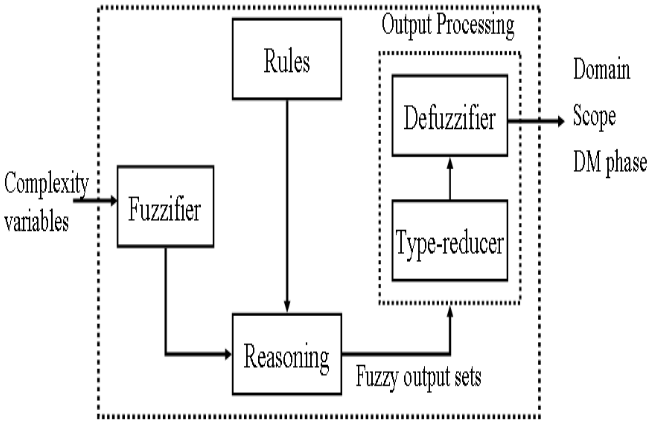


**Figure 1.** A type-2 FLS for IT Governance concerns

An interval type-2 fuzzy set is a special case of type-2 fuzzy sets in which the secondary membership functions are defined by $f_x(u)=1$, $\forall u \in J_x \subseteq [0,1]$. For $x=x'$, the primary membership value $u$ can be represented as an interval $[l, r]$. Since $X' \in X$, we can then drop the prime notation and refer to $\mu_A(x)$ as a secondary membership function. The type-2 fuzzy set can be defined as

$$A = \int_{x \in X} \mu_A(x)/x = \int_{x \in X} [\int_{u \in J_x} f_x(u)/u]/x,\ J_x \subseteq [0,1]$$

The domain of a secondary membership function is called the primary membership of $x$. $J_x$ is the primary membership of $x$, where $J_x \subseteq [0,1]$ for $\forall x \in X$. As in type-1 fuzzy logic, once the fuzzy set is defined, the fuzzy inference can be obtained based on the fuzzy set and the choice of operators for operations on the fuzzy set.

A type-2 fuzzy logic system is a rule-based system comprising five components: fuzzifier, fuzzy rules, inference, type-reducer and defuzzifier, as shown in Fig. 1. All the rules have antecedents and consequents. Based on the input and the antecedents of the rules, the fuzzy inference process will compute a 'firing level' for each rule, combine the consequents of the rules according to the firing level and then generate the resulting type-2 fuzzy set. The type-reducer and defuzzifier will perform the type-reduction and defuzzification to get a crisp value from the type-2 fuzzy set. This crisp value is the output of the type-2 fuzzy logic system.

## 3. Type-2 Fuzzy Logic for IT Governance Concerns

In this section, we discuss the type-2 fuzzy logic system in conjunction with our application. The type-2 fuzzy logic system developed for IT Governance concerns has the following five assumptions:

1. All the type-2 fuzzy sets are interval type-2 fuzzy sets.
2. Antecedent and consequent membership functions are Gaussian primary membership functions.
3. Input membership functions are Gaussian primary membership functions, with uncertain standard deviation.
4. The fuzzy operations use product implication and t-norm.
5. The type-reduction uses a centre-of-sets method and the defuzzification process uses a simple average method.

It is in general difficult to determine the exact probability density function for such a system. The interval type-2 fuzzy set and Gaussian primary membership functions are quite robust compared with other choices. These assumptions are made for simplifications in computation.

### 3.1 Membership functions

The interval type-2 fuzzy set has an upper membership function and a lower membership function. This property can be conveniently utilized to generate a prediction interval. In our formulation, Gaussian primary membership functions are used in two ways. We consider the use of a Gaussian primary membership function with a fixed standard deviation, $\sigma$, but uncertain mean in the following form:

$$\mu_A(x) = \exp\left[-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2\right], \; m \in [m_1, m_2]$$

Denote exp(·) by N($m$, $\sigma$; $x$). For each value of m, there is a corresponding membership curve. The choice of $m_1$ and $m_2$ is based on the historical information. In the case of the interval type-2 fuzzy set, the upper membership function is defined by

$$\overline{\mu}_A(x) = \begin{cases} N(m_1, \sigma; x), & x \le m_1 \\ 1, & m_1 \le x \le m_2 \\ N(m_2, \sigma; x), & x > m_2 \end{cases}$$

whereas the lower membership function is defined by

$$\underline{\mu}_A(x) = \begin{cases} N(m_2, \sigma; x), & x \le (m_1 + m_2)/2 \\ N(m_1, \sigma; x), & x \ge (m_1 + m_2)/2 \end{cases}$$

Similarly, we can consider the use of a Gaussian primary membership function with fixed mean, $m$, but uncertain standard deviation:

$$\mu_A(x) = \exp\left[-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2\right], \; \sigma \in [\sigma_1, \sigma_2]$$

For each value of $\sigma$, there is a corresponding membership curve. In the case of the interval type-2 fuzzy set, the upper membership function is

$$\overline{\mu}_A(x) \equiv N(m, \sigma_2; x)$$

The lower membership function is defined by

$$\underline{\mu}_A(x) \equiv N(m, \sigma_1; x)$$

### 3.2 Design of an IT Governance model based on type-2 fuzzy logic

There are five steps involved in the design of a type-2 fuzzy logic-based IT Governance concerns model:

#### 3.2.1 Design of the fuzzifier

The input data of a fuzzy logic system are a set of crisp values. The function of the fuzzifier is to transform the crisp values into a set of fuzzy values, that is, variables with a fuzzy membership function. In the IT Governance model, the fuzzifier will take the IT Governance concerns variable $x_k$ at the $k^{th}$ interval, $x'_k$, as an input to generate a type-2 fuzzy set. The membership functions used in our model are Gaussian primary membership functions with uncertain standard deviations given by

$$\mu_k(x_k) = \exp\left[-\frac{1}{2}\left(\frac{x_k - x_k'}{\sigma}\right)^2\right], \sigma \in [\sigma_1, \sigma_2]$$

It is reasonable to make the fuzzifier interval dependent since the mean concerns variables at different intervals are very different. The variance of the concerns variable, however, falls into a range with its boundary values determined from data across different data sets. The resulting membership function is shown in Fig. 2. For simplicity in notation, we omit the subscript to denote days here as well as in the following subsections.

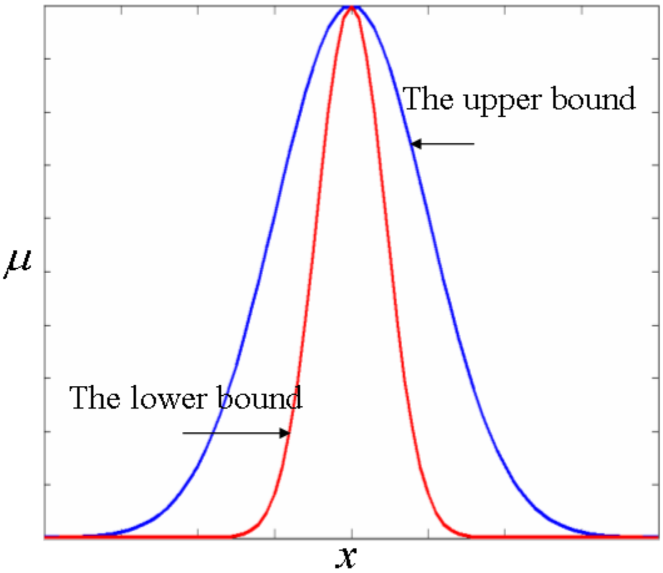


**Figure 2.** The type-2 fuzzy membership function with uncertain variance

### 3.2.2 The Construction of fuzzy rules

Once the set of type-2 fuzzy membership functions is defined, the next step is to construct the fuzzy rules for processing the fuzzy input. In our case, concerns data are used to build the fuzzy rules. It is similar to a training process in which data sets are utilized one by one to establish the centre of the fuzzy sets that appear in the antecedents and consequents of the rules. The $l^{th}$ fuzzy rule in the set with a total of $M$ rules has the format:

$R_l$ : IF $x_1$ is $F_1^{\,l}$ and . . . and $x_p$ is $F_p^{\,l}$

THEN $y$ is $G^l$, $l$=1, . . . , $M$

where $F_i^{\,l}$ is the antecedent, $G^l$ is the consequent of the $l^{th}$ fuzzy rule, $x_1$, . . . , $x_p$ are the input of the fuzzy logic system and $y$ is the output for this rule, which will be utilized in fuzzy inference. In our model, rules are mostly developed based on historical information. Since the data available to us are quite limited, we use a single data set to construct a single fuzzy rule. This does not have to be the case. There are many alternative ways to construct fuzzy rules.

### 3.2.3 Design of the fuzzy inference engine

Fuzzy inference is the key component of the fuzzy logic system. Based on the input and the antecedents of the rules, it calculates a 'firing level' for each rule and then applies these firing levels to the consequent fuzzy sets. First, the union of two type-2 fuzzy sets $A$ and $B$ is expressed by

$$\mu_{A\cup B}(x) = \mu_A(x) \vee \mu_B(x), x \in X$$

The symbol $\vee$ denotes the join operator. Computationally, for any $x$, the join operator will enumerate all the possible combinations of u and w, take the maximum of $u$ and $w$ as the resulting primary membership value and take the minimum or product of the two secondary grades, $f_x(u)$ and $g_x(w)$, as the resulting secondary grade. In the interval type-2 fuzzy set, the join operator will be simplified as $\vee F_i$, $i$=1, …, $n$, representing the join of $n$ interval type-1 sets $F_1$, …, $F_n$, having domains [$l_1$, $r_1$], …, [$l_n$, $r_n$], respectively, or [($l_1 \vee l_2 \vee \ldots \vee l_n$), ($r_1 \vee r_2 \vee \ldots \vee r_n$)].

The intersection of two type-2 fuzzy sets, $A$ and $B$, is expressed by

$$\mu_{A\cap B}(x) = \mu_A(x) \wedge \mu_B(x), x \in X$$

The symbol $\wedge$ here denotes the meet operator. Computationally, for any $x$, the meet operator will enumerate all the possible combinations of $u$ and $w$, take the minimum or product of $u$ and $w$ as the resulting primary membership value and take the minimum or product of the two secondary grades, $f_x(u)$ and $g_x(w)$, as the resulting secondary grade. This operation will give a new type-2 fuzzy set. For the interval type-2 fuzzy set, the meet operator will be simplified as $\wedge F_i$, $i$=1, …, $n$, representing the meet of n interval type-1 sets $F_1$, …, $F_n$, having domains [$l_1$, $r_1$], …, [$l_n$, $r_n$], respectively, or [($l_1 \wedge l_2 \wedge \ldots \wedge l_n$), ($r_1 \wedge r_2 \wedge \ldots \wedge r_n$)].

In type-2 fuzzy logic systems, the output type-2 fuzzy set of the fuzzy inference of the $l^{th}$ fuzzy rule is:

$$\mu_{B^l}(y) = \mu_{G^l}(y) \wedge \{[\vee_{x_1} \mu_{X_1}(x_1) \wedge \mu_{F_1^l}(x_1)] \wedge \cdots$$
$$\wedge [\vee_{x_p} \mu_{X_p}(x_p) \wedge \mu_{F_p^l}(x_p)]\}, \, y \in Y$$

where $\mu_{X_i}(\cdot)$ is the type-2 membership function of the input, $\mu_{F_i^l}(\cdot)$ is the type-2 membership function of the antecedent $i$ of the $l^{th}$ rule, and $\mu_{G^l}(\cdot)$ is the type-2 membership function of the consequent of the $l^{th}$ rule. The above equation can be written as

$$\mu_{B^l}(y) = \mu_{G^l}(y) \wedge F^l(x')$$

where $F^l(x')$ is the firing level of the input data.

Since the interval type-2 fuzzy sets are used for IT Governance concerns, the firing level will also be an interval set:

$$F^l(x') = [\underline{f}^l(x'), \overline{f}^l(x')] \equiv [\underline{f}^l, \overline{f}^l]$$

Here

$$\underline{f}^l = \int_{\{x_1 \in X_1, \ldots, x_p \in X_p\}} [\underline{\mu}_{X_1}(x_1) \wedge \underline{\mu}_{F_1^l}(x_1)] \vee \cdots \vee$$
$$[\underline{\mu}_{X_p}(x_p) \wedge \underline{\mu}_{F_p^l}(x_p)] / (x_1, \ldots, x_p)$$

$$\overline{f}^l = \int_{\{x_1 \in X_1, \ldots, x_p \in X_p\}} [\overline{\mu}_{X_1}(x_1) \wedge \overline{\mu}_{F_1^l}(x_1)] \vee \cdots \vee$$
$$[\overline{\mu}_{X_p}(x_p) \wedge \overline{\mu}_{F_p^l}(x_p)] / (x_1, \ldots, x_p)$$

### 3.2.4 Type-reduction

For fuzzy reasoning of IT Governance concerns, the type-2 fuzzy set generated from the previous steps needs to be converted to a crisp value. This is realized through Steps 4) and 5), type-reduction and defuzzification. Type-reduction generates the centroid type-1 fuzzy set of a type-2 fuzzy set. There are several other methods for type-reduction, such as centre-of-sums type-reduction, height type-reduction, modified height type-reduction and centre-of-sets type-reduction. For the sake of computational efficiency, we employ the centre-of-sets type-reduction method. Instead of combining the type-2 sets from the fuzzy inference of all the rules before reduction, the centre-of-sets type reduction makes use of the centroid method to reduce the resulting type-2 sets from each rule and obtain a type-1 set [$L_i$, $R_i$] for each rule $i$.

The weighted combination of these type-1 sets is then used to get the final type-1 set [$y_L$, $y_R$]:

$$y_L = \sum_{i=1}^{M} f_l^i L_i / \sum_{i=1}^{M} f_l^i$$

$$y_R = \sum_{i=1}^{M} f_r^i R_i / \sum_{i=1}^{M} f_r^i \cdot$$

Here $f_l$ and $f_r$ are the firing level corresponding to $y$ of rule $i$ that will maximize $y_L$ and minimize $y_R$. Each $f$ can be enumerated in the interval $[\underline{f}^i, \overline{f}^i]$.

### 3.2.5 Defuzzification

Defuzzification is the last step to get the final forecast result. The defuzzification of a type-2 fuzzy logic system is identical to the defuzzication of a type-1 fuzzy logic system. There are also several methods for the defuzzification of a type-1 or a type-2 fuzzy logic system, such as the centroid defuzzifier, centre-of-sums defuzzifier, height defuzzifier, modified height defuzzifier and centre-of-sets defuzzifier. A commonly used defuzzification is the centroid method

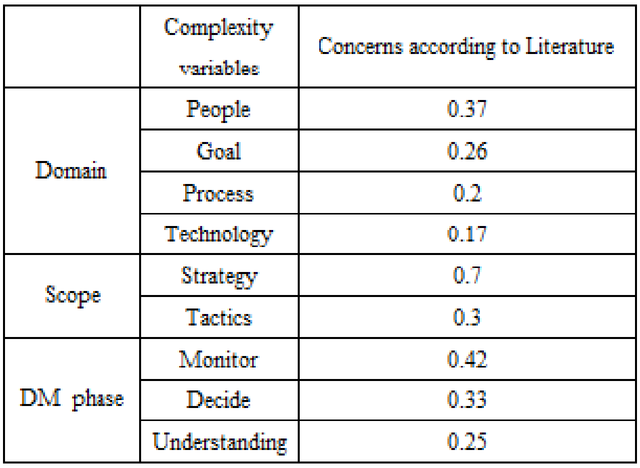
$$y_c(x) = \sum_{i=1}^{M} y_i \mu_B(y_i) / \sum_{i=1}^{M} \mu_B(y_i) ,$$

in which the range of $y$ is discretized into $M$ points. The subscript 'c' denotes the centroid method. In the case of the interval set, we can defuzzify the interval [$y_L$, $y_R$] from type-reduction using the average of $y_L$ and $y_R$.

## 4. Application

The basis for IT governance application is the theoretical IT governance concerns. 100 sources of information on IT governance were identified when conducting an extensive literature search. The forums in which the articles have been published include the MIS Quarterly, Information Systems Control Journal, Information Systems Research, International Journal of Information Management, International Journal of Accounting Information Systems, and the Hawaii International Conference on System Sciences. 50 of the sources were selected randomly and analyzed in order to find common concerns. All statements used to create the IT governance complexity were again analyzed in order to extract the theoretical IT governance knowledge according to literature. The statements were classified and the number of times that each dimensional complexity was mentioned explicitly or implicitly was counted. figure 3 shows the results for these theoretical complexity variables, i.e. literature's concerns of IT governance. The total score for each dimension (e.g. Domain) is 100%.

The theoretical IT governance concerns show that the dimensional variables "People," "Strategic," and "Monitor" were most frequently used within the 50 articles and within their dimensions respectively. IT governance mainly comprises strategic concerns according to literature. The daily use of IT, all the operational concerns for bread-and-butter IT are surely important, but they are not in the scope of IT governance. Regarding the decision-making phases, monitoring of IT-related decisions is emphasized. In literature, IT control frameworks and legislations stipulating the need for internal control are often referred to, which is clearly reflected to in the figure. Technology issues are not the mayor concerns to decide upon, and literature rather stresses the importance of establishing roles and responsibilities, and an accountability framework that supports the business goals.



| | Complexity variables | Concerns according to Literature |
|---|---|---|
| Domain | People | 0.37 |
| | Goal | 0.26 |
| | Process | 0.2 |
| | Technology | 0.17 |
| Scope | Strategy | 0.7 |
| | Tactics | 0.3 |
| DM phase | Monitor | 0.42 |
| | Decide | 0.33 |
| | Understanding | 0.25 |

**Figure 3.** IT Governance concerns according to literature

Today, computing with words must still be done using numbers, and, therefore, numeric intervals must be associated with words. An earlier paper (Mendel, 1999) reported on an empirical study that was performed to determine how the scale 0–10 can be covered with words (or phrases). In typical engineering applications of fuzzy logic, we do not worry about this, because we choose the number of fuzzy sets that will cover an interval arbitrarily, and then choose the names for these sets just as arbitrarily. This works fine for many engineering applications when rules are extracted from data. One of the most striking conclusions drawn from the processed data is that linguistic uncertainty appears to be useful in that it lets us cover the 0~10 range with a much smaller number of terms than without it. Figure 4 depicts this for five terms (see Mendel, 2002). Solid lines are drawn between the sample means for the interval end-points and dashed lines are for the appropriate standard deviation about each mean end-point. Although five labels cover 0~10, there is not much overlap between some of them. It is when the standard deviation information is used that a sufficient overlap is achieved.

For simplicity, Figure 5 only illustrates fuzzy type-2 sets of the input variable "Process" and the output variable "Domain." They have each been divided into five overlapping sets labeled "none to very little," "some," "a moderate amount," "a large amount" and "a maximum amount." For the fuzzy sets of the input variable, we use the default ±0.1 standard deviation about each mean end-point. For the fuzzy sets of the output variable, we use the standard deviations corresponding to Fig.5. We can construct the fuzzy type-2 sets of Scope and DM complexity in a similar manner. Related to Domain complexity, twelve rules are defined in the rule base as shown below. We use the normalized rule weights for fuzzy pieces of IT governance concerns where twenty rules apply to the same conclusion.

If (Technology is none) then (Domain is none) (0.46)
If (Technology is some) then (Domain is none) (0.46)
If (Technology is moderate) then (Domain is some) (0.46)
If (Technology is large) then (Domain is some) (0.46)
If (Technology is maximum) then (Domain is moderate) (0.46)

If (Process is none) then (Domain is none) (0.54)
If (Process is some) then (Domain is none) (0.54)
If (Process is moderate) then (Domain is some) (0.54)
If (Process is large) then (Domain is moderate) (0.54)
If (Process is maximum) then (Domain is moderate) (0.54)
If (Goal is none) then (Domain is none) (0.7)
If (Goal is some) then (Domain is some) (0.7)
If (Goal is moderate) then (Domain is some) (0.7)
If (Goal is large) then (Domain is moderate) (0.7)
If (Goal is maximum) then (Domain is large) (0.7)
If (People is none) then (Domain is none) (1)
If (People is some) then (Domain is some) (1)
If (People is moderate) then (Domain is moderate) (1)
If (People is large) then (Domain is large) (1)
If (People is maximum) then (Domain is maximum) (1).

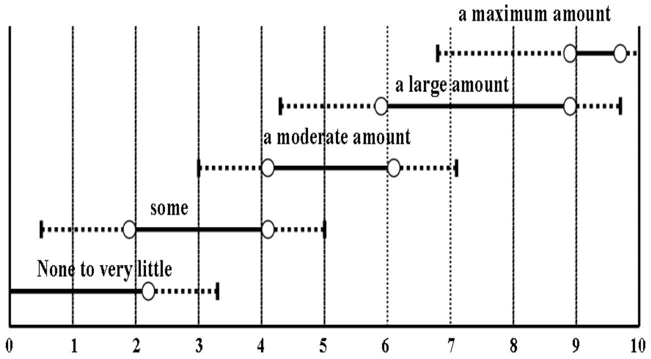


**Figure 4.** 5 labels and their intervals and uncertainty bands.

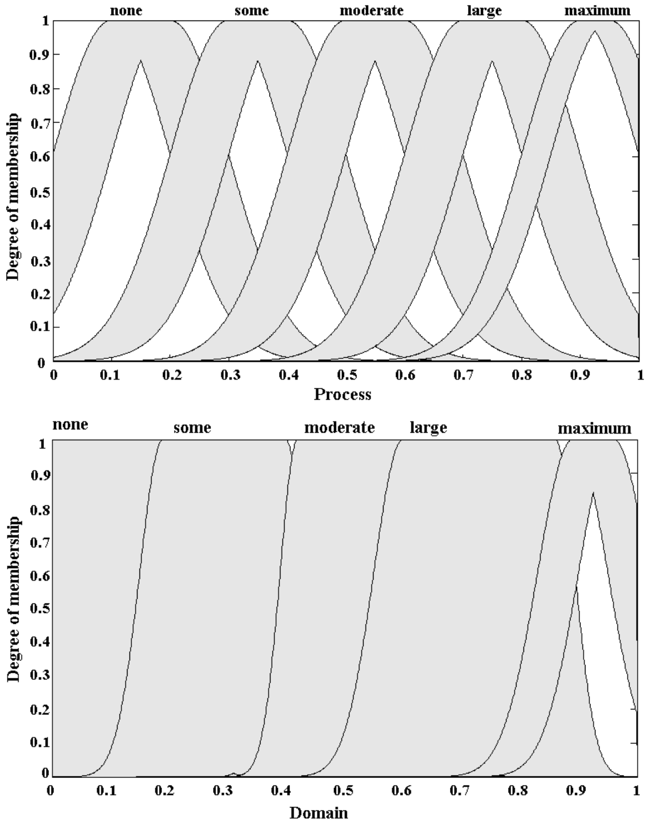
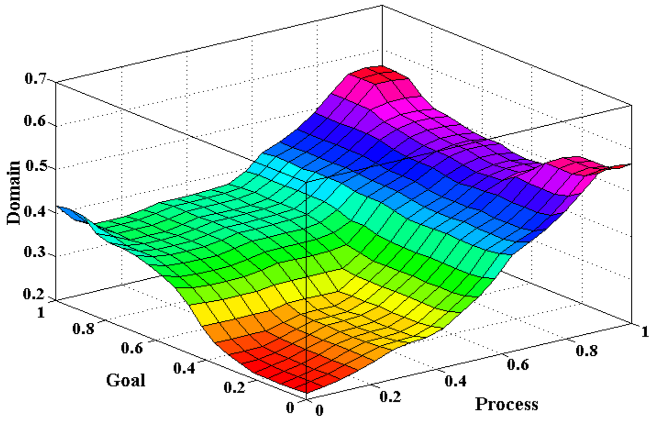
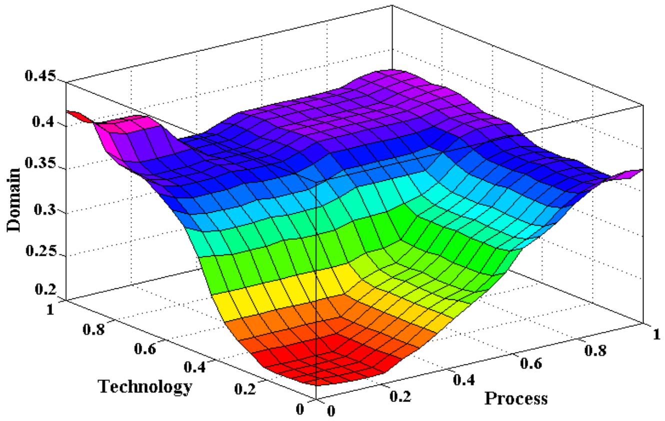
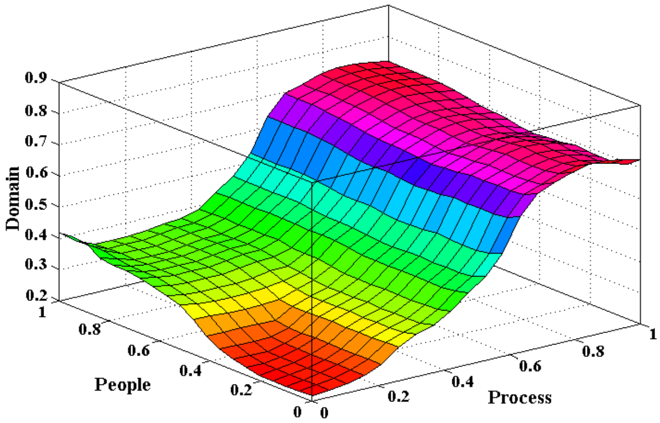
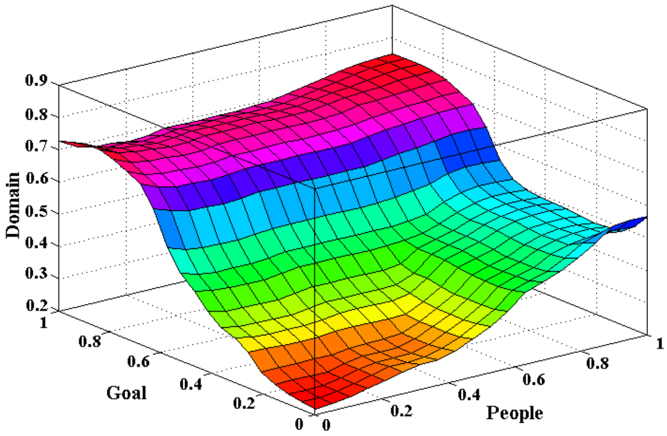


**Figure 5.** Fuzzy type-2 sets for “Process” and “Domain”









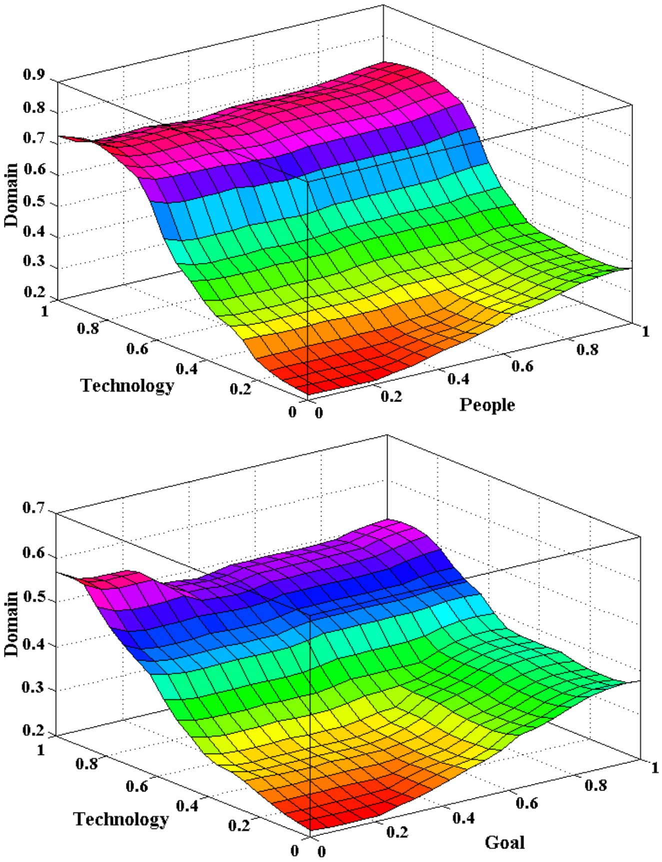
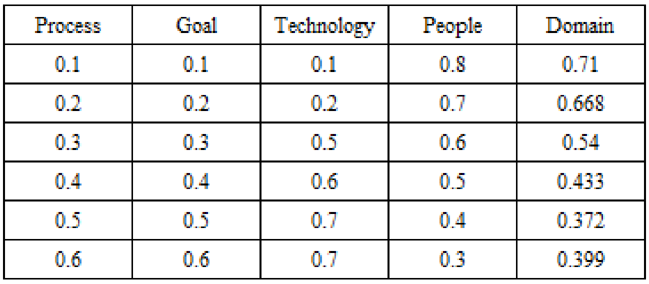


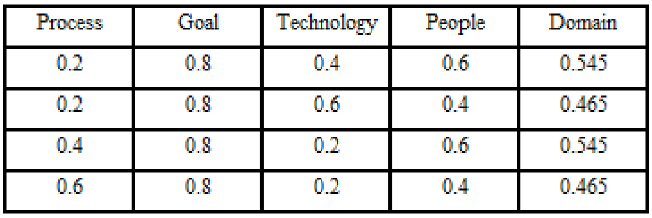
**Figure 6.** Mapping surface of Domain complexity.

Figure 6 shows the surface plots between the variables of Domain complexity. Clearly it is evident from the plot that "People" is more significant than other input variables. IT governance concerns in Literature denotes that "Technology" is less significant than other ones. But, considered as a whole, "Process" is less significant than other ones, c.f. (Fig 7). In particular, in proportion as "Goal" rises "Technology" concerns increase. Figure 7 illustrates the comparison of values estimated by using four input variables. According to the survey with practitioners, practitioner's concerns were mainly about IT goal setting, while IT processes and technology issues were less stressed. Figure 8 illustrates the comparison of values estimated by our fuzzy model. Here, "Goal" is more stressed. The result implies that the collective behaviors in which the concerns of practitioners affect other parts of the theoretical concerns must be no more complex. Generally speaking, it denotes that practitioners are faithful to the theoretical concerns. Compared to the concerns identified in literature, Cobit was focused on decisions regarding the processes while people receive less attention. Figure 9 illustrates the comparison of values estimated by our fuzzy type-2 model. The result denotes that there is discrepancy in the range of the concerns identified in literature.



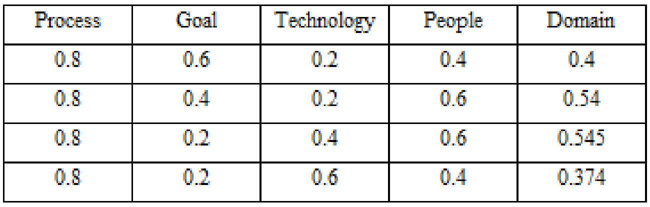
| Process | Goal | Technology | People | Domain |
|---|---|---|---|---|
| 0.1 | 0.1 | 0.1 | 0.8 | 0.71 |
| 0.2 | 0.2 | 0.2 | 0.7 | 0.668 |
| 0.3 | 0.3 | 0.5 | 0.6 | 0.54 |
| 0.4 | 0.4 | 0.6 | 0.5 | 0.433 |
| 0.5 | 0.5 | 0.7 | 0.4 | 0.372 |
| 0.6 | 0.6 | 0.7 | 0.3 | 0.399 |

**Figure 7.** Comparison of values by fuzzy model.



| Process | Goal | Technology | People | Domain |
|---|---|---|---|---|
| 0.2 | 0.8 | 0.4 | 0.6 | 0.545 |
| 0.2 | 0.8 | 0.6 | 0.4 | 0.465 |
| 0.4 | 0.8 | 0.2 | 0.6 | 0.545 |
| 0.6 | 0.8 | 0.2 | 0.4 | 0.465 |

**Figure 8.** Comparison of values by practitioners' concerns.

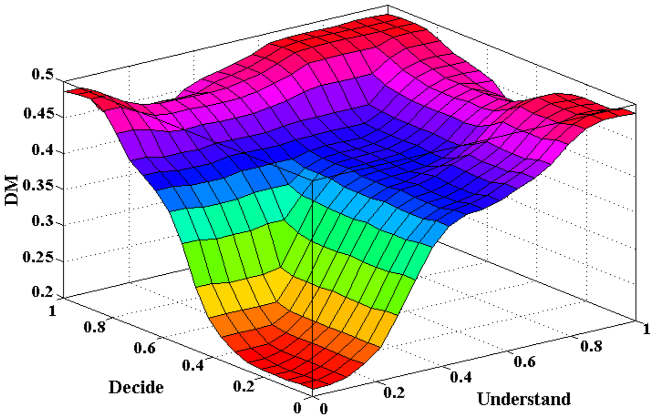


| Process | Goal | Technology | People | Domain |
|---|---|---|---|---|
| 0.8 | 0.6 | 0.2 | 0.4 | 0.4 |
| 0.8 | 0.4 | 0.2 | 0.6 | 0.54 |
| 0.8 | 0.2 | 0.4 | 0.6 | 0.545 |
| 0.8 | 0.2 | 0.6 | 0.4 | 0.374 |

**Figure 9.** Comparison of values by Cobit.

Figure 10 shows the surface plots between input variables of DM and scope complexity, respectively. For DM complexity, the following fifteen rules and normalized weights are included in the fuzzy rule system.

If (Understand is none) then (Decision-making is none) (0.6)
If (Understand is some) then (Decision-making is none) (0.6)
If (Understand is moderate) then (Decision-making some) (0.6)
If (Understand is large) then (Decision-making is moderate) (0.6)
If (Understand is maximum) then (Decision-making is large) (0.6)
If (Decide is none) then (Decision-making is none) (0.79)
If (Decide is some) then (Decision-making is some) (0.79)
If (Decide is moderate) then (Decision-making is moderate) (0.79)
If (Decide is large) then (Decision-making is moderate) (0.79)
If (Decide is maximum) then (Decision-making is large) (0.79)
If (Monitor is none) then (Decision-making is none) (1)
If (Monitor is some) then (Decision-making is some) (1)
If (Monitor is moderate) then (Decision-making is moderate) (1)
If (Monitor is large) then (Decision-making is large) (1)
If (Monitor is maximum) then (Decision-making is maximum) (1)



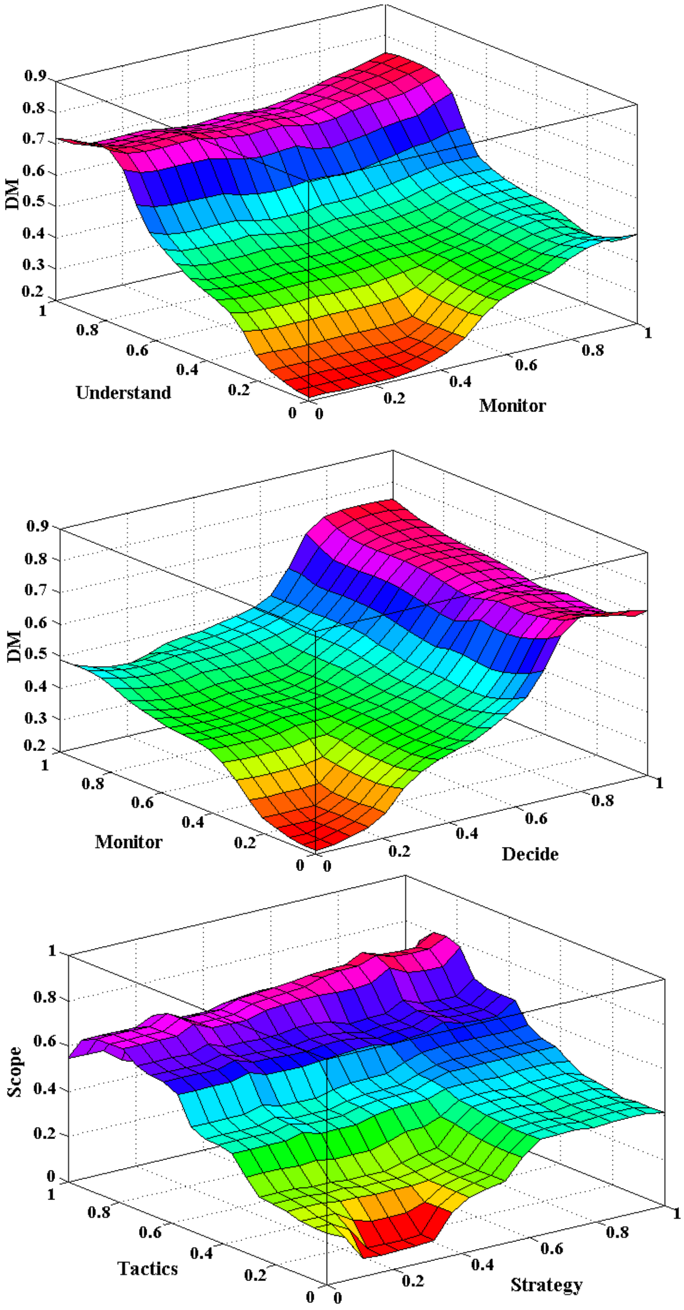


**Figure 10.** Mapping surface of DM and Scope complexity.

The theoretical concerns showed that the dimensional variable “Monitor” was more frequently used within the DM complexity. But, monitoring the implementation of decisions already made receives somewhat less attention from the practitioners, according to the survey. Also, comparing Cobit’s concerns of IT governance to literature, it showed that Cobit does support most needs, but lacks in providing information on how decision-making structures should be implemented. Applied to our fuzzy type-2 model, the dimension variables of DM complexity are almost uniformly stressed. The relative concerns for the DM complexity remain a bit more uncertain. The difference seems to lie in their interconnection weights (and interactions) between the concerns of IT governance. For scope complexity, strategic concerns are most often dealt with, while tactical concerns are only briefly discussed. The following ten rules and normalized weights are included in the fuzzy rule system.

If (Strategy is none) then (Scope is none) (1)
If (Strategy is some) then (Scope is some) (1)
If (Strategy is moderate) then (Scope is moderate) (1)
If (Strategy is large) then (Scope is large) (1)
If (Strategy is maximum) then (Scope is maximum) (1)
If (Tactics is none) then (Scope is none) (0.43)
If (Tactics is some) then (Scope is none) (0.43)
If (Tactics is moderate) then (Scope is some) (0.43)
If (Tactics is large) then (Scope is some) (0.43)
If (Tactics is maximum) then (Scope is moderate) (0.43)

IT governance mainly comprises strategic concerns according to literature. According to the practitioners responding the survey, IT governance decision making is mainly a strategy issue while tactical decisions are less important. Similarly, Cobit spends more effort in discussing strategic concerns and less on tactical concerns. But, according to the mapping surface of Figure 6, strategic and tactical concerns that make up a large collective behavior must be correlated and not independent.

## 5. Summary

This paper presented a fuzzy type-2 model to understand the relationship among IT governance concerns. The fuzzy type-2 logic presents a specific architecture for making judgments by computing with words. To assess the IT governance concerns is equal to make the judgments through computing with words. Compared with our previous work, results showed the more exact relations of the IT governance concerns than using fuzzy type-1 model. Similarly, the application results showed that the major differences exist within the concerns of the domain complexity in the case of Cobit. Really, Cobit was focused on decisions regarding the processes while people receive less attention. In conclusion, an analysis based upon these mathematical models suggested that IT governance itself is an organism capable of behaviors that are of greater complexity than those of an individual IT governance concerns. What makes something complex is that the concerns of IT governance is evolving. There is still room to further improve the performance of the proposed model. For example, currently in the model-building process we have made no attempt to tune the parameters to optimize the performance of the model. This is in part because there are only limited data sets available to us. In future studies, we should obtain data from other sites and automate the model calibration process. Research is ongoing to test other methods for constructing fuzzy rules and tie the use of a specific method with the characteristics of the IT governance concerns.

## References

[1] Cumps B, Viaene. S, Dedene. G, and Vandenbulcke. J, “An Empirical Study on Business/ICT Alignment in European Organizations.” Proceedings of the 39th Hawaii International Conference on System Sciences, 2006.

[2] Dahlberg T, Kivijärvi. H, “An Integrated Framework for IT Governance and the Development and Validation of

an Assessment Instrument." Proceedings of the 39th Hawaii International Conference on System Sciences, 2006.

[3] Debraceny R. S, "Re-engineering IT Internal Controls: Applying capability Maturity Models to the Evaluation of IT Controls", Proceedings of the 39th Hawaii International Conference on System Sciences, 2006.

[4] De Haes S, Van Grembergen. W, "IT Governance Structures, Processes and Relational Mechanisms – achieving IT/Business alignment in a major Belgian financial group." Proceedings of the 38th Hawaii International Conference on system Sciences, 2005.

[5] Guldentops E, "Governing Information Technology through COBIT." In Van Grembergen, W. (Ed.): Strategies for Information Technology Governance. Idea Group Publishing, 2004.

[6] Holm Larsen M, Kühn Pedersen. M, Viborg Andersen. K, "IT Governance – Reviewing 17 IT Governance Tools and Analysing the Case of Novozymes A/S." Proceedings of the 39th Hawaii International Conference on System Sciences, 2006.

[7] IT Governance Institute (ITGI), COBIT, 4th Edition, December 2005. Available online at http://www.isaca.org.

[8] Johansson E, Assessment of Enterprise Information Security – How to make it Credible and Efficient. Ph.D. Thesis at the Department of Industrial Information and Control Systems, Royal Institute of Technology, Stockholm, Sweden, 2005.

[9] Kaplan R, Norton. D, "The Balanced Scorecard. Harvard Business School Press, 1996 Office of Government Commerce (OGC)", IT Infrastructure Library Service Delivery. The Stationery Office, 2002.

[10] Liang Q, Mendel J. M, "Interval type-2 fuzzy logic systems: Theory and design." IEEE Trans. Fuzzy Systems, Vol. 8, No. 5, pp. 535–550, 2000.

[11] Cho S. E, Lee S. H, Moon K. I, "FUZZY DECISION MAKING OF IT GOVERNANCE ". International Conference on e-Business, Ice-b 2010, pp. 132-136, 2010.

[12] Mendel J. M, "An architecture for making judgments using computing words." Int. J. Appl. Math. Comput. Sci., Vol.12, No.3, 325–335, 2002.

[13] Mendel J. M, "Uncertainty, type-2 fuzzy sets, and footprints of uncertainty." In Proc. 9th Int. Conf. Information Processing and Management of Uncertainty in Knowledge-Based Systems, Annecy, France, pp. 325–331, 2002.

[14] Ribbers P. M. A., Peterson, R.R., and Parker, M.M., "Designing information technology governance processes: Diagnosing contemporary practices and competing theories." Proceedings of the 35th Hawaii International Conference on System Sciences, 2002.

[15] Remenyi D. A. H, Money, et al. "The effective measurement and management of IT costs and benefits". Computer weekly professional series. Oxford ; Boston, Butterworth-Heinemann. ISBN 0-7506-4420-6, 2000.

[16] Simonsson M, Ekstedt M, "Getting the Priorities Right - Literature versus Practice on IT Governance." Accepted for publication at Portland International Conference on Management of Engineering and Technology, Istanbul, July 9-13, 2006.

[17] Van Grembergen W, Saull R, De Haes S, "Linking the IT Balanced Scorecard to the Business Objectives at a Major Canadian Financial Group." In (Ed. Van Grembergen, W., Strategies for Information Technology Governance. Idea Group Publishing, 2004.

[18] Warland C, Ridley G, "Awareness of IT control frameworks in an Australian state government: A qualitative case study." Proceedings of the 38th Hawaii International Conference on System Sciences, 2005.

[19] Webb P, Pollard C, Ridley G, "Attempting to define IT Governance: Wisdom or Folly" Proceedings of the 39th Hawaii International Conference on system Sciences, 2006.

[20] Weill P, Ross J. W, "IT Governance: How Top Performers Manage IT Decision Rights for Superior Results", Harvard Business School Press, Boston, 2004.

## Author Biographies

**Sang Hyun LEE** He received the BS and MS in Department of Computer Engineering from Honam University. in 2002 and 2004, respectively. He received Ph.D. degrees in Computer Science from Chonnam National Univ. in 2009. His research interests include artificial intelligence, Software Engineering, Early Warning System, claim analysis.

**Kyung-li Moon** He received a Ph.D. Ph.D, is a professor at the Department of Computer Engineering, Honam University in Gwang-Ju, Korea. His theoretical work began at Seoul University as a statistical computing scientist, and then expanded into complexity science, chaos theory, and cognitive science –"generative" sciences.

# A Goal Programming Model for Solving Environmental Risk Production Planning Problem in Dairy Production System

Suresh Chand Sharma[1], Devendra Singh Hada[2], Sanjay Kumar Bansal[3] and Shilpa Bafna[4]

[1] Department of Mathematics, University of Rajasthan, Jaipur, Rajasthan, India.
[2] Department of Mathematics, Kautilya Institute of Technology and Engineering,
ISI – 16, RIICO Institutional Complex, Sitapura, Jaipur, Rajasthan, India.
[3] Department of Mathematics, Bansal School of Engg. & Technology,
Renwal Road, Sanganer, Jaipur, India.
[4] Department of Management, Kautilya Institute of Technology and Engineering,
ISI – 16, RIICO Institutional Complex, Sitapura, Jaipur, Rajasthan, India.
Suresh chand 26@gmail.com, dev.singh1978@yahoo.com,
Corresponding Authors bansalindian@gmail.com & kiteshilpa@gmail.com

## ABSTRACT

Environmental risk production planning and decision making is needed to analyze several alternatives in terms of multiple non commensurate criteria which involve conflicting preferences of different stakeholders.

In this paper, a goal programming model for tracking and tackling such environmental risk production planning problem that includes minimization of damages and wastes in the milk production system has been proposed. This model is explained by taking “SARAS” dairy production system in India. An interactive method which combines A.H.P. Strategy has been developed to solve this model.

**Keywords:** dairy, goal programming, production planning, A.H.P.

## INTRODUCTION

Management of dairy and related products is a complex environmental issue. The products can have both positive and negative environmental consequences which have substantial benefits and thus result in severe environmental degradation. Also the rapid growing population and economic development are supplementing to the effects on the environmental degradation.

Due to the uncontrolled urbanization and industrialization, Forest and Agricultural land degradation, Resource depletion (water, mineral, forest, sand, rocks etc.,), Environmental degradation, Public Health, Loss of Biodiversity, Loss of resilience in ecosystems, Livelihood Security for the Poor Ramesha Chandrappa and Ravi.D.R [2] have become the major environmental issues. Conjunction with the growing environmental resources depletion, human toxicity levels and ecosystem quality deterioration have made our governments and corporate sector more attentive towards the environmental damage and they are investing more in the assessment of environmental impact of their products and services to reduce such impacts.

Industrial eco-systems are the environmental friendly systems for industrial waste recycling, resembling the food chains, food webs and the nutrient recycles in natural environment Liu and Shyng, [7]. Because it transforms the harmful component of waste into usable substance and slows down the depletion of primary resources. The win-win solutions for business and the environment seem quite elusive in practice, in particular for considerable reductions on environmental pressure Walley and Whitehead, [14].

Consequently, the dairy industry is facing a potential impact of an individual operation on the environment which varies with animal concentration, weather, and numerous other conditions. Some questions which cropped up:

1. What is the trade-off between the environmental pressure of an economic activity and its costs?
2. How much we need to spend to reduce the wastage produced during the dairy processing?
3. What efforts required to be put in to reduce the environmental impact due to dairy functions and processing?
4. What are the “best” solutions balancing ecological and economic concerns? (Quariguasi Frota Neto [8], [9])

On the normative and qualitative field, these question have led to the concept of trade-offs and efficient frontiers for business and the environment (Huppes and Ishikawa [3]), Bloemhof-Ruwaard et al. [1]) the rationale is to determine the set of solutions towards the reduction wastage and increase environmental quality without much/ any increase in costs. In order to explore the efficient frontier in feasible time (for the intractability of determining all extreme efficient solutions in a multi objective linear program, see Steuer, R.E [13] and Steuer et al. [12].

In every production process, inputs are used to create finished product or commodity. Inevitably, some inputs are not fully used and are released into the environment in forms that may be considered pollutants. Similarly in a dairy production process also, some inputs which are not fully used or wasted during transport and processing are released into the environment in forms that may be considered as pollutants and which specifically are categorized as wastes. Thus, whenever the level of wastage and pollution exceeds the environmental ability to absorb and process dairy discharge, environmental risks develops.

In this paper, a goal programming model has been proposed for managing business environmental risks and wastage thus produced in a dairy products processing organization which consists of making the production process more efficient in such a way to limit its environmental consequences while increasing profitability irrespective of the fact that no production process is 100% efficient.

## 1. <u>DAIRY PRODUCTION SYSTEM</u>

The dairy production system has two major divisions that further leads to the division of the fragmentation of the dairy industry into two main production areas:

1. Primary sector

2. Processing sector
3. Distribution sector

The Primary sector involves into the following activities:

1. Milking of cows, goats, sheep, buffalos and camels.
2. Feeding these milk producing animals

The processing sector involves in the following activities:

1. Heat treatment of milk (to ensure that milk is safe for human usage and to elongate its preservation period)
2. Preparation of a range of dairy product for the consumption by human being. This includes:
   i. Semi-hydrated dairy products
   ii. Dehydrated dairy products

The distribution sector of the dairy industry includes transportation of milk to the collection centers for various treatments and quality tests and up-gradation. It also includes the distribution of the variety of milk products to the various retail outlets for disbursement to the consumers.

This paper is focused on the dairy production process, the products thus manufactured and the wastage then produced. The primary sector is not considered here in this paper as it is more related to agriculture sector. We do will take into consideration the processing and the distribution sector to identify the actual wastage areas and then will take out the measures to minimize that loss by the use of goal programming.

## 3. DAIRY INDUSTRY IN INDIA

**"INDIA IS THE WORLD'S HIGHEST MILK PRODUCER"**

Dairy is a place where handling of milk and milk products is done and technology refers to the application of scientific knowledge for practical purposes. Dairy technology has been defined as that branch of dairy science, which deals with the processing of milk and the manufacture of milk products on an industrial scale.

In India, dairying has been regarded as a rural cottage industry profession since the remote past. Semi-commercial dairying started with the establishment of military dairy farms and co-operative milk unions throughout the country towards the end of the nineteenth century. During the earlier years, each household in the country maintained its 'family cow' or secured milk from its neighbor who supplied those living close by. As the growth of urban population, fewer households could keep a cow for private use. The high cost of milk production, problems of sanitation etc., restricted the practice; and gradually the "family cow" in the city was eliminated and city cattle were all sent back to the rural areas. Gradually farmers within easy driving distance began delivering milk over regular routes in the cities. This was the beginning of the fluid milk-sheds which surround the large cities of today. Prior to the 1850's, most milk was necessarily produced within a short distance of the place of consumption because of lack of suitable means of transportation and refrigeration.

The Indian dairy industry has made rapid progress since independence. A large number of modern milk plants and product factories have since been established. These organized dairies have been successfully engaged in the routine commercial production of pasteurized bottled milk/ packed milk and various western and Indian dairy products. With modern knowledge of the protection of milk during transportation, it became possible to locate dairies where land was less expensive and crops could be grown more economically.

In India, the market milk technology may be considered to have commenced in 1950, with the functioning of the central dairy of array milk colony, and further the country stepped into milk product technology in 1956 with the establishment of Amul Dairy, Anand. To fulfil the national objective of making India self sufficient in milk production, a small step was taken in March, 1975 when Jaipur Zila Dugdh Utpadak Sahakari Sangh Ltd., Jaipur (popularly known as Jaipur dairy) was registered under cooperative act 1965 to work in Jaipur district. Initially this union did not have the processing facilities. It started with a modest beginning of procuring 250 liters of milk per day, which has increased manifolds with the passage of time.

### 3.1 <u>HISTORY OF INDIAN MARKET MILK INDUSTRY</u>

Beginning in organized milk handling was made in India with the establishment of military dairy farms. Handling of milk in co-operative milk unions established all over the country on a small scale in the early stages. Long distance refrigerated rail-transport of milk from Anand to Bombay since 1945

Pasteurization and bottling of milk on a large scale for organized distribution was started at Aarey (1950), Calcutta (Haringhata, 1959), Delhi (1959), Worli (1961), Madras (1963) etc.

Establishment of milk plants under the five-year plans for dairy development all over India. These were taken up with the dual object of increasing the national level of milk consumption and ensuing better returns to the primary milk producer.

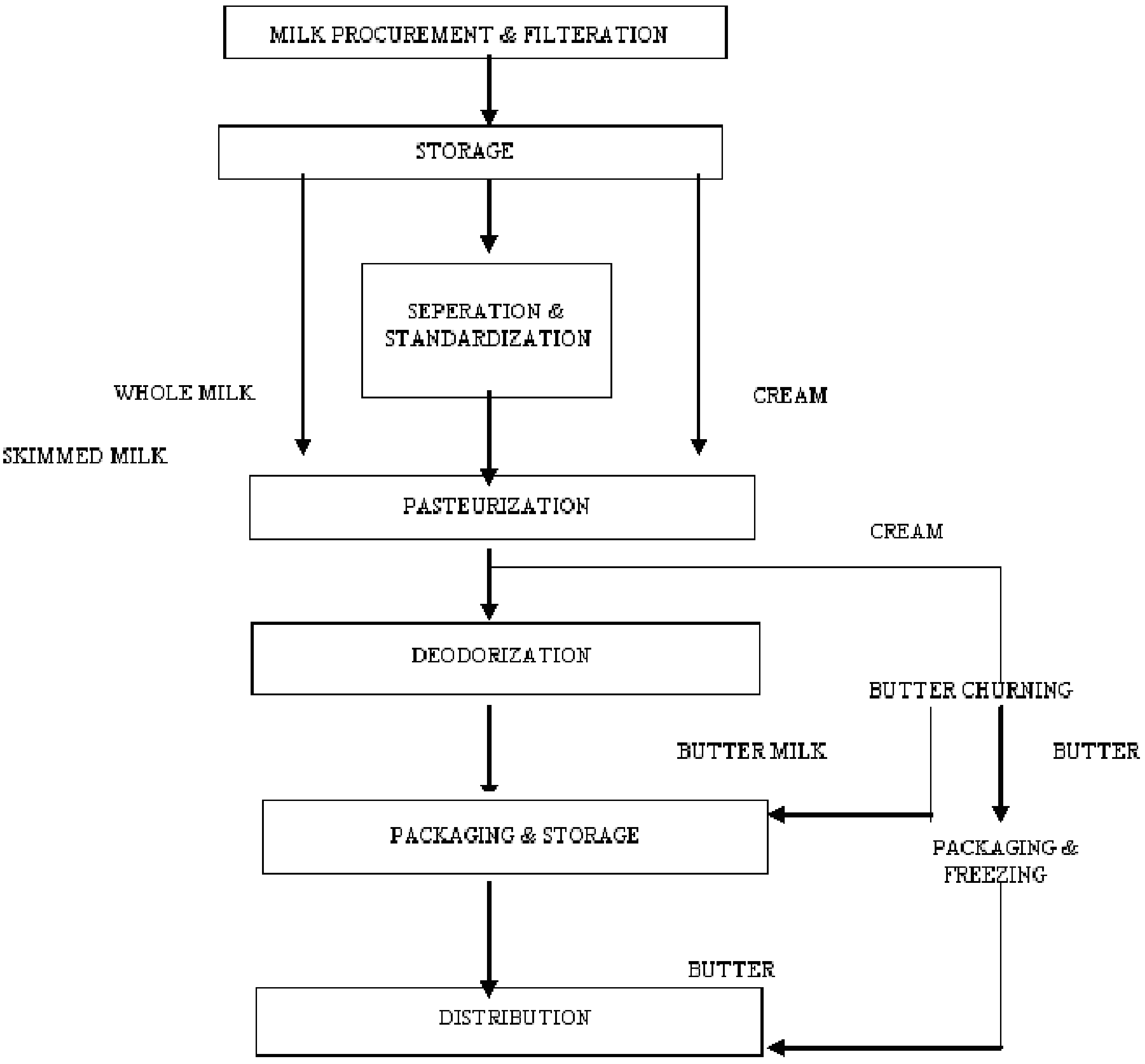
The main aim has been to produce more, better and cheaper milk which was possible only when the wastage was minimized and the available milk is packaged and distributed so that it could be made usable even after a considerably longer gestation period.

## 4. <u>PRODUCTION PROCESS OVERVIEW 'SARAS' DAIRY PRODUCTS</u>

This production process has been classified into following steps:

1. Milk procurement
2. Processing – this includes chilling, seperation & standardization, pasteurization, sterilization and deodorisation
3. Production of milk product range :
   i. Ghee
   ii. butter (salted / unsalted)
   iii. Skimmed milk powder(SMP)
   iv. Indigenous fresh milk products (paneer, shrikhand, chhach (plain / salted), lassi, mawa (khoa)& dahi (plain / mishti) and aseptic milk (which was handed over to Jaipur dairy only in 1997-98), ice-cream, shakes, etc.
4. Packaging
5. Storage (includes before and after storage both)
6. Distribution

**A MILK PROCESSING PLANT**



MILK PRODUCT RANGE
MILK - SKIMMED & UN-SKIMMED,
FLAVOURED & OTHERS
CREAM
BUTTER
BUTTERMILK
LASSI
CHEESE
GHEE
ICE CREAM
MILK POWER
ETC.

Similar plants have been developed for the processing of butter, cheese, ghee and the various other products of the SARAS dairy.

## 5. <u>WASTAGE ANALYSIS AT EACH LEVEL</u>

THE FOLLOWING WASTES TO BE DISPOSED:

In the associated milk processing factories, most of the waste is washing water that is treated, usually by composting, and returned to waterways. This is much different from half a century ago, when the main products were butter, cheese and casein, and the rest of the milk had to be disposed of as waste (sometimes as animal feed).

In many cases, modern farms have very large quantities of milk to be transported to a factory for processing. If anything goes wrong with the milking, transport or processing facilities it can be a major disaster trying to dispose of enormous quantities of milk. If a road tanker overturns on a road, the rescue crew is looking at accommodating the spill of 5 to 10 thousand gallons of milk (20 to 45 thousand litres) without allowing any into the waterways. A derailed rail tanker-train may involve 10 times that amount. Without refrigeration, milk is a fragile commodity, and it is very damaging to the environment in its raw state. A widespread electrical power blackout is another disaster for the dairy industry, because both milking and processing facilities are affected.

1. Water wastage- water is used for cleaning, cooling and maintaining hygine standards
2. Solid wastes- include plastic bags, bottles, packs, cartons, etc.

LEVEL WASTE ANALYSIS

1. At milk procurement level – poor drainage of tankers, spilling any leakage in pipes and tankers, foaming of milk, wastage in cleaning operations
2. At pasteurization and heat treatment level – leaking, deposits on the equipment, foaming, cleaning operations
3. At separation and filtration level – foaming, cleaning operation, spilling, cleaning operation, left over
4. At deodorization level – vacreation, cleaning operations
5. Miscellaneous wastes – start-up and shut-down process loss, damaged packaging, overfilling, plant malfunctioning, manhandling losses, bagging losses, incomplete separation of whey from curd, etc.

## 6. MODELING MULTI-OBJECTIVES OPTIMIZATION OF THE PRODUCTION

Quariguasi Frota Neto [8] was the first approach to define the theoretical frontier of Huppes And Ishkawa [3]. A cradle to grave approach is used to determine the eco-efficient frontier regarding business and the environment for the design of sustainable logistic networks. In this work, the diverse phases of a product: raw material, extraction, manufacturing, transportation, use and end-of-use alternatives are accounted determine the optimal solutions. In order to assess the trade-offs and determine the optimum configurations, multi-objective programming is used.

A multi objective programming is denoted by (Steuer et al. [12]):

$$\text{Min}\{c'X = z_1\}$$

$$\ldots.$$

$$\text{Min}\{c^k x = z_k\}$$

$$\text{s.t. } \{x \in R^n \mid AX \leq b, b \in R^m, x \geq 0\}$$

where,

k is number of objectives

a point $x \in SR^n$ is efficient iff there is no $x \in S$ such that $c^i \geq x^i$ and there is atleast one $c^i x < x^i$

The efficient set or efficient frontier is the set of all efficient solutions.

In our formulation, $c^1x$ represents total cost of a certain economic configuration, $c^2x$ the cumulative impact to the environment, $c^3x$ the respective wastage.

The economical objective function is the sum of the cost of the following activities:

1. minimized wastage
2. waste processing
3. reuse of the industrial wastage
4. production based on clean technology

The constraints towards attainment of the functional objectives are:

1. capital required
2. labor handling
3. flow conditions
4. storage conditions
5. time lag during transportation
6. quality of raw material

## 6. DESCRIPTIONS

1. GOAL PROGRAMMING – The goal programming technique is an analytical framework that a decision-maker can use to provide optimal solutions to multiple and conflicting objectives. The GP and its variants have been applied to wide-ranging problems ( Ignizio, [4]; Ijiri, [5]; Lee, [6]; Romero, [10] ). it is a branch of multi-objective optimization, which in turn is a branch of Multi-Criteria Decision Analysis (MCDA), also known as Multiple-Criteria Decision Making (MCDM). This is an optimization programme. It can be thought of as an extension or generalization of linear programming to handle multiple, normally conflicting objective measures. Each of these measures is given a goal or target value to be achieved. Unwanted deviations from this set of target values are then minimized in an achievement function. This can be a vector or a weighted sum dependent on the goal programming variant used. As satisfaction of the target is deemed to satisfy the decision maker(s), an underlying satisfying philosophy is assumed.
2. A.H.P. - ANALYTIC HIERARCHY PROCESS (AHP), as one of multi-attribute decision making (MADM), is a structured technique for dealing with complex decisions. AHP provides the decision maker an approach to find results that best suits their requirement and their understanding of the problem, i.e. rather than prescribing a "correct" decision, the AHP helps the decision makers to find the decision most suitable to him. A.H.P was proposed by Saaty [11] 20 year ago and therefore referred as ‘Saaty’ method. It is a widely used technique for MADM, which is based upon pair wise subjective judgment of element used to complete a matrix. The Eigen value for each element is then used to asses the contribution of that element to the overall component.

In the context of this paper, a dairy production system on the basis of several criteria such as cost and wastage can be taken as a typical example. We would need to determine the relative contribution of cost and wastage to the overall decision and also the relative degree to which dairy manufacturer processes each criterion.

Assume that there are n elements, then we require (n (n-1))/2 pair-wise judgments to complete the matrix, where each judgments reflects the perception of the ratio of the relative contributions of elements i & j to the overall components be assessed so $a_{ij} = ( w_i / w_j )$ , subject to the following constraint:

$a_{ij} > 0$ , $a_{ij} = 1$ when i = j elsewhere $a_{ji} = (1 / a_{ij})$.

Saaty argues that each of these judgments assign a number on a scale. A basic and reasonable assumption is that if attribute A is absolutely more important than attribute B and is rated at 9, then B must be absolutely less important than A and is valued at 1/9. The technique can only be effectively used where the elements are homogenous, that is with in the same order of magnitude, and hence the ratio must range from 1/9 to 9.

Through the literature review, we analyzed that Some researchers attach semantic labels such as "equal" where ratio is 1 ,"Slightly more important " where it is 2 and so forth for instance, if we considered wastage to be "Slightly more important " than cost , one would assign the value to the appropriate cell in the matrix. In this case matrix would be completed as follows:

| | Wastage | Cost |
|---|---|---|
| Wastage | 1 | 2 |
| Cost | 0.5 | 1 |

Each component has a priority scale that is derived ratio scale, to measure the contribution of each element to that component. This is based upon the approximate Eigen value (i.e. divide the sum of the row by n) of each element.

One problem that can occur, especially since the judgments are subjective is that the values assigned are inconsistent. For example, one would expect to observe transitivity. Consistency can be measured as a deviation of the principle Eigen value of the matrix from the order of the matrix.

The consistency index, CI, is calculated as follows.

$$\text{CI} = (\lambda_{\max} - n)/(n-1)$$

Where $\lambda_{\max}$ is the maximum principle Eigen value of the judgment matrix. The nearer CI is to zero the more consistent the judgments. The CI can be compared with the consistency index of a random matrix (RI). The ratio CI/RI is known as the consistency ratio (CR) Saaty suggest CR should be less than 0.1, although one should be cautious about attaching undue significance to this value.

GP Model as follows :

**Minimize :** $\sum^{4}_{k=1} P_k(\eta_k + \rho_k)$

**Subject to :**

$\sum^{n}_{j=1} C_{j\ \text{wastage}} X_j - \eta_1 + \rho_1 = G_W$ **}wastage target (P1)**

$\sum^{n}_{j=1} C_{j\ \text{waste processing}} X_j - \eta_2 + \rho_2 = G_{WP}$ **}waste processing target (P2)**

$\sum^{n}_{j=1} C_{j\ \text{reuse of the industrial wastage}} X_j - \eta_3 + \rho_3 = G_{RW}$ **}reuse of the industrial wastage target (P3)**

$\sum^{n}_{j=1} C_{j\ \text{production based on clean technology}} X_j - \eta_4 + \rho_4 = G_{CT}$ **} clean technology target (P4)**

$\sum^{n}_{j=1} C_j X_j \leq B_{cr}$ **}capital required constraint**

**$\sum^{n}_{j=1} C_j X_j \le B_{lh}$ }labor handling constraint**

**$\sum^{n}_{j=1} C_j X_j \le B_{fc}$ }flow conditions constraint**

**$\sum^{n}_{j=1} C_j X_j \le B_{sc}$ }storage conditions constraint**

**$\sum^{n}_{j=1} C_{ij} X_j \le B_{tl}$ } time lag during transportation constraint**

**$\sum^{n}_{j=1} C_{ij} X_j \le B_{qm}$ } quality of raw material constraint**

**$\eta_i, \rho_i, X_j \ge 0$ } non–negativity constraint**

**$\eta_i * \rho_i = 0$ } complementary constraint**

**$i = 1,2,3,\ldots\ldots..8$ , $j = 1,2,3,\ldots\ldots..n$**

We use Analytic Hierarchy Process to determine the level of priority. A hierarchy of importance among goals is established by assigning to each of them a pre-emptive priority factor, Pj These pre-emptive priority factors reflect the hierarchical relationships in such a way that P1 represents the highest priority, P2 the second highest, and P3 third highest, P4 forth highest priority. A positive deviational variable ($\eta_i$) represents overachievement of the goal. A negative deviational variable ($\rho_i$) represents underachievement of the goal. If the desire is not to underachieve the goal, d should be driven to zero. To the contrary, if d is driven to zero, the overachievement of the goal will not be realized. Deviational variables are mutually exclusive.

Where X1,X2,…….,Xn represent decision variable, (Cij ) represent the contribution coefficientof each decision variable. GW ,GWP ,GRW ,GCT represent the goals for the minimized wastage ,waste processing , reuse of the industrial wastage , production based on clean technology respectively.

## 7. RECOMMENDATIONS

Feeding to increase productive life by reducing culling rates, improving herd health status, maintaining fertility, reducing mastitis and somatic cell count, and increasing milk production are possible goals on dairy farms. Dairy managers, veterinarians, and nutritionists can review the following outline of key points and nutrient guideline table for phase feeding.

1. Monitoring dry matter intake
2. Optimizing rumen fermentation
3. Strategies with transition feeding program
4. Balancing and meeting nutrient requirements
5. Benchmarking cow performance

## 8. CONCLUSION

The purpose of the paper was to suggest another method for the minimization of the wastage during the dairy products' processing in the SARAS dairy. We suggest the use and the application of goal programming method for the reduction of the wastage of the milk and its products or by-products. The goal programming model for managing the wastage during the dairy processing consists of making the production process more efficient in such a way as to limit the wastes in the processing.

## 9. References

**URL's**

http://www.milkacademy.com

http://www.indiaagronet.com/indiaagronet/DAIRY/Dairy.htm

http://www.jaipurdairy.com/

[1] Bloemhof-Ruwaard, J.M, Krikk, H., and Van Wassenhove L.N OR models for eco-eco closed-loop supply chains optimization, volume 1 of Reverse Logistics: Quantitative Models for Closed-Loop Supply Chain. Springer-Verlag, Berlin / Heiderberg, 1 edition, 2004.

[2] Environmental Issues, Law and Technology - An Indian Perspective, Ramesha Chandrappa and Ravi.D.R, Research India Publication, Delhi, 2009

[3] Huppes , G and M. Ishikawa. A framework for quantified eco-efficiency analysis, Journal of Industrial Ecology, 9(4), 2005, 41.

[4] Ignizio, J.P., Goal Programming and Extensions, Massachusetts, Lexington Books, 1976.

[5] Ijiri, Y., Management Goals and Accounting for Control ,Amsterdam, North-Holland, 1965.

[6] Lee,S.M. ,Goal Programming for Decision Analysis ,Philadelphia, Auerbach, 1972.

[7] Liu, K and Shyng , Q., Eco-system in the steel industry . Published in Proceeding of International Conference on Cleaner Production and Sustainable Develoment 99. Decemder 1999. Taipei, Taiwan,1999.

[8] Quariguasi Frota Neto, J. Alternative targets for data envelomet analysis through multiobjective linear programming: Rio de janerio odontological public health system case study . Journal of Operations Research Society , 2005.

[9] Quariguasi Frota Neto, J, J. Bloemhof, J.A.A.E. van Nunen, and E. van Heck. Designing and evaluating sustainable logistics networks. International Journal of Production Economics ( in print ), 2007.

[10] Romero, C. ,Handbook of Critical Issues in Goal Programming Oxford, Pergamon Press, 1991.

[11] Saaty, T.L., The Analytic Hierarchy Process . McGraw-Hill, New York,1980.

[12] Steuer,R.E and C.A. Piercy. A regression study of the number of efficient extreme points in multiple objective linear programming. European Journal of Operational Research, 162 (2) 2005, 484-496.

[13] Steuer, R.E. Random problem generation and the computation of efficient extreme points in multiple objective linear programming. Computational Optimization and Applications, 3(1994), 333-347 .

[14] Walley, N and B. Whitehead. Its not easy being green. Harvard Business Review , 72(3) 1994, 46.

**AUTHOR BIOGRAPHIES**:

**DEVENDRA SINGH HADA**

Born in Kota (Rajasthan, INDIA) on 8th October 1978 and brought up in Kota. Pursued B.Sc. from University of Rajasthan in 1998 followed by M.Sc. in year 2000. The year 2006 is marked for pursuing M.Phil. from Alagappa University, which is Merit rated by the National Council for evaluation and accreditation (NAAC) in Karaikudi in Tamil Nadu, Tiruchirappalli. At this time, research work on the topic “Study of interdisciplinary areas using Goal Programming” is being pursued under the guidance of Prof. Suresh Chand Sharma, affiliated in University of Rajasthan in Department of Mathematics. Since 2005, have been working as Associate Professor of Mathematics at Kautilya Institute of Technology and Engineering, ISI – 16, RIICO Institutional Complex, Sitapura, Jaipur, Rajasthan, India. Email : dev.singh1978@yahoo.com

**SANJAY KUMAR BANSAL**

Born in Sikar (Rajasthan, INDIA) on 8th June 1978. Pursued B.Sc. from University of Rajasthan in 1998 followed by M.Sc. in year 2000. The year 2002 is achived his M.Phil. from University of Rajasthan, which is Merit rated by the National Council for evaluation and accreditation (NAAC). At this time, research work on the topic " A Study of generalized Mellin-Barnes Contour integrals and Geometric Function Theory" pursued under the guidance of Prof. S. P. Goyal, affiliated in University of Rajasthan in Department of Mathematics. Since 2008, have been working as Associate Professor of Mathematics at Bansal School of Engineering & Technology, Renwal More, Sanganer, Jaipur, Rajasthan, India. He has having teaching experience of more than 12 years at graduate & undergraduate level. He has 8 research papers published in journals of international repute. Email: bansalindian@gmail.com

**SHILPA BAFNA**

Born in Jodhpur (Rajasthan, INDIA) on 29th December 1983 and brought up in Jaipur, where pursued B.Com. with average 70% in the year 2005 along with diploma courses in Computer. Subsequently, completed PGDBM course from Symbiosis, Pune in 2007. Since June 2007, have been working as Asst. Professor of Finance at Kautilya Institute Of Technology And Engineering And School Of Management, ISI – 16, RIICO Institutional Complex, Sitapura, Jaipur, Rajasthan, India. Email: kiteshilpa@gmail.com

# Prediction of Hypo/Hyperglycemia through System Identification, Modeling and Regularization of Ill- Posed Data

S.Shanthi[1],Dr.D.Kumar[2],Prof.S.Varatharaj[3],S.Santhana Selvi[4]

[1]. Asst.Prof., Dept. of ECE, JJCET, doing part time PhD at Anna University- Tiruchirappalli, Tamilnad, India.

[2]. Dean-Research,Periyaar Maniammai University,Vallam,Tanjore, Tamilnad, India.

[3].Prof & Head, Dept. of Mathematics,JJCET, Tiruchirappalli, Tamilnad, India.

[4]. JJCET, Tiruchirappalli, Tamilnad, India.

{ sshanthi289@hotmail.com,kumar_durai@yahoo.com,varaj48@gmail.com,sanselvi@gmail.com}

***Abstract:*** A clinically important task in Diabetes management is prevention of Hypoglycemic events. Continuous Glucose Monitoring (CGM) devices have been used to find the trend and temporal variability of the glucose levels of a Diabetic person. This CGM data can be used to identify the impending Hypo/Hyperglycemia well in advance with preprocessing the CGM data and System identification and Modelling. In this paper we have tried with ARIMA model and Tikhonov regularization. The prediction process is done with 10,20 and 30 minutes ahead time slots and their RMSE have been calculated. Results show that the preprocessed data with proper system modeling give accurate prediction values with clinically acceptable time lags.

**Key words :** Diabetes, Hypo glycemic Alerts, Ill posed problem, System identification, Mathematical modeling, Regularization.

## 1. Introduction

Diabetes Mellitus is a Metabolic disorder characterized by the inability of the Pancreas to regulate blood glucose concentration. High blood glucose levels lead to chronic diseases such as Cardio vascular , Retinal, Renal and Nervous disorders. Low blood glucose levels lead to immediate effects like Seizures and Short term Coma. According to DCCT (Diabetic Complications and Trial)[1] the risks of Diabetes can be prevented by proper blood glucose monitoring and regulation. Conventional method is to use blood Glucometers which use the capillary blood obtained through finger prick. This method is associated with pain and inconvenience and gives only the instantaneous values. Whereas Continuous Glucose Monitoring represents a significant advancement in the technology because it provides real time information about the current blood glucose. Eventhough the CGM devices measure the Interstitial fluid glucocose, they provide information about magnitude, direction, duration and frequency of fluctuations in the glucose levels. The device can give an alert at instances of unacceptable high or low glucose levels. Instead of giving alerts at that instances of blood glucose excursions, it would be advantageous if the alerts are given in advance so that the regulation of blood glucose can be done in a proactive manner. Efficient generation of Hypoglycemic alert in advance is of much importance due to the dangerous nocturnal hypoglycemia.

## 2. Predictive monitoring

Many researchers are contributing their work in this area. Since the CGM time series is of ill posed in nature, the systems are said to be ill conditioned. The numerical treatment of ill conditioned linear system is more complicated. Therefore much effort is required in system identification and regularization. System identification is the art of building mathematical models of dynamic systems from observed input – output data. It can be as the interface between the real world of applications and mathematical world of control theory and model abstractions. A model gives a relationship between observed quantities. Model allows for prediction of properties or behaviors of the object. Data driven models represent a class of modeling techniques where the relationships between input and output process variables that characterize the underlying phenomenon being modeled are learned during the training phase. Then that model can be used for prediction of future values.

The first question on prediction was raised by Bremer and Gough[2] whether the CGM data could be used for prediction of near future glucose levels. According to them if the recent blood glucose history is not random but has an exploitable structure, it might be possible to predict the near future blood glucose values based on previous values. They used 10- min data from ambulatory Type I Diabetes Mellitus patients and identified Autoregressive models. They explored 10-min,20-min,ans 30-min prediction horizons and report that the 10- min ahead predictions are accurate. But no quantification for their predictions were given. Bellazzi et al., [3] used non uniformly and sparsely sampled T1DM (Type I Diabetes Mellitus – onset of diabetes before the age of 25) subject data collected in ambulatory conditions, linearly interpolated at 2 hour intervals, to identify low order ARX models whose inputs included meals and a filtered insulin input. They investigated with 2-h, 4-h and 6-h prediction horizons. The prediction metrics were summarized with 1 step ahead prediction for best case subject, worst case subject and mean case of 60 subject data bank. The marked difference between the results for the best case and worst case subjects illustrates a fundamental and significant inter-subject variability. But the results were somewhat positively based due to the linear interpolation. Hovorka et al., [4] performed experiments in 10 T1DM patients under clinical conditions, using their own physiological model to make predictions of 15 minutes glucose data upto 4 steps (i,e, 60 minutes) into the future. The glucose was measured intravenously, but delayed by 30 minutes to mimic subcutaneous measurement. The model parameters were recursively estimated using a sophisticated Bayesian method. The predictions of the resulting models had RMSE values of 8.6, 13.0 and 17.3 mg/dL for 2 step, 3 step and 4 step predictions respectively. Trajanosoki et al.,[5] proposed a neural predictive controller for closed loop control of glucose in subcutaneous route. The control strategy is based on off line system identification using neural networks and non linear model predictive controller design. The proposed framework combines the concept of Non linear Auto Regressive with eXogenous inputs model using a regularization approach for constructing Radial Basis Function Neural Network. The drawback of this method is that the training of neural network requires the solution of a non convex optimization and the resulting network weights or lack of model coefficient. Dua et al., [6] employ a Kalman filter to adjust the parameters of first principles model for the prediction and control of blood glucose. The performance was tested with simulated data. The Kalman filter had different implementation challenges. They require the availability of a high fidelity first principle model capable of accounting for meals and physical activity. Robert S.Parker, Doyle III [7] and their group worked on model based predictive control algorithm which was developed to maintain normoglycemia in Type 1 Diabetes patients using a closed loop Insulin infusion pump. Compartmental modeling technique was used in this work. A 19'th order non linear Pharmaco kinetic – Pharmaco dynamic representation was used in controller synthesis. Linear identification of an i/p – o/p model from noisy patient data was performed by filtering the impulse response coefficient via projection onto Laguerre basis. Palerm et al.,[9**]** have demonstrated the effect of of sampling frequency, threshold selection and prediction horizon on the sensitivity and specificity of prediction of hypoglycemia. In their view, an optimal estimator could be structured to estimate not only the value of interest (i.e. glucose concentration ) but also its rate of change. They extended this to estimate the rate of change of rate of change( second derivative ) to improve prediction particularly for longer prediction horizons. The same group in their earlier work [10], proposed an algorithm based on the real time glucose sensor signals and optimal estimation theory (Kalman filtering) to predict hypoglycemia. The algorithm was validated in simulation based studies. In this current work, they further refined and validated the prediction algorithm based on the analysis of clinical hypoglycemia clamp data from 13 subjects. The result of this work was that for a 30 minute prediction horizon and alarm threshold of 70 mg/dl, the sensitivity and specificity were 90 and 79% respectively. Indicating that a 21% flase alarm rate must be tolerated to predict 90% of hypoglycemic events 30 minutes ahead of time. Shorter prediction horizons yield a significant improvement in sensitivity and specificity. Palerm et al., had two challenges in when testing a real time BG prediction algorithm with clinical data. First is the necesssacity of having frequently sampled reference BG values for comparison and next is the need to separate performance of sensor from that of prediction algorithm. For their study they used the hypoglycemic clamp data from CGMS® of Medtronic Minimed. Sparacino et al. , [11] used two prediction strategies based on the description of past glucose data. One is the first order polynomial and the other is the first order Auto Regressive model. Both the methods have time varying parameters estimated by Weighted Least Squares. In both the methods, at each sampling time, a new set of model parameters is first identified by means of WLS technique. Then the model is used to forecast glucose level for a given prediction horizon. The prediction algorithm was tested with Glucoday CGM system data from 28 type1 diabetic patients for a duration of 48 hours collected at a frequency of 3 minutes. Mean Square Error and Energy of Second Order Differences (ESOD) were taken as the performance metrics. Results proved that the performance of prediction algorithm is adequate for prediction of glucose from past data is feasible for preventing hypo/ hyperglycemic events. The importance of using a time varying approach was witnessed in this work. Reifman et al., [12**]** investigated the capabilities of data driven AR models to Capture the correlations in glucose time series data, make accurate predictions as a function of prediction horizon and be made portable from individual to individual without any need for model tuning. They had made investigation with CGM data of 9 Type 1 diabetic subjects in a continuous 5 day period. The predicted glucose values were analyzed with Clarke's Error Grid. The study shows that, for a 30 minute prediction horizon data driven AR models provide sufficiently accurate estimates of glucose levels, for timely proactive therapy and AR model can be considered as the modeling engine for predictive monitoring of patients with Diabetes Mellitus. It also suggests that AR models can be made portable with minor performance penalties which greatly reduces the burden associated with model tuning and data collection for model development. Finan et al., [13] obtained data set of 2 Type I Diabetes subjects for a period of 5 days with values taken in 5 minutes span. They also generated simulated data for reality check from a non linear physiological model of TIDM. Each data set was divided into 2 halves. First half used for calibration i.e,model identification  and second half used for validation. They identified 3 types of dynamic models - AR, ARX and ARMAX. Model identification procedure is by estimating model parameters such that one step prediction error are minimized. FIT value is the metric used to quantify the accuracy of model predictions. It is the measure of how much variability in the data has been explained by the model predictions. RMSE can also be used. Predictions deteriorate as the validation prediction horizon is extended to 24 steps (120 minutes). ARX model has increased complexity than  AR model. For simulated data, the best modeling results were achieved with ARMAX model. **Cobelli group ( Sparacino et al., )**[11] had also suggested the use of CGM and AR models for short term glucose level predictionsof Type 1 diabetic patients. Although they found AR models to provide adequate results for 30 minute ahead predictions. But their modeling formulation is significantly different. They found that the models with order larger than one and with fixed parameters to be unstable and yield unacceptable prediction delays. Their AR model of order m=1 is updated continuously ( for each individual ) as each new observation becomes available and to avoid model " over fit " the parameter update balances the weight amoung current and prior observations. This is in contrast with the Reifman's group where an AR model is developed once for individual and same model is applied to other individuals without any modifications. A.Gani et al.,[14] combined the predictive data driven models and the frequent blood glucose measurements to provide an early warning of the impending glucose excursions and proactive regulatory actions. By simulation they proved that stable and accurate models for near future glycemic predictions with clinically acceptable time lags obtained by smoothing the raw glucose data and regularizing the model coefficients. This has to be validated for real time implementation. This group has worked with AR model of higher orders. C.Perez-Gandia et al., [15] have implemented an artificial neural network algorithm for online glucose prediction from continuous glucose monitoring. The predictor is implemented with artificial neural network model (NNM). In all these approaches the large time lags reduce the clinical benefits of predictions.

In this paper we propose an Auto Regressive Integrated Moving Average model (ARIMA) for the prediction of near future glucose concentration. We compared our results with 1 step, 2 step and 3 step time ahead (i.e., 10-min, 20-min and 30-min) predictions and their corresponding relative absolute differences and the RMSE have been analysed.

## 3. Methodology

A stochastic model that is extremely useful in the representation of certain practically occurring series is the *Auto Regressive* model[16][17]. In this model, the current value of the process is expressed as a linear aggregate of previous values of the process. Another kind of model is the *Moving Average* model which depends on the previous deviations. To achieve greater flexibility in fitting of actual time series, it is advantageous to include both *AutoRegressive and Moving Average* terms in the model. Many time series data obtained practically are of non stationary in nature. ARIMA models are the most general class of models for forecasting a time series which can be stationarized by transformations such as differencing and logging. ARIMA models are fine tuned versions of random walk and random trend models. The fine tuning consists of adding lags of the differenced series and/or lags of the forecast errors to the prediction equation. The first step in fitting an ARIMA model is the determination of the order of differencing needed to stationarize the series. The optimal order of differencing is often the differencing at which the standard deviation is minimum.

### 3.1 Regularization

One important property of mathematical problems is the stability of their solutions to small changes in the initial data. Problems that fail to satisfy this stability condition are said to be *ill posed*. The main objective of *regularization* is to incorporate more information about the desired solution in order to stabilize the ill posed problem and to find an useful solution. The additional information is usually in the form of a penalty for complexity such as restrictions for smoothness or bounds on the vector space norm.The most common and well known form of regularization is that of Tikhonov.[18]

Noise in the raw data should be removed so that the estimated coefficients would reflect the underlying physiologic dependency. Smoothing of raw data removes the high frequency noise[19]. A linear smoother estimates the function value

$$\hat{Y}_j = x(t_j) \qquad (1)$$

by a linear combination of the discrete observations

$$x(t_j) = \sum_{l=1}^{n} S_j(t_l)\, y_l \qquad (2)$$

where $S_j(t_l)$ weights the $l$'th discrete data values in order to generate the fit to $y_j$. In matrix terms,

$$x(t\ \ ) = \ Sy \qquad (3)$$

Where x(t) is a column vector containing the values of the function 'x' at each sampling point '$t_j$'.

$$S = S_d = \Phi\ (\Phi'\Phi)^{-1}\ \Phi'. \qquad (4)$$

In the context of least squares estimation, the smoothing matrix has the property of being a projection matrix. This means that it creates an image of data vector 'y' on the space spanned by the columns of matrix 'Φ'such that the residual vector

$$e = y – \check{y} \qquad (5)$$

is orthogonal to the fit vector ў. The key idea in Tikhonov's method is to incorporate a priori assumptions about the size and smoothness of the desired solution in the form of smoothing function. The smoothed signal is given by

$$\check{y} = S_d\ \ W \qquad (6)$$

where $S_d$ is the integral operator and W denotes the estimates of glucose signals first derivatives. The derivatives' estimate yield excellent data smoothing and do not introduce lag on the smoothed signal relative to the raw signal. To estimate the signal's derivative W, the functional f(W) is minimized**.** [18]

$$f(W) = \|\ Y\text{-}S_d\ \|^{2} + \lambda_d^{2}\ \|\ L_dW\ \|^{2} \qquad (7)$$

where 'Y' is the N*1 vector of raw CGM data and '$S_d$ ' is the N*N integral operator, 'W' represents the rate of change of glucose with time. $\lambda_d$ is the regularization parameter and $L_d$ denotes a well conditioned matrix chosen to impose smoothness on 'W'. $L_d$ is typically either an identity matrix, a diagonal weighting matrix or a p*n discrete approximation of a derivative operator in which case L is a banded matrix with full row rank.

$$L_2 = \begin{pmatrix} 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \end{pmatrix} \qquad (8)$$

$$L_d = L_2 * (\text{second derivative } N \times 1 \text{ matrix}) \qquad (9)$$

The regularization parameter λ controls the weight given to minimization of the regularization term relative to the minimization of the residual norm. The most convenient graphical tool for analysis of the discrete ill posed problems is the so called L-curve which is a plot for all regularization parameters of the discrete smoothing norm. The L-curve clearly displays the compromise between minimization of these two quantities, which is the heart of any regularization method.

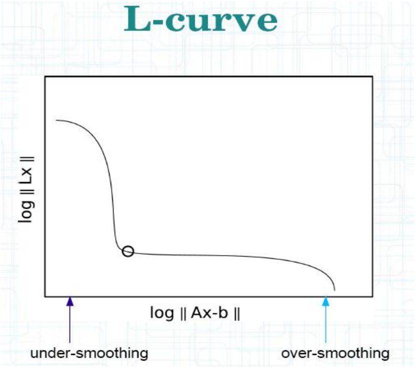


**Fig 1.**Graph used for selection of Regularization parameter.

The selection of the regularization parameter can either by Pragmatic parameter choice method or by Discrepancy principles or by methods based on error estimation. In our work, the optimum value of λ is obtained by minimizing the RMSE between the smoothed signal and the predicted signal.

## 3.2 ARIMA Modelling

After stationarizing the data by preprocessing i.e, through regularization, the next step is to fitting in an ARIMA model. The more systematic way to do this through Auto correlation and Partial Auto correlation plots of the regularized data. ACF plot is merely a bar chart of the coefficients of correlation between the time series and lags of itself. PACF plot is a plot of partial correlation coefficient between the series and lags of itself. The terms corresponding to exponential decline in ACF and peak in PACF would contribute to AR processes and Peak in ACF and exponential decline in PACF would contribute for MA processes. The next step is to determine the coefficients of model parameters by Maximum likelihood estimation. A conditional likelyhood function is selected in order to get good starting point to ontain an exact likelihood function.

Then the diagnosis check is carried out to validate the model. In successive trials the observation of the residuals obtained can help to refine the structure of the functions in the model[20][21]. An ARIMA model is generally given by

$$\Phi(B)\, g(t) = \theta(B)\varepsilon(t) \qquad (10)$$

Where g(t) is the glucose level at time 't', Φ(B) and θ(B) are the parameters of AR and MA processes involved and ε(t) is the error term. Φ(B) and θ(B) are functions of backward shift operator i.e.,

$$B^l\, g_t = g_{t-l} \qquad (11)$$

$$\Phi(B) = 1 - \textstyle\sum_{l=1}^{\Phi} \Phi l B l \qquad (12)$$

$$\theta(B) = 1 - \textstyle\sum_{l=1}^{\theta} \theta l B l \qquad (13)$$

The Third order ARIMA model has been selected first through empirical approach and then confirmed with optimization. The prediction efficiency of the model has been validated initially with simulated data and then with five real life subjects' data who were using the Minimed MedtronicCGM device.

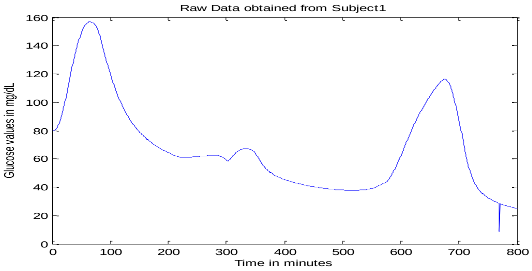


**Fig. 2** Fluctuation in Glucose Profile of a Diabetic Subject

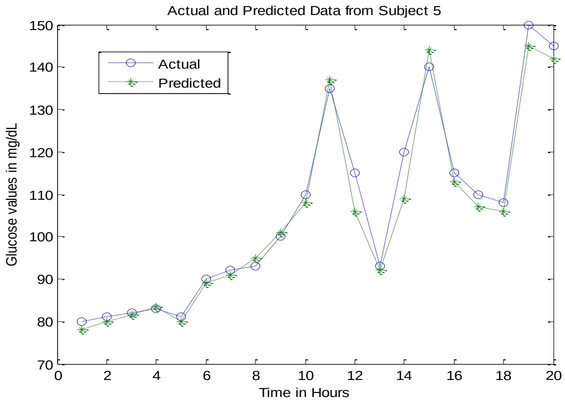


**Fig .3**. Actual and Predicted glucose profiles

First half of the data is used for training and the second half data is used for validation. RMSE between the predicted glucose levels and the actual value have been studied under various prediction horizons such as 10min,20-min and 30-minutes ahead. It is observed that the short term predictions provide accurate results. However the RMSE obtained for the 2 step and 3 step prediction horizons are much reduced compared to earlier approaches.

## Results :

Performance Metric as Root Mean Square Error

| Subject | RMSE in mg/dL for Prediction Horizons of | | |
|---|---|---|---|
| | 10-min | 20-min | 30-min |
| 1 | 0.5 | 1.2 | 2.4 |
| 2 | 0.6 | 1.4 | 3.1 |
| 3 | 0.9 | 2.7 | 4.2 |
| 4 | 0.4 | 1.1 | 2.1 |
| 5 | 0.5 | 2.3 | 3.1 |

## 4. Conclusion

This paper proposes that a lower order ARIMA model could used for the prediction of near future blood glucose concentration so that the impending dangerous hypo/hyper glycemia can be inferred well in advance and preventive actions can be taken. The methodology is validated with simulated data as well as with real patient data. Avoiding false alarms and improvement in accuracy atleast by a factor of 5% is a great thing in these works. CGM devices with prediction capability will be much helpful in improving the quality of life of Diabetic society.

## 5. References

[1] The Diabetes Control and Complications Trial Research Group: The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin dependent diabetes mellitus. *N Engl J Med* 329 : 977 – 986 , 1993.

[2] R. Bellazzi, C. Siviero, M. Stefanelli, and G. De Nicolao, Adaptive Controllers for Intelligent Monitoring, *Artif. Intell. Med.*, vol. 7, 1995, pp. 515–540.

[3] T. Bremer and D.A. Gough, Is Blood Glucose Predictable from Previous Values?, *Diabetes*, vol. 48, 1999, pp. 445–451.

[4] R. Hovorka, V. Canonico, L.J. Chassin, U. Haueter, M. Massi- Benedetti, M.O. Federici, T.R. Pieber, H.C. Schaller, L. Schaupp, T. Vering, and M.E. Wilinska, Nonlinear Model Predictive Control of Glucose Concentration in Subjects With Type 1 Diabetes, *Physiol. Meas.*, vol. 25, 2004, pp. 905–920.

[5] Z. Trajanoski, "Simulation studies on neural predictive control of glucose using the subcutaneous route," *Comput. Methods Programs Biomed.*, vol. 56, pp. 133–139, 1998.

[6] P. Dua, F. J. Doyle, III, and E. N. Pistikopoulos, "Model-based blood glucose control for type 1 diabetes via parametric programming," *IEEETrans. Biomed. Eng.*, vol. 53, no. 8, pp. 1478–1491, Aug. 2006.

[7] Robert S.Parker, Francis J.Doyle III, Nicholas A.Peppas ," Model Based Algorithm for Blood Glucose control in

Type I Diabetes Patients", IEEE Transactions on Bio Medical Engineering, Vol.46, No.2, Feb-1999.

[8] Riccardo Bellazzi,"Bayesian analysis of BG Time series from Diabetes Home Monitoring. ", *IEEE Trans.Biomed. Eng*, July 2000.

[9] Cesar C.Palerm, B.Wayne Bequette," Issues in Hypo glycemia Prediction and Detection *IEEE Trans. on Biomed Eng*, 2004.

[10] Cesar C.Palerm & B.Wayne Bequette," Hypoglycemia Detection and Prediction using continuous Glucose Monitoring – A study on Hypoglycemic Clamp Data",*Journal of Diabetes Science and Technology*, Vol.1, Issue 5, Sep – 2007.

[11] G. Sparacino, F. Zanderigo, S. Corazza, A. Maran, A. Facchinetti, and C. Cobelli, Glucose Concentration Can Be Predicted Ahead in Time from Continuous Glucose Monitoring Sensor Time-Series, *IEEETrans. Biomed. Eng.*, vol. 54, 2007, pp. 931–937.

[12 ] J. Reifman, S. Rajaraman, A. Gribok, and W.K. Ward, Predictive Monitoring for Improved Management of Glucose Levels, *J. Diabetes Sci. Technol.*, vol. 1, 2007, pp. 478–486.

[13] Daniel A.Finan, Cesar C.Palerm, Francis J.DoyleIII,Howard Zisser, Lois Jovanovic, Wendy C.Bevier and Dale E.Seborg, "Identification of Empirical Models From type I Diabetes Subject Data", 2008 American Control Conference, Washington , USA

[14] Adiwanata Gani, Andrei V.Gribok, Srinivasan Rajaraman, W.Kenneth Ward, and Jaques Reifman, " Predicting Subcutaneous Glucose Concentration in Humans: Data-driven Glucose Modeling", *IEEE Trans.Biomed.Eng.*, vol.56, No.2, Feb.2009

[15] C.Perez-Gandia, A.Facchinetti, G.Sparacino, C.Cobelli, E.J. Gomez, M.Rigla, A.de.Leiva, M.E.Hernando, "Artificial Neural Network Algorithm for Online Glucose Prediction from Continuous Glucose Monitoring", *Diab.Tech.& Therap*., vol.12, No.1, 2010.

[16] L. Ljung, *System Identification: Theory for the User*. Upper Saddle River, NJ: Prentice-Hall, 1999.

[17] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*. Upper Saddle River, NJ: Prentice-Hall, 1999.

[18] A. N. Tikhonov and V. Y. Arsenin, *Solutions of Ill-Posed Problems*. Washington, DC: Winston, 1977.

[19] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis*. New York: Springer, 2005.

[20] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis Forecasting and Control, Third ed. Englewood Cliffs, NJ: Prentice*-Hall, 1994.

[21] L. Liu and G. P. Hudak, *Forecasting and Time Series Analysis Using the SCA Statistical System: Scientific Computing Associated, 1994.*

# Automatic Threshold Selection Based on Histogram Gaussian Estimation Method in FPGA

Deng-Yuan Huang[1], Ta-Wei Lin[1] and Wu-Chih Hu[2]

[1]Department of Electrical Engineering, Dayeh University,
168 University Rd., Dacun, Changhua 515, Taiwan

[2]Departement of Computer Science and Information Engineering, National Penghu University of Science and Technology,
300 Liu-Ho Rd., Makung, Penghu 880, Taiwan

{kevin@mail.dyu.edu.tw, daweimailbox@gmail.com, wchu@npu.edu.tw}

***Abstract*:** A fast and efficient algorithm called HGEM (Histogram-based Gaussian Estimation Method) based on an FPGA (Field Programmable Gate Array) is developed to automatically determine a threshold value for a Sobel edge detector. In comparison with Otsu's method based on a discriminant criterion, the proposed method is more efficient in computing performance. The proposed method is also simple to be implemented on the FPGA since it avoids the repetitious iterations and complex arithmetic operations in Otsu's thresholding procedures. The relative error (RE) of HGEM to Otsu's method is utilized to measure the closeness of the thresholds obtained by the two methods. The relative error is less than 1.50% for all the test images, indicating that the proposed method has the approximately same accuracy as that of Otsu's method. Timing simulations show that the FPGA circuits can run at a speed of up to 193.9 MHz, which is equivalent to a theoretical frame rate of 1,479 frame/s for a gray-level image of 256×256. This result confirms that the proposed hardware architecture can achieve the requirements for a real-time image processing system.

***Keywords*:** Otsu's method; binary thresholding; image segmentation; field programmable gate array (FPGA)**.**

## 1. Introduction

Automatic thresholding is a very straightforward and effective technique used in the fields of image processing, pattern recognition and computer vision. However, it requires an adequate threshold value to extract objects of interest from their background, since objects in an image have their own distinct gray-level distributions. Thresholding methods are widely used in many application domains, such as human action recognition [1], optical character recognition (OCR) [2],[3], automatic defect inspection [4],[5], video change detection [6]-[8], moving object segmentation [9]-[12], and medical image diagnoses [13],[14]. As a fundamental task for image preprocessing, many researchers pay much attention to the method of how to determine appropriate thresholds.

These applications demand real-time performance and hardware implementation, especially for an FPGA, is essential to increase the computational efficiency of thresholding procedures. Hence, the choice of a thresholding method for implementation on an FPGA board is important. In binary thresholding for image segmentation, Otsu's method [15] is a very popular global automatic thresholding technique; it selects an optimum threshold by maximizing the between-class variance in a gray-level image. However, the basic Otsu thresholding computations involve repetitious iterations of the zero- and first-order cumulative moments of a gray-level histogram, which requires a great number of complex arithmetic operations such as multiplications and divisions. The heavy computational resource makes Otsu's method unsuitable for a high-speed low-cost implementation in FPGA.

Otsu's method is simple to be implemented in software, but it is less efficient when implemented in FPGA circuits. Tian *et al.* [16] introduced a binary logarithmic conversion unit (LCU) to implement Otsu's method by eliminating the complex divisions and multiplications in the computations of between-class variances. The hardware was synthesized with Synplicity Synplify Pro 7.0.3 targeted at the Xilinx Virtex XCV800 HQ240-4 FPGA device. The results for implementations on the FPGA platform showed that their method is 2.75 times faster because it occupies only 1/6th of the FPGA slices required by a direct implementation. The introduction of an LCU can avoid the complex computations of divisions and multiplications, but repetitious computations are still required to search for the maximum between-class variance to determine an optimum threshold.

To eliminate both the repetition and complex arithmetic operations in the computations of between-class variance, we present a fast algorithm called HGEM (Histogram-based Gaussian Estimation Method), which is based on the Gaussian distributions of a histogram to determine an optimal threshold for gray-level images. The proposed method is relatively simple and efficient for implementation on an FPGA platform when compared to the basic Otsu thresholding procedures. We also develop a Sobel-based edge detector in the FPGA circuits as a target platform for the HGEM. To detect the presence of an edge pixel in an image, an appropriate threshold value is required to compare it with the magnitude of the Sobel gradient. Therefore, HGEM can be used to choose the optimum threshold for the Sobel-based edge detector.

Most algorithms for edge detection need to perform a convolution with an image in the spatial domain using a specific mask like a Sobel operator. Benkrid *et al*. [17] proposed a general framework which is built on a library of hardware skeletons for FPGA-based image processing. Two methods, online arithmetic and 2's complement LSBF (least significant bit first) with bit serial transfer, were presented to implement the Sobel-based edge detector on an FPGA board. Time simulations revealed that for a 256×256 gray-level image, the Sobel-based edge detector can run at a speed of 75 MHz, which leads to theoretical frame rates of 88 and 104 frame/s for online arithmetic and 2's complement LSBF methods, respectively. However, [17] did not explicitly describe the determination of the threshold required to establish the presence of an edge pixel. Other studies [18]-[20] have also omitted this description.

Rosas *et al*. [18] utilized a SIMD (single instruction multiple data) architecture based on an FPGA which was connected to two external RAMs modules. One of the RAMs modules was used to store the image captured by a CMOS sensor, and the other was used to store the image processed by the FPGA. In our study, instead of using external RAMs modules, the proposed hardware architecture uses built-in dual-port block RAMs to implement the HGEM algorithm targeted at a Sobel-based edge detector because it can store great amounts of data and access it quickly.

To evaluate the accuracy of the optimum threshold obtained for an image, Sezgin and Sankur [21] employed the following five methods to assess 40 existing thresholding algorithms: misclassification error (ME), edge mismatch (EMM), region non-uniformity (NU), relative foreground area error (RAE), and modified Hausdorff distance (MHD). In this paper, the method of ME is used to evaluate the accuracy of HGEM and the Otsu method.

This paper presents an efficient framework for threshold determination based on the FPGA using the HGEM algorithm. The rest of this paper is organized as follows: Section 2 briefly describes the proposed system architecture. Section 3 then gives a detailed description of the proposed HGEM method. The experimental results are discussed in Section 4, and Section 5 contains the concluding remarks of this work.

## 2. System Architecture

The proposed architecture for an FPGA-based image processing system is shown in Fig. 1. The hardware was implemented using a Xilinx ISE 8.1i IDE (Integrated Development Environment) tool on the ML401 Xilinx Virtex-4 (XC4VLX25) FPGA based board. This system consists of the following primary components: a data transmission unit, an image segmentation unit, a moving window generator, a three-stage Sobel pipeline unit, and a threshold estimation unit. The data transmission unit is designed to transfer original image pixels from a PC to the FPGA through a UART (universal asynchronous receiver and transmitter) module. To perform convolution, an input image with $m \times n$ pixels has to be convoluted with a $p \times q$ mask. The moving window generator is used to sequentially extract a $p \times q$ window of neighboring pixels from the input image. The three-stage Sobel pipeline unit is adopted to carry out the convolution of the Sobel gradient operator with the window of image pixels acquired by the moving window generator. The threshold estimation unit is an implementation of the proposed HGEM algorithm, which performs histogram statistics, threshold search, and threshold determination. The image segmentation unit is to binarize input image by a threshold that is determine by the threshold estimation unit.

A threshold value is required for segmenting a gray-level image. An algorithm called HGEM is proposed to replace the Otsu's method based on the feasibility of implementation on the FPGA device. The proposed algorithm HGEM, as implemented in the threshold estimation unit, involves the computing procedures including histogram statistics, threshold search, and threshold determination. As illustrated in Fig. 1, when the convolution operation is complete, the processed image (or output image) is stored in block RAMs and then sent back to the PC through the UART module for further verification. The details of the data transmission unit, the moving window generator, and the three-stage Sobel pipeline unit are described in the following sections.

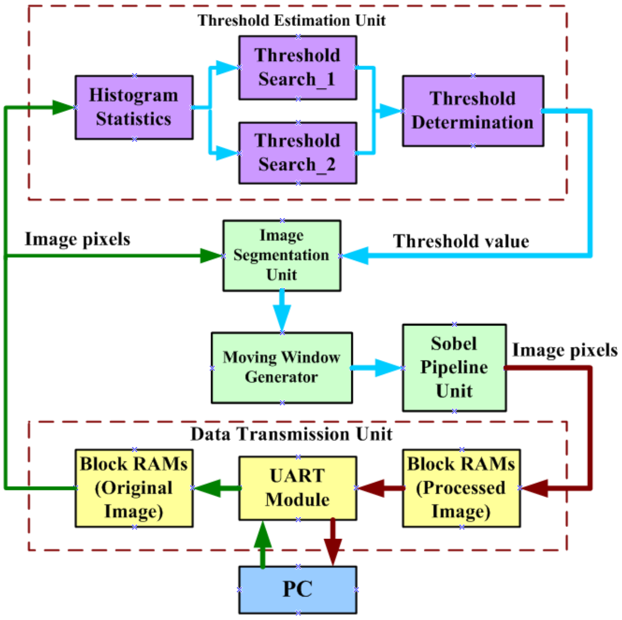


**Figure 1.** System architecture for threshold determination targeted at the Sobel edge detector

### 2.1 Data transmission unit

To verify the correctness of the proposed architecture shown in Fig.1, the output image stored in block RAMs, which can be transmitted to a PC by the UART module, is compared with that calculated by software. The architecture of data transmission unit is shown in Fig. 2. This unit primarily serves the following two functions (1) to transmit the original image from a PC to the FPGA, and (2) to send back the output image from the FPGA to the PC.

The byte data of the image is sent by a PC through the UART pin RXD using serial transmission to the UART_Rx module on the FPGA, as shown in Fig. 2. When the UART_Rx module has completely received one byte of image data, it sends the RXD_data[7:0] and RXD_ready signals. RXD_ready is used to trigger the address controller to address the memory locations of block RAMs A to store the image data that is contained in register RXD_data[7:0]. When the image data has been fully transmitted, the address controller stops the action of writing the image data into block RAMs A to avoid writing error. The image can then be processed in the FPGA circuits.

When the output image has been produced, one can push the transmit button with debouncing capability to trigger the signal of TXD_start to initiate the transmission of the image data from the UART_Tx module to a PC. When one byte of the image data has been transmitted, the signal of TXD_busy is pulled down to a low level to trigger the address controller to acquire the next byte of the image data from block RAMs B into register TXD_data[7:0]. Then, the UART_Tx module sends the processed image data back to the PC until the total image has been completely transmitted. With the aid of this unit, the transmission of the original image and the verification of the output image can be easily achieved.

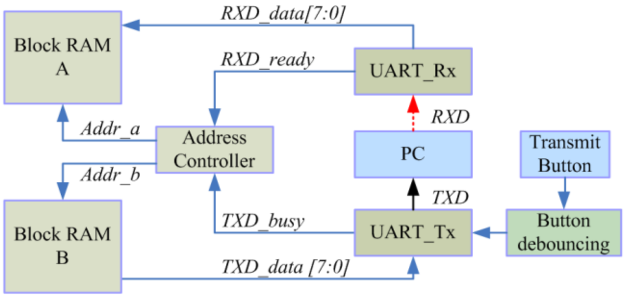


**Figure 2.** Architecture of the data transmission unit

### 2.2 Moving window generator

To compute the Sobel gradient, a 3x3 (i.e., p=q=3) window of neighboring pixels extracted from the input image is required for convoluting with a Sobel gradient operator. This neighborhood window then moves over the whole image until an output has been produced for all pixels. Generally, it is not practical to store the whole image in RAMs before starting computations due to the limited CLBs (Configurable Logic Blocks) in FPGA. A better way is to only store the image pixels required to perform the current convolution operation.

Figure 3 shows the architecture of the moving window generator which comprises nine flip-flops and two line buffers (or FIFOs; first in first out). The line buffers (i.e., FIFO A and B) and flip-flops are used to store one row of image data with a dimension of *n*, with each grayscale pixel represented by 8 bits. Generally, when a $p \times q$ convolution mask is applied, $[(p-1) \times n+q] \times 8$ registers are required. Line buffers can be implemented using either shift registers or block RAMs in FPGA. Generally, when an FPGA has no built-in block RAMs, the only way to implement the line buffer is to use shift registers [17],[20],[22],[23], which usually consume a large number of FPGA gates. For example, when a 3x3 convolution window is applied to a 256x256 (i.e., $m=n=256$) image with 8-bit pixels, (2x256+3)x8=4,120 flip-flops are required. That is about 19% (4,120/21,504*100%) of all the available flip-flops in the Virtex-4 (XC4VLX25) FPGA used in this case. Furthermore, if the size of the convolution mask or image becomes larger, the required gate counts or CLBs of the FPGA will increase accordingly. To reduce the consumption of FPGA CLBs, [24] utilized block RAMs to implement the line buffers.

This paper also adopts the block RAMs to implement the line buffers using a framework similar to that used in [24]. Using block RAMs to implement the line buffers not only reduces the consumption of FPGA gates, but also lowers the required routing of logic elements. Less routing of logic elements implies that a higher operation speed of the FPGA circuits can be achieved. This is verified by Figs. 4 and 5.

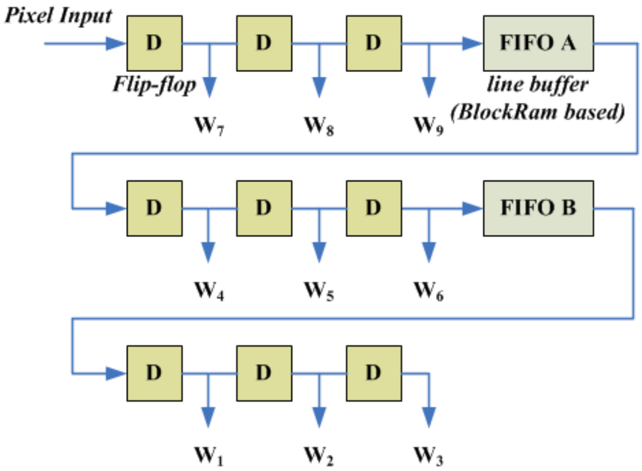


**Figure 3.** Architecture of the moving window generator

The FPGA gate counts for block RAM-based and shift register-based FIFOs under various image sizes are shown in Fig. 4. The results show that the required gate count increases from 4,000 to 38,000 when the image size grows from 32x32 to 512x512 pixels for the case of a shift register-based FIFO. However, when block RAMs is used, the required gate count remains approximately constant at about 2,000 with increasing image size. Figure 5 shows the effects of image size on the operational speed of the FPGA circuits. Generally, the total speed decreases when the image size becomes larger for both shift register-based and block RAM-based FIFOs. Sharp declines of speed can be observed when the image size increases from 128x128 to 512x512. In addition, the average speed for a block RAM-based FIFO is much higher than that of a shift register-based one. Figs. 4 and 5 show that the operational speed can be increased by 15.6% and that the logic elements of the FPGA can be reduced by 74.2% when block RAMs is used to replace shift registers to implement the FIFOs in the moving window generator.

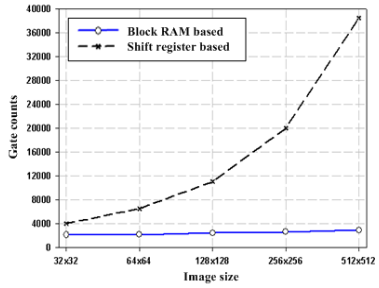


**Figure 4.** Effect of image size on the gates consumed in FPGA

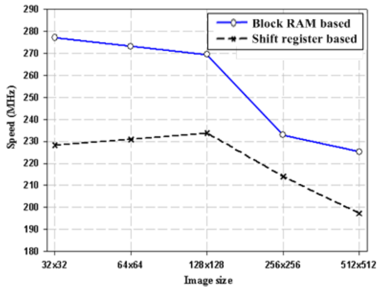


**Figure 5.** Effect of image size on the operational speed of the FPGA chip system

### 2.3 Three-stage Sobel pipeline unit

Many methods for edge detection have been implemented with convolution masks, and most are derived from the differential operators, which measure the rate of change in brightness of an image. Generally, a large change in brightness in an image over a short spatial distance (typically one pixel) reveals the existence of an edge. The most popular convolution mask used in edge detection is the Sobel gradient operator, which looks for edges in both the horizontal and vertical directions and then combines this information into a single metric, as shown in Fig. 6. The convolution can be carried out with the Sobel gradient operator as follows:

$$G_x = (w_7 + 2w_8 + w_9) - (w_1 + 2w_2 + w_3) \qquad (1)$$

$$G_y = (w_3 + 2w_6 + w_9) - (w_1 + 2w_4 + w_7) \qquad (2)$$

where Gx and Gy are called the “row mask” and the “column mask,” respectively. Since both have the same computational complexity in performing the convolution operation, this paper only implements horizontal edge detection, i.e., Gx, to avoid detecting redundant edge information.

| | | |
|---|---|---|
| w1 | w2 | w3 |
| w4 | w5 | w6 |
| w7 | w8 | w9 |

Coefficient matrix

| | | |
|---|---|---|
| -1 | -2 | -1 |
| 0 | 0 | 0 |
| 1 | 2 | 1 |

Gx

| | | |
|---|---|---|
| -1 | 0 | 1 |
| -2 | 0 | 2 |
| -1 | 0 | 1 |

Gy

**Figure 6.** Sobel masks used to compute gradients Gx and Gy

Figure 7 shows the architecture of the three-stage Sobel pipeline unit. In this study, three pipelines are employed to improve the performance of the system. First, stage-1 pipelining deals with the additions and multiplications of the image window pixels with the Sobel gradient operator. In this stage, four additions and two multiplications are required, where multiplication is performed using one shift-left operation rather than using a multiplier to save logic elements. Stage-2 determines the absolute value for Gx, and stage-3 outputs an enhanced edge detection image. When synthesizing individually, this unit can run at a speed of up to 238.2 MHz.

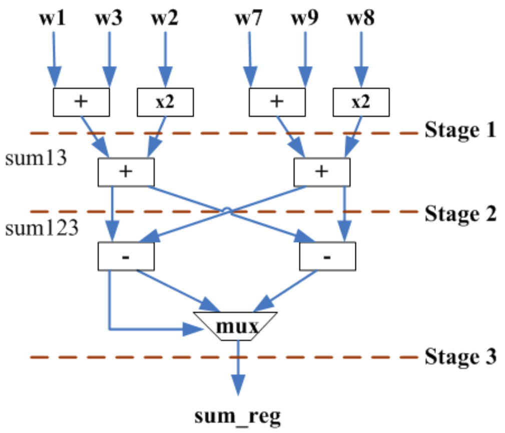


**Figure 7.** Architecture of the three-stage Sobel pipeline unit

## 3. Threshold selection algorithm

In an image processing system, the success of image segmentation highly depends on the capabilities of the thresholding method to determine an optimum threshold. Lee *et al.* [25] adopted the histogram concavity technique to locate the optimal threshold value. In their method, the slopes of all the line segments are calculated from the starting gray level. Then, the background peak, Bp, with the greatest slope can be obtained. Similarly, the object peak, Op, can be secured starting from the opposite direction. As a result, the optimal threshold can be found somewhere between Bp and Op. However, this method fails to find the optimal threshold for the special case when Bp meets Op due to extremely high histogram data in a gray level.

Some heuristic approaches have been presented to determine the optimal threshold. El-Khamy *et al.* [26] proposed a so called “Modified Fuzzy Sobel” method that uses a fuzzy reasoning-based algorithm to detect the edges of an image. They first divided an image into two fuzzy regions, i.e., the Fuzzy Smooth region and the Fuzzy Edge region, and then constructed a difference histogram from the input image. The four threshold values used to define the boundaries of the image fuzzy region were used to build a membership function to determine the optimal threshold.

The methods proposed by [25],[26] are relatively simple to implement in software, but they are quite difficult to implement in FPGA circuits due to the determination of varying slopes for [25], and the calculation of the difference histogram for [26]. To achieve much higher accuracy in thresholding estimation, a lot of studies [27]-[29] have used the entropic thresholding technique to find an optimal threshold instead of adopting histogram shape-based methods. However, this method needs to calculate the probability distributions for the edge and non-edge pixels, making it computationally expensive and hard to implement on an FPGA. To balance the computational cost and thresholding estimation accuracy, the present study proposes a histogram-based Gaussian estimation method (HGEM), which is not only easily implemented in FPGA circuits, but also provides a more reasonable threshold value.

### 3.1 Histogram-based Gaussian estimation method (HGEM)

HGEM is based on the analysis of the gray-level probability density function (pdf) for an image[30]. When the histogram is modeled as two different Gaussian functions, as shown in Fig. 8, with means and variances ( $\mu_1$ , $\sigma_1^2$ ) and ( $\mu_2$ , $\sigma_2^2$ ), respectively, the histogram function becomes:

$$p(z) = \frac{P_1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(z-\mu_1)^2}{2\sigma_1{}^2}} + \frac{P_2}{\sqrt{2\pi}\sigma_2} e^{-\frac{(z-\mu_2)^2}{2\sigma_2{}^2}} \qquad (3)$$

and

$$P_1 + P_2 = 1 \qquad (4)$$

where z denotes gray level values, and $P_1$ and $P_2$ are the probabilities of occurrence of the two classes of pixels, respectively. To find the optimal threshold value *T* in Fig. 8, erroneous classifications, which assign a background pixel to the object, and vice versa, should be minimized.

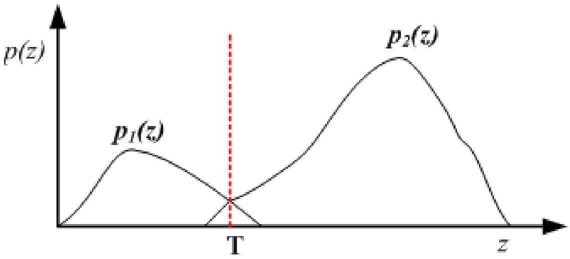
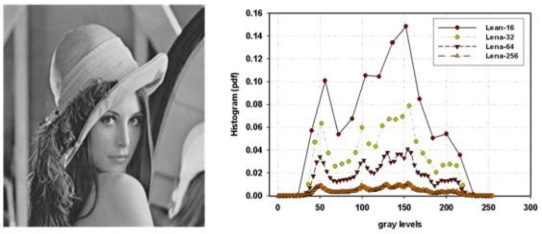


**Figure 8.** Graph of the probability density function (pdf) for gray-level distribution

To greatly reduce the hardware resources required to implement HGEM on the FPGA, we performed histogram binning by employing wider bin widths. In this study, 16 bin groups, which contain 16 gray levels in every group, are employed to compute the histogram of gray levels to find the optimal threshold. One may argue when the bin width is beyond a certain limit, it may destroy the modes or the valleys in between. However, if the bin width is constrained within a reasonable range, the fine characteristics of the histogram in an image can still be retained. Hence, the operation of "histogram binning" greatly decreases the computational complexity and significantly reduces the required logic elements in the FPGA. Figure 9 shows the histogram of various bin groupings, i.e., 16, 32, 64, and 256, for the test image "Lena". As shown in this figure, the fine characteristics (i.e., valleys in between) of the histogram are similar for all the cases even when the widest bin (i.e., Lean-16) is used. Consequently, the bin width adopted falls into a reasonable range without missing the fine characteristics of the histogram.



(a) Lena (b) Histogram for various bin groupings

**Figure 9.** Histogram for various bin groupings for test image "Lean"

To efficiently determine the optimal threshold T shown in Fig. 8, the 16 bin grouping is employed. The index value (0 to 15) of the counter and count value of the histogram are used to estimate the Gaussian distribution of an image. Based on the index value and count value, the Gaussian distributions can be categorized into four types. The details of the HGEM algorithm used to determine the optimal threshold are described below.

To complete the histogram of gray levels in an image, the counters are first labeled C0 to C15. Hence, when the values of gray levels are in the ranges of 0 to 15, 16 to 31, …, and 240 to 255, they will be grouped into counters C0, C1, …, and C15, respectively. Then, the histogram can be modeled as two distinct Gaussian functions, divided into the left region, i.e., C0 to C7, and the right region, i.e., C8 to C15. Next, we can search for the index values, max1_1 and max1_2, corresponding to the first two largest count values in the direction from C7 to C0 in the left region. Similarly, the index values, max2_1 and max2_2, can be obtained from C8 to C15 in the right region. Typical search results are shown in Fig. 10. Two of the four index values, i.e., th1 and th2, can be selected based on the type of Gaussian distribution to which the shape of the image histogram belongs. Finally, the threshold value *T* can be calculated as 16×(th1+th2)/2. The HGEM algorithm for finding th1 and th2 is described in the style of the C-language for the following four cases.

$$\begin{array}{l} \textit{if}\ \ (\max1\_1 \neq 7\ \textit{and}\ \ \max2\_1 \neq 8) \\ \quad th1 = \max1\_1\ \textit{and}\ \ th2 = \max2\_1 \end{array} \qquad (5)$$

$$\begin{array}{l} \textit{if}\ \ (\max1\_1 = 7\ \textit{and}\ \ \max2\_1 = 8) \\ \quad th1 = \max1\_2\ \textit{and}\ \ th2 = \max2\_2 \end{array} \qquad (6)$$

$$\begin{array}{l} \textit{if}\ \ (\max1\_1 \neq 7\ \textit{and}\ \ \max2\_1 = 8) \\ \quad th1 = \max1\_1 \\ \textit{if}\ \ (C[\max2\_1] \neq 0\ \textit{or}\ \ \max1\_2\ >\ \max1\_1) \\ \quad th2 = \max2\_1 \\ \textit{else} \\ \quad th2 = 1\ \text{(minimum)} \end{array} \qquad (7)$$

$$
\begin{aligned}
&\mathit{if}\ \ (\max 1\_1 = 7\ \mathit{and}\ \ \max 2\_1 \neq 8) \\
&\quad th2 = \max 2\_1 \\
&\mathit{if}\ \ (C[\max 1\_1] \neq 0\ \mathit{or}\ \ \max 2\_2\ <\ \max 2\_1) \\
&\quad th1 = \max 1\_1 \\
&\mathit{else} \\
&\quad th1 = 14\ (\text{maximum})
\end{aligned}
\tag{8}
$$

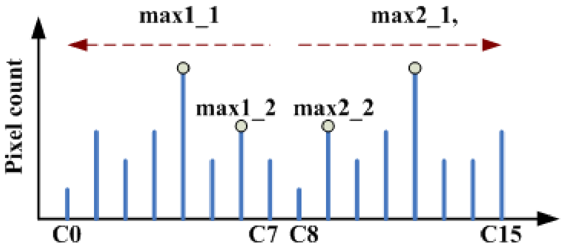
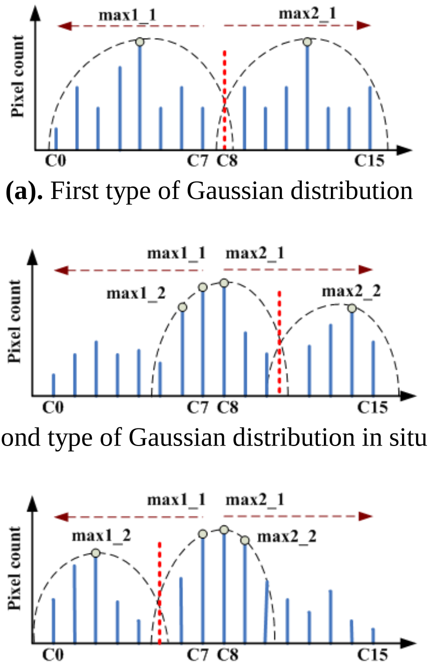


**Figure 10.** Typical distribution of the histogram with 16 counters

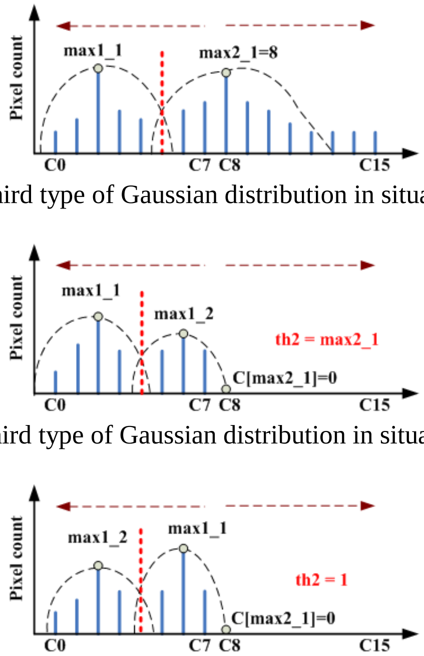
Equation (5) represents the first type of Gaussian distribution shown in Fig. 11(a). If the two largest count values with corresponding index values max1_1 and max2_1 are found on the opposite side of the histogram, the possible Gaussian distributions can be estimated around max1_1 and max2_1, as indicated in Fig. 11(a). Hence, the index values th1 and th2 can be selected as max1_1 and max2_1, respectively. Equation (6) denotes the second type of Gaussian distribution with the first two largest count values in the central region, i.e., corresponding to index values 7 and 8, as shown in Fig. 11(b) and (c). Thus, one possible Gaussian distribution can be estimated in the central region, but the other may be in the right region, as shown in Fig. 11(b), or in the left region, as shown in Fig. 11(c). Consequently, the index values of th1 and th2 should be determined as max1_2 and max2_2, respectively. However, if we choose th1 as max1_1 and th2 as max2_1, the threshold value, 16×(th1+th2)/2, must be in the central region of the histogram, implying that a greatly erroneous classification will be raised.



**(a).** First type of Gaussian distribution

**(b).** Second type of Gaussian distribution in situation-1

**(c).** Second type of Gaussian distribution in situation-2



**(d).** Third type of Gaussian distribution in situation-1

**(e).** Third type of Gaussian distribution in situation-2

**(f).** Third type of Gaussian distribution in situation-3

**Figure 11.** Analysis of Histogram-based Gaussian Estimation Method

The third type of Gaussian distribution is more complicated than the first and second ones, as shown in Fig. 11(d)-(f). If the first two largest count values are not next to each other in the central region, say max1_1≠7 but max2_1=8, one possible Gaussian distribution can be modeled around max1_1, as indicated in Fig. 11(d)-(f), and then th1 can be evaluated as max1_1. However, when the count value C[max2_1] is not equal to zero, the other possible Gaussian distribution can be estimated around max2_1 (see Fig. 11(d)), and th2 should be selected as max2_1. On the other hand, when the count value C[max2_1] is zero, no gray levels are larger than 128, as shown in Fig. 11(e) and (f). Hence, the other possible Gaussian distribution can be modeled around max1_2. As a result, th2 can be chosen as max2_1 when max1_2>max1_1, or 1 when max1_2<max1_1. As can be expected, the third type of Gaussian distribution often occurs in a darker image. The searching method of th1 and th2 (see Eq. (8)) for the fourth type of Gaussian distribution is similar to that of the third type but there are no gray levels smaller than 128. This always happens in a brighter image. Since the searching method is similar to that of the third type, its discussion is omitted here.

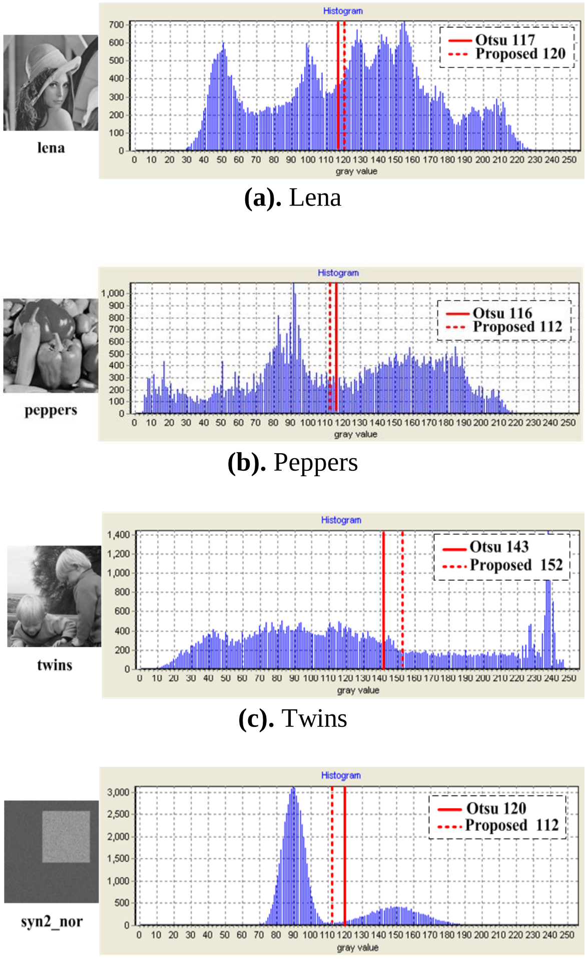
### 3.2 Comparison of HGEM and Otsu’s method

Sezgin and Sankur *et al.* [21] conducted an exhaustive survey of 40 selected image thresholding methods. The results confirm that the thresholding evaluation rank of 40 NDT (nondestructive testing) images according to the overall average quality score for Otsu’s method is relatively high, with a rank of 6 and an average score of 0.318. This indicates that Otsu’s method can provide a reasonable threshold value for image segmentation. Here, the threshold estimations of HGEM are compared with those of the Otsu method.

The testing images used in this study consists of natural images (see Fig. 12(a)-(c)) and artificial images (see Fig.

12(d)-(g)), where Fig. 12(d) and (e) are adopted from [31]. To evaluate the accuracy of threshold estimations by HGEM, an artificial image is much better than a real-world one. The threshold values estimated by HGEM for images Lena and Peppers, as shown in Fig. 12(a) and (b), respectively, are very close to those evaluated by Otsu's method. For the image Twins, as shown in Fig. 12(c), HGEM provides a more reasonable threshold estimation, which is much closer to the deeper valley than that of Otsu's method, implying that HGEM can find a satisfactory threshold value even for an image histogram with a wide flat valley. A similar result was obtained for Fig. 12(d). That is, HGEM provided a threshold value exactly in the valley between the two peaks of Gaussian distributions. However, the threshold value estimated by Otsu's method was shifted to the edge part of the right Gaussian distribution. Furthermore, when three objects appear in one image, as shown in Fig. 12(e), an appropriate threshold value was obtained by both methods.

Images with different luminance levels were also examined by HGEM to verify the robustness of the threshold estimations. Images with low luminance, as shown in Fig. 12(f), and with high luminance, as shown in Fig. 12(g), were tested using HGEM and Otsu's method. The results indicate that a reasonable threshold value can be obtained using either method, even for images with large variations in luminance. However, HGEM is computationally efficient; it avoids both repetitious iterations and complex arithmetic operations that are required to compute the between-class variance when using Otsu's method.

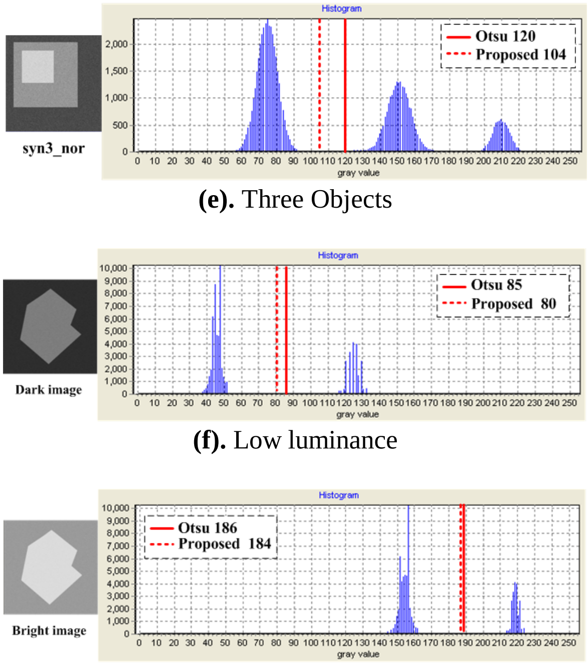


**(a).** Lena

**(b).** Peppers

**(c).** Twins

**(d).** Two Objects



**(e).** Three Objects

**(f).** Low luminance

**(g).** High luminance

**Figure 12.** Testing images for threshold estimation

To visually compare the segmented results obtained by HGEM and Otsu's method, three 256×256 test images (i.e., Lena, Peppers, and Twins), with each pixel represented by 8 bits, were used. The segmented images with their corresponding thresholds are shown in Fig. 13. The figure shows that the thresholds evaluated by HGEM are very close to those of Otsu's method, indicating that a closely visual perception between them can be achieved.

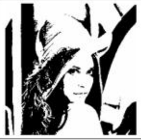
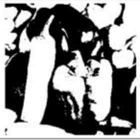
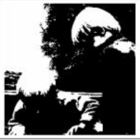
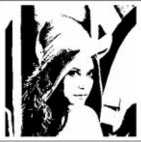
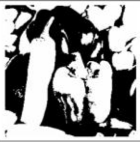
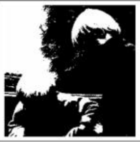
| | Lena | Peppers | Twins |
|---|---|---|---|
| Original Image | | | |
| Otsu |  |  |  |
| threshold | 117 | 116 | 143 |
| HGEM |  |  |  |
| threshold | 120 | 112 | 152 |

**Figure 13.** Segmented images of binary thresholding for HGEM and Otsu's method

A comparison of accuracy using the ME method [21] for HGEM and Otsu's method was performed. Gray-level images extracted from the CEDAR database of handwritten words with 29 test images [32], and from the FVC2000 database of fingerprints with 70 test images [33], were employed. Some typical samples are shown in Figs. 14 and 15 with the corresponding ground-truth images. The ground-truth images can be obtained by visually determining the valley of the histogram of the test images. The average results of ME for 29 test images of the CEDAR database and 70 test images of the FVC2000 database were used to evaluate the accuracy of bi-level thresholding for the two methods.

**Figure 14.** Typical images in the CEDAR database of handwritten words

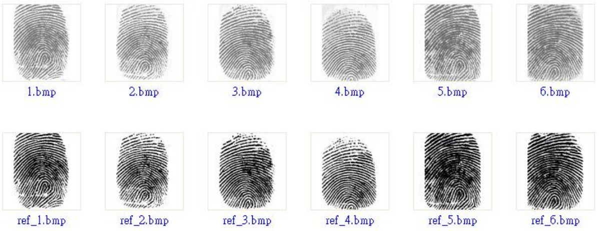


**Figure 15.** Typical images in the FVC2000 database of fingerprints

The index of ME is quite useful for quantifying the percentage of background pixels wrongly assigned to the foreground, and vice versa. For bi-level segmentation, ME can be simply represented as

$$ME = 1 - \frac{|B_O \cap B_T| + |F_O \cap F_T|}{|B_O| + |F_O|} \quad (9)$$

where $B_O$ and $F_O$ denote the background and foreground of the original (ground-truth) image, respectively, and $B_T$ and $F_T$ denote the background and foreground pixels in the test image, respectively. Note that the value of ME varies from 0 for a totally well classified image to 1 for a completely wrongly binarized image.

Table 1 shows the results of ME of bi-level thresholds for the two methods using the test images of handwritten words taken from CEDAR and fingerprints taken from FVC2000. As indicated in Table 1, the values of ME are very close for fingerprints images, but there are small differences for handwritten words images under the cases of no noise. The results of ME after adding Gaussian noise with standard deviations of   and   were also examined. Approximate ME values were obtained by the two methods although there was some noise in the test images.

The relative error (RE) of HGEM to Otsu's method, defined in Eq. (10), can be used to measure the closeness of the threshold values obtained by the two methods, where (1-ME) means the percentage of the correct classification of image pixels. As indicated in Table 1, the maximum RE with cases of no noise is 1.494% for handwritten words images. However, when Gaussian noise was added to the test images, the maximum RE is only 1.104% in the CEDAR database. Consequently, the relative errors of HGEM to Otsu's method in all cases are less than 1.50%.

$$RE = \frac{|(1 - ME_{Otsu}) - (1 - ME_{TSMO})|}{1 - ME_{Otsu}} = \frac{|ME_{Otsu} - ME_{TSMO}|}{1 - ME_{Otsu}} \quad (10)$$

**Table 1.** Comparisons of ME and RE with cases of no noise, σ=10, and σ=20

| Methods | Handwritten words | | | Fingerprints | | |
|---|---|---|---|---|---|---|
| | No noise | σ=10 | σ=20 | No noise | σ=10 | σ=20 |
| ME(Otsu) | 0.0162 | 0.0079 | 0.0221 | 0.0298 | 0.0225 | 0.0442 |
| ME(HGEM) | 0.0309 | 0.0163 | 0.0113 | 0.0295 | 0.0315 | 0.0444 |
| RE(%) | 1.494 | 0.847 | 1.104 | 0.031 | 0.921 | 0.021 |

### 3.3 Comparison of HGEM and Otsu's method

The corresponding FPGA circuits for the proposed HGEM method were designed based on the architecture of the threshold estimation unit shown in Fig. 16. This unit comprises the following three modules: histogram statistics, threshold searching, and threshold determination. The histogram statistics module is used to divide the number of gray levels into 16 counters (C0 to C15) as described earlier; it then outputs the resulting histogram to the threshold searching module. As indicated in Fig. 16, the threshold searching module consists of two sub-modules, namely, threshold search1 used to find max1_1 and max1_2, and threshold search2 used to find max2_1 and max2_2. Then, the threshold determination module determines th1 and th2 based on the proposed HGEM method (see Eqs. (5) – (8)). Finally, the threshold value can be evaluated as 16×(th1+th2)/2 in this module.

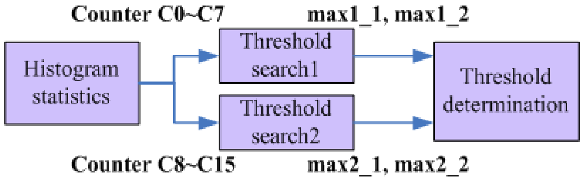


**Figure 16.** Architecture of the threshold estimation unit

Some issues, such as latency in the design of the FPGA, should be considered carefully. Theoretically, when a 256×256 gray-level image is used, the width of the counters must be 16 bits to avoid overflow. Therefore, 16-bit wide comparators are required to find the first two largest count values in the submodules: threshold search1 and search2. However, the latency of the synthesized circuit in the FPGA is very serious due to the larger number of bits, i.e., 16 bits, used in the comparators. To improve the latency of the threshold searching module, the number of bits was reduced to 10 in the comparators. By only comparing the results in higher bits of the counters and discarding lower bits with allowable losses in accuracy, the number of bits used in the comparator can be reduced from 16 to 10, which increases the performance of this system from 184.6 MHz to 193.9 MHz.

The total cycles required for threshold estimation in the FPGA includes the time consumed by the three modules mentioned above, where the time required by the histogram statistics module highly depends on the input image size; for example, 65,536 cycles are needed to complete the histogram when the input image is 256×256 pixels. Moreover, eight cycles are needed for the threshold searching module to determine the index values corresponding to the first two largest count values for both regions; max1_1 and max1_2 for the left region, and max2_1 and max2_2 for the right region. Four cycles are required for the threshold determination module to determine th1 and th2, and to complete the calculation of the threshold value, 8×(th1+th2). In this study, the threshold searching and determination modules were designed using state machines. Consequently, the total time needed by the threshold estimation unit for a 256×256 image is 65,536+8+4=65,548 cycles.

## 4. Experimental results

The proposed system architecture consists of the moving window generator, the three-stage Sobel pipeline unit, and the threshold estimation unit. The components were integrated and implemented into a chip system with a Xilinx Virtex-4 (XC4VLX25) FPGA. The total execution time can be evaluated as the cycles consumed by the moving window generator (=65,536 cycles), the three-stage Sobel pipeline unit (=4 cycles), and the threshold estimation unit (=65,548 cycles), a total of 65,536+4+65,548=131,088 cycles. The synthesized results of this FPGA chip system including the UART data transmission unit reveal that the required gate count is only 17,101, and that the operation speed can reach up to 193.9 MHz, which is equivalent to the processing rate of 1,479 (=193.9MHz · 106/131,088) frame/s for a 256 · 256 image, as indicated in Tables 2 and 3.

**Table 2.** Comparison of performance of Sobel-based edge detector on FPGA

| Architecture | Image size | Operation speed | Frame/s |
|---|---|---|---|
| **Proposed system** | 256x256 | 193.9 MHz | 1,479 |
| **K. Benkrid, 2002[17]** | 256x256 | 75 MHz | 104 |
| **X. Li, 2003[19]** | 256x256 | 40 MHz | 610 |
| **R.L. Rosas, 2005[18]** | 640x480 | 13.2 MHz | 43 |

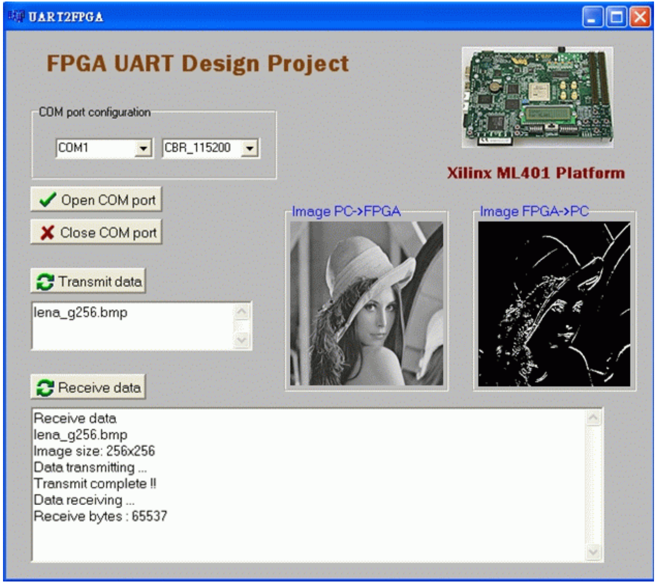
**Table 3.** List of synthesized resources for individual components on FPGA

| Module | **Slice flip flops** | **4 input LUTs** | **Occupied slices** | **BRAM/ FIFOs** | **Speed (MHz)** | **Gate Count** |
|---|---|---|---|---|---|---|
| UART | 55 | 121 | 64 | 0 | 249.6 | 1,203 |
| Histogram statistic | 257 | 49 | 146 | 0 | 324.9 | 5,846 |
| Threshold search 1/2 | 77 | 194 | 116 | 0 | 266.6 | 1,816 |
| Threshold segment | 10 | 50 | 31 | 0 | 260.2 | 804 |
| Moving window generator | 140 | 189 | 131 | 2 | 237.9 | 2,382 |
| Sobel pipeline unit | 51 | 71 | 44 | 0 | 238.2 | 1,305 |
| Total system | 995<br>(4%) | 1,116<br>(5%) | 930<br>(8%) | 38<br>(52%) | 193.9<br>(%) | 17,101<br>(%) |
| Virtex-4XC4VLX25 | 21,504 | 21,504 | 10,752 | 72 | 500 | - |

Table 2 compares the performance of the Sobel-based edge detectors implemented by [17],[18],[19] for some fixed image sizes. However, the edge detectors [17]-[19] were implemented on a pre-specified threshold value. The operational speed of the proposed architecture, i.e., 193.9 MHz, is about 2.5 to 4.8 times greater than those of [17],[19] for 256×256 images. The processing rate, i.e., 1,479 frame/s, far exceeds the requirement of a real time image processing system. This implies that highly efficient image processing can be achieved using the proposed architecture.

Table 3 lists the synthesized resources of the FPGA chip system for individual components. As indicated in Table 3, the histogram statistics module, implemented using only 16 counters with a width of 16 bits, consumes the most resources of the FPGA, with about 5,846 gates (about 34% of the total system), to complete a histogram of a gray-level image. That is why this work does not use 256 counters, corresponding to 256 gray levels, to implement this module. Furthermore, since the function of this module is relatively simple, the operational speed, i.e., 324.9 MHz, is the highest among the modules. Although all components can perform individually with an operation speed higher than 235 MHz, the total speed of the integrated system is only 193.9 MHz when all components have been interconnected. Moreover, to readily access the image data, 38 blocks of RAMs were used in the FPGA chip system, where two of the blocks were used by the moving window generator, and the others were employed for temporarily storing the image data.

Figure 17 shows the software used for interfacing a PC and the FPGA chip system, developed using Borland C++ Builder 6.0. This software can be used to verify the correctness of the edge detection obtained by the proposed architecture. As indicated in Fig. 18, the edge detection of the image "camera man" by the software (Fig. 18(b)) and FPGA (Fig. 18(c)) is the same, implying that a highly accurate threshold value can be obtained using the proposed architecture.

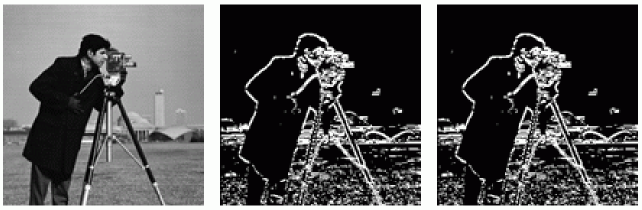


COM port configuration: COM1, CBR_115200

lena_g256.bmp

```
Receive data
lena_g256.bmp
Image size: 256x256
Data transmitting ...
Transmit complete !!
Data receiving ...
Receive bytes : 65537
```

**Figure 17.** Developed software for threshold estimation by HGEM

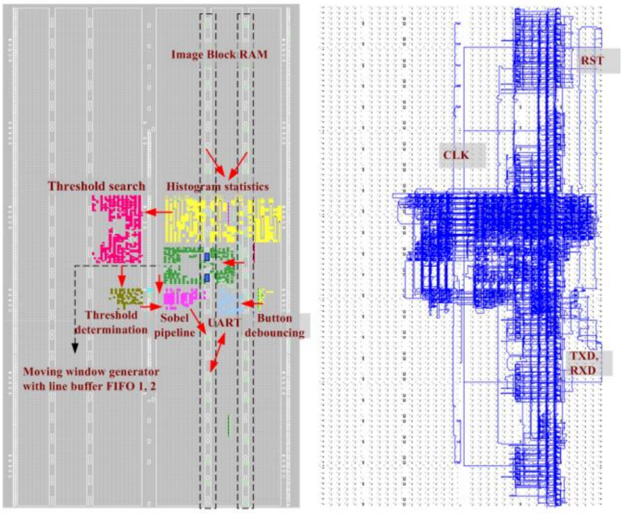


**(a).** Original image **(b).** Software **(c).** FPGA

**Figure 18.** Comparison of edge detection by software and FPGA

Figure 19 shows the placement and routing for all the components in the FPGA after floorplanning. As indicated in

Fig. 19(a), the components in FPGA can be arranged compactly by specifying the positions of CLBs (Configurable Logic Blocks) to reduce the required areas. The block RAMs, indicated by two rectangles with dashed lines (see Fig. 19(a)), was used to store the input and output images temporarily. The histogram statistics module accesses the block RAMs to acquire the image data and then groups it into 16 counters to complete the histogram of an input gray-level image. The placement of this module should be as close as possible to the block RAMs to save routing resources. However, since the synthesized logic of this module is relatively large, its placement may occupy the positions of some block RAMs, making this memory space unavailable for other modules. To reduce the number of occupied blocks of RAMs, the height of the floorplanning area of this module was shortened by area constraints, and the width was adjusted to cross over three CLB columns so that the input and output images could be completely stored in neighboring two columns of the block RAMs, greatly reducing the routing paths between this module and the block RAMs. Additionally, because the line buffers of the moving window generator need two blocks of RAMs, the placement of the moving window generator also needs to cross over two CLB columns to reduce routing paths.



**(a).** Placement **(b).** Routing

**Figure 19.** Placement and routing in the FPGA after floorplanning

## 5. Conclusion

A fast and efficient algorithm called HGEM, which is based on the Gaussian distributions of a histogram, was developed to determine a threshold for a gray-level image of which value is close to that of Otsu's method. The proposed method is simple and efficient for implementation on an FPGA, since it avoids the repetitious iterations and complex arithmetic operations, such as multiplication and division, when compared to the basic Otsu thresholding procedures. To use hardware resource more effectively, the block RAMs was used to implement the line buffers (FIFO A and B) of the moving window generator. The synthesized results indicate that the operation speed can be increased by 15.6%; the logic elements of the FPGA were reduced by 74.2%, when using the block RAMs to replace the shift registers.

Misclassification error (ME) was used in the evaluations of the accuracy for the proposed method. The maximum ME for HGEM in all test cases with or without noise (σ=10 and σ=20) was only 0.044, which is very close to the value obtained by Otsu's method. The relative errors (REs) were all less than 1.50% for the test images, indicating that a comparable threshold can be obtained by HGEM when compared to Otsu's method. Therefore, the proposed method is very efficient with an accuracy equivalent to that of Otsu's method.

The hardware architecture of the Sobel-based edge detector with an optimal threshold determined by HGEM comprises four major components: the UART transmission unit, the moving window generator, the three-stage Sobel pipeline unit, and the threshold estimation unit. The components were integrated and implemented into a single chip system with a Xilinx Virtex-4 (XC4VLX25) FPGA. The synthesized results reveal that the total required gates amount to 17,101, and that the total operation speed can run at up to 193.9 MHz, which is equivalent to a theoretical processing rate of 1,479 frame/s for 256×256 images. This result confirms that the proposed architecture on FPGA can easily achieve the requirements for a real-time image processing system.

## References

[1] S. Arseneau, J.R. Cooperstock, "Real-Time Image Segmentation for Action Recognition," in: Proc. IEEE Pacific Rim Conf. on Communications, Computers and Signal Processing, Victoria, B.C., Canada, pp. 86-89, 1999.

[2] Hammouche K, Diaf M, Siarry P, "A multilevel automatic thresholding method based on a genetic algorithm for a fast image segmentation," Comput. Vis. Image Und., Vol.109, No. 2, pp. 163-175, 2008.

[3] D.Y. Huang, C.H. Wang, "Optimal multi-level thresholding using a two-stage Otsu optimization approach," Pattern Recogn. Lett., Vol. 30, No. 3, pp. 275-284, 2009.

[4] D. Aiteanu, D. Ristic, A. Graser, "Content based threshold adaptation for image processing in industrial application," in: Int. Conf. Control and Automation, Budapest, Hungary, pp. 1022-1027, 2005.

[5] H.F. Ng, "Automatic thresholding for defect detection," Pattern Recogn. Lett., Vol. 27, No. 14, pp. 1644-1649, 2006.

[6] G. Jing, D. Rajan, C.E. Siong, "Motion Detection with Adaptive Background and Dynamic Thresholds," in: IEEE Int. Conf. Information, Communications and Signal Processing, Bangkok, Thailand, pp. 41-45, 2005.

[7] E.P. Ong, B.J. Tye, W. S. Lin, M. Etoh, "An efficient video object segmentation scheme," in: IEEE Proc. Int. Conf. Acoustics, Speech, and Signal Processing, Orlando, Florida, USA, pp. 3361-3364, 2002.

[8] C. Su, A. Amer, "A Real-Time Adaptive Thresholding for Video Change Detection," in: IEEE Int. Conf. on Image Processing(ICIP), Atlanta, Georgia, USA, pp. 157-160, 2006.

[9] A. Amer, "Memory-based spatio-temporal real-time object segmentation for video surveillance," in: Proc.

SPIE Int. Symposium on Electronic Imaging, Conf. on Real-Time Imaging VII, Santa Clara, CA, USA, pp. 10-21, 2003.

[10] S.Y. Chien, Y.W. Huang, B.Y. Hsieh, S.Y. Ma, L.G. Chen, “Fast video segmentation algorithm with shadow cancellation, global motion compensation, and adaptive threshold techniques,” IEEE Trans. Multimedia, Vol. 6, No. 5, pp. 732-748, 2004.

[11] O. Sukmarg, K.R. Rao, “Fast object detection and segmentation in MPEG compressed domain,” in: IEEE Proc. TENCON, Kuala Lumpur, Malaysia, pp. 364-368, 2000.

[12] D. Zhang, G. Lu, “Segmentation of Moving Objects in Image Sequence: A Review. Circuits Syst,” Signal Process, Vol. 20, No. 2, pp. 143-183, 2001.

[13] M.S. Atkins, B.T. Mackiewich, “Fully Automatic Segmentation of the Brain in MRI,” IEEE Trans. Med. Imaging, Vol. 17, No. 1, pp. 98-107, 1998.

[14] P.K. Saha, J.K. Udupa, “Optimum Image Thresholding via Class Uncertainty and Region Homogeneity,” IEEE Trans. Pattern Anal. Mach. Intell., Vol. 23, No. 7, pp. 689-706, 2001.

[15] N. Otsu, “A threshold selection method from gray-level histogram,” IEEE Trans. Syst. Man Cybern., Vol. 9, No. 1, pp. 62-66, 1979.

[16] H. Tian, S. K. Lam, T. Srikanthan, “Implementing Otsu’s thresholding process approximation unit using area-time efficient logarithmic,” in: Proc. Int. Symposium Circuits and Systems (ISCAS), Bangkok, Thailand, pp. IV-21-IV-24, 2003.

[17] K. Benkrid, D. Crookes, A. Benkrid, “Towards a general framework for FPGA based image processing using hardware skeletons,” Journal of Parallel Computing, Vol. 28, No.7-8, pp. 1141-1154, 2002.

[18] R.L. Rosas, A. de Luca, F.B. Santillan, “SIMD architecture for image segmentation using Sobel operators implemented in FPGA technology,” The 2nd International Conference on Electrical and Electronics Engineering (ICEEE) and XI Conference on Electrical Engineering, Mexico City, Mexico, pp. 77-80, 2005.

[19] Xue Li, Rongchun Zhao, Qing Wang, “FPGA based Sobel algorithm as vehicle edge detector in VCAS,” in: Proc. IEEE International Conference on Neural Networks and Signal Processing, Nanjing, China, pp. 1139-1142, 2003.

[20] P.Y. Hsiao, C.H. Chen, H. Wen, S.J. Chen, “Real-time realisation of noise-immune gradient-based edge detector,” IEEE Proceedings-Computers and Digital Techniques. Vol. 153, No. 4, pp. 261-269, 2006.

[21] M. Sezgin, B. Sankur, “Survey over image thresholding techniques and quantitative performance evaluation,” J. Electron. Imaging, Vol. 13, No. 1, pp. 146-165, 2004.

[22] A. Benedetti, A. Prati, N. Scarabottolo, “Image convolution on FPGAs: the implementation of a multi-FPGA FIFO structure,” Proceedings of the 24th Euromicro Conference, Vasteras, Sweden, pp. 123-130, 1998.

[23] B. Bosi, G. Bois, Y. Savaria, “Reconfigurable pipelined 2-D convolvers for fast digital signal processing,” IEEE Trans. Very Large Scale Integr. (VLSI) Syst., Vol. 7, No. 3, pp. 299-308, 1999.

[24] V. Muthukumar, D.V. Rao, “Image processing algorithms on reconfigurable architecture using HandelC,” Proceedings of the Euromicro Symposium on Digital System Design, Rennes, France, pp. 218-226, 2004.

[25] C.K. Lee, F.W. Choy, H.C. Lam, “Real-time thresholding using histogram concavity,” Proceedings of the IEEE International Symposium on Industrial Electronics, Xian, China, pp. 500-503, 1992.

[26] S.E. El-Khamy, M. Lotfy, N. El-Yamany, “A modified Fuzzy Sobel edge detector,” 7th National Radio Science Conference, Minufiya University, Egypt, pp. C32 1-9, 2000.

[27] Jianping Fan, Walid G. Aref, Mohand-Said Hacid, Ahmed K. Elmagarmid, “An improved automatic isotropic color edge detection technique,” Pattern Recogn. Lett., Vol. 22, No. 13, pp. 1419-1429, 2001.

[28] C.H. Li, P.K.S. Tam, “An iterative algorithm for minimum cross-entropy thresholding,” Pattern Recogn. Lett., Vol. 19, No. 8, pp. 771-776, 1998.

[29] Shu Yang, Ying Han, Cai-Rong Wang, Xiao-Wei Wang, “Fast selecting threshold algorithm based on one-dimensional entropy,” Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, China, pp. 4554-4557, 2005.

[30] R.C. Gonzalez and R.E. Woods, Digital image processing, 2nd edition, Prentice Hall, Upper Saddle River, New Jersey, 2002.

[31] URL: http://www.csse.monash.edu.au/hons/projects/2002/Laura.Frost/index.html

[32] N.S. Sargur, Center of Excellence for Document Analysis and Recognition (CEDAR), 1991. Available from:< http://www.cedar.buffalo.edu>

[33] D. Maltoni, D. Maio, A.K. Jain, S. Prabhakar, FVC2000 Fingerprint Database, 2000. Available from: http://bias.csr.unibo.it/fvc2000/default.asp

## Author Biographies

**Deng-Yuan Huang** received his Ph.D. degree in Aeronautic and Astronautic Engineering from National Cheng Kung University, Tainan, Taiwan, in 1994. He was with the steel and alumina R&D department at CSC Inc. in Taiwan for several years as an associate scientist specializing in process control in steel-making. He joined the Department of Electrical Engineering at Dayeh University in 2002 and is currently an assistant professor. He has published over 40 papers in journals and conference proceedings since 2002. His major research interests include FPGA chip design, image processing, pattern recognition, and computer vision.

**Da-Wei Lin** received his M.S. degree in computer science and information engineering in 2009 from Dayeh University, Changhua, Taiwan. Currently, he is working toward the Ph.D. degree in electrical engineering at the same university. His current research interests include image processing and computer vision.

**Wu-Chih Hu** received his Ph.D. degree in electrical engineering from the National Taiwan University of Science and Technology, Taipei, Taiwan, in 1998. From 1998, he worked at the National Penghu University of Science and Technology for 12 years. He is currently an associate professor in the Department of Computer Science and Information Engineering. He has published more than 70 papers in journals and conference proceedings since 1998. His current research interests include computer vision, image processing, pattern recognition, digital watermarking, visual surveillance, and video processing.

# An automated system for Health Care and Monitoring driven by intelligent agents

Dr S Ram Reddy[1], Raviprakash[2], Yogesh Karunakar[3], N T Markad[4]

[1]University of Mumbai phdr_reddy@sify.com
[2]University of Mumbai j_ravi54@rediffmail.com
[3]University of Mumbai askyogi@gmail,com
[4]GGSIP University nt_markad@yahoo.co.in

***Abstract:*** Logical health care system is designed to diagnose specific disease where input is sequence of symptoms. Symptom is associated with a numerical value 0-1 and a random mathematical formulated value will reveal disease. The diagnosed disease would have another unique numerical value. Different categories have made to put similar symptoms corresponding to particular disease. Role of intelligent agents comes from mathematical formulated inputs which on the later phase diagnose disease. Taking symptoms as data samples the HCS trains itself then after refinement and preprocessing with factual representation directs towards result as per as minimum error. The accuracy of result depends on the degree of training conducted on the data samples. Another issue which makes result having minimal error is as number of operations HMS performs with data samples; it becomes more consistent and generates result with less error and adequate feasibility**.** Here authors try to develop a logical system for health care to promote availability of medical experts in rural areas where demand is more but infrastructure is less.

***Key words:*** LHCMS, Symptoms, HMS, CDSS

## 1. Introduction

Many developed countries have announced initiatives to modernize their health care systems with investments in health information technology (IT). The goal of these initiatives is to use technology to improve the health care system by reducing costs, increasing patient safety and improving quality of care. Improving health care is a common goal for these countries, but there are wide disparities in the success with which nations have pursued this goal [1]. In particular, countries such as the United States have lagged behind some European nations in the adoption of health IT, such as electronic health records. Interoperable electronic health records are a prerequisite for a modern health care system and the key to delivering a number of benefits to health care patients and payers. For example, the computerized decision support systems used in hospitals provide patients the most benefit when they use a complete and accurate set of patient data. These systems can help ensure a return to the core principle of evidence-based medicine—that patients and doctors have the best evidence available when making a decision about treatment. While much attention has been paid to the degree to which nations have made progress with investment in health IT, less attention has been paid to the level of investment in health IT research. Yet evidence-based medicine relies on high quality medical research. Moreover, as we enter an increasingly digital world, the amount of health data that will be available to medical researchers will be increasing substantially. While past medical researchers had only a few limited data points recorded on paper on which to base their

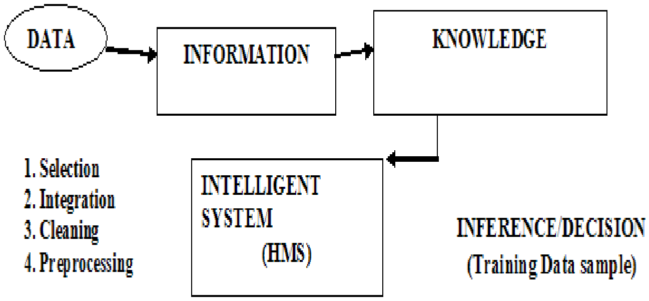


**Fig1 . Flow of data in LHCMS**

hypotheses, in the future researchers will have massive online databases containing terabytes of data for their analysis. Some of the major benefits from modernizing our health care system are expected to come from the improvements in medical research that it will enable. For example, medical researchers will be able to use rapid-learning health networks to determine the effectiveness of a particular treatment for a certain population or to discover

## 2. Background

### *2.1. Informatics in Health Care*

Health care is becoming an increasingly data-intensive field as doctors and researchers generate gigabytes of medical data on patients and their illnesses. While a patient visiting the doctor 20 years ago may have only generated a few data points— basic information such as weight, blood pressure, and symptoms—a medical encounter today may leave a long trail of digital data from the use of high-definition medical imaging to implantable or wearable medical devices such as heart monitors. More importantly, as doctors and hospitals transition away from paper medical records, this data is increasingly being collected and made available in an electronic format. The availability of large data sets of digital medical information has made possible the use of informatics to improve health care and medical research. Often referred to as "in silico" research, informatics offers a new pathway for medical discovery and investigation. Informatics focuses on developing new and better ways of using technology to process information. Today, informatics is being applied at every stage of health care from basic research to care delivery and includes many specializations such as bioinformatics, medical informatics, and biomedical informatics. Medical informatics, or clinical informatics, focuses on using information processing to improve health care delivery. It covers various applications including using information technology within the clinical setting for medical billing, patient and resource scheduling, and patient care. An example of medical informatics is the use of clinical decision support systems (CDSS) which provide feedback and instruction to health care workers at the point of care. Such a system may, for example, provide warnings of potential drug interactions to a prescribing doctor based on a patient's existing medical history and known allergies. By integrating patient information with harmful side-effects of a drug. While some of this research will occur in the private sector, for example through private pharmaceutical research, public investment in this area will also be important. Already a variety of projects offer a glimpse into the possibilities that IT will allow for future medical research. But achieving this vision will require substantial leadership and effort on the part of nations to overcome the technical and social hurdles ahead clinical guidelines, health care providers can help reduce medical errors. Adverse drug events alone account for an estimated 19 percent of injuries in hospitalized patients in the United States and cost hospitals over $2 billion per year, excluding medical malpractice expenses [5]. Biomedical informatics is a unique discipline that bridges multiple fields including medical research, clinical care and informatics. At its core, the objective of biomedical informatics is to develop new tools and technology to better collect, display, retrieve and analyze biomedical data. Such research can lead to new treatments, diagnostic tests, personalized medicine and better understanding of illnesses.

### 2.2. ***Building the Digital Platform for Medical Research***

Achieving this vision of an intelligent and fully-connected health care research infrastructure has not yet been realized. While various pilot projects have shown success and have demonstrated the potential benefits that can emerge from a ubiquitous deployment of informatics in health research, many technical obstacles still need to be overcome. These obstacles include making data accessible, connecting existing data sources, and building better tools to analyze medical data and draw meaningful conclusions. Much medical research data is not accessible electronically. For example, one challenge for the United States and the United Kingdom are the low rates of adoption of electronic health records among primary care providers and in hospitals. Electronic health records provide a complete medical history for a patient, including a full account of the patient's illnesses, treatments, laboratory results, medication history and known allergies. Among primary care providers, approximately one quarter use an EHR system in the United States and 89 percent use them in the United Kingdom. At hospitals, the rate of use is much lower with only about 10 percent or fewer of the hospitals in the United States and the

United Kingdom having adopted EHR systems [19]. Achieving the widespread use of electronic health records is a necessary requirement for creating the underlying data sets needed for biomedical informatics research. Access to the electronic health records of large populations will help researchers apply informatics to various problems including clinical trial research, comparative effectiveness studies, and drug safety monitoring. However, collecting medical data in electronic format is only the first step. Interoperability poses a substantial challenge for biomedical research. The vast amount of electronic medical data cannot fully be utilized by researchers because the data resides in different databases. Even when the organizations that collect and distribute biomedical data are willing to share data, incompatible data formats or data interfaces can create challenges for analyzing data across multiple data sets. As a result, researchers wishing to use multiple data sets must devote significant resources simply to managing the differences between the data and, as a result, have fewer resources available for working with the data [6]. For many years individuals in the research community have called for increased coordination and interoperability among data repositories to advance the use of informatics in health care. They have proposed various options to address interoperability although, to date, no proposal has achieved universal acceptance [6, 7]. One interim solution has been the development of online communities to share programming code to reduce the burden of working with diverse data sets. The most notable, Bio*, is a collection of open-source biomedical informatics projects that provide re-usable code for researchers to use that automate common computing tasks. For example, the project includes modular programming code to manipulate DNA sequences or combine data sets from different data sources [6].

## 3. A General Model

Healthcare applications have a number of additional requirements beyond the basic functions and representations that are common to many cognitive-system theories. (The “Related Work in Multi agent Healthcare Systems” sidebar describes four multi agent healthcare systems that, in different ways, illustrate these requirements.) On the basis of our experience with healthcare systems, we’ve identified three key requirements over and above the basic domino model:

• A communication capability for interactions between agents, which is important for multiagent systems but not supported by the formalisms proposed for modeling clinical guidelines, workflows, and so on.

• A well-developed model of decision making under uncertainty, which is generally regarded as fundamental to dealing with the complexities of clinical practice. Researchers have described how to embed this capability in an agent system and incorporate logical argumentation techniques for decision making.1, 6, and 7.

• The ability to access or communicate the knowledge and arguments used in specific decisions, a requirement that supports collaborative decision-making in multi agent applications. We extended the domino model to meet these requirements. Figure 1 shows the extended model. It’s built around the six basic entities of the original domino model, which can itself be seen as an extension of the BDI agent model. In our terminology, beliefs are aspects of the agent’s environment or its own mental states, which the agent holds to be true (that is, the agent will act upon them while they continue to hold). Goals are equated with “desires” and plans with “intentions.” We view intentions as commitments to new beliefs or to carrying out certain plans or pursuing new goals in the future.
A fundamental capability of the agent model is the ability to make decisions under uncertainty—that is, to make choices between competing beliefs or alternative plans given a lack of certain knowledge about the true state of the environment or about the consequences of possible actions on the environment. The model introduces a four-step decision procedure in which an agent can identify decision options (competing beliefs or plans), construct arguments for and against the options, assess the relative strength of the sets of arguments for alternative options, and commit to the most-preferred option.1,6 The decision procedure reflects our primary ASPIC project activity—namely, to develop an agent framework that can integrate the different roles of argumentation in a principled way. Two features of the extended model accommodate this activity:

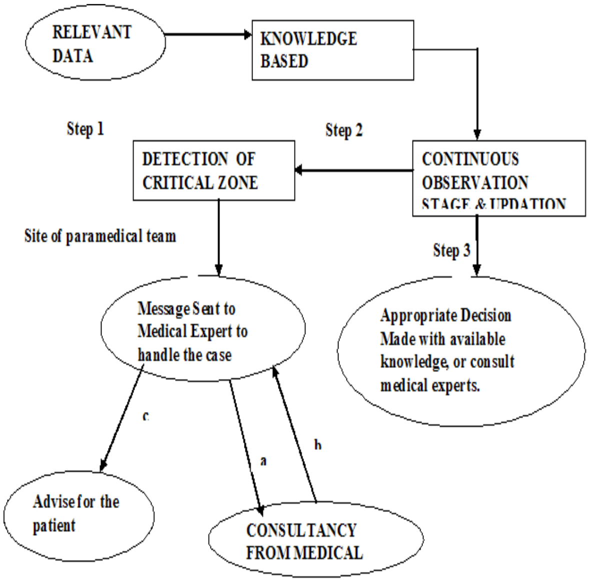

RELEVANT DATA
KNOWLEDGE BASED
Step 1
Step 2
DETECTION OF CRITICAL ZONE
CONTINUOUS OBSERVATION STAGE & UPDATION
Site of paramedical team
Step 3
Message Sent to Medical Expert to handle the case
Appropriate Decision Made with available knowledge, or consult medical experts.
c
b
a
Advise for the patient
CONSULTANCY FROM MEDICAL

**Fig 2. Building Blocks of LHCS**

***3.1 Interagent dialogue models.*** Project partners are developing and formalizing interagent dialogue models and we're incorporating the results into our extended model for use in the Carrel and CREDO applications described in the sidebar. eg we're extending standard FIPA-like per formatives to include those that facilitate coordination on collaborative tasks, such as joint decision making or service negotiation, where deliberative or dialectical argumentation between agents is required.

***3.2 Machine learning.*** Project partners are investigating the relationship between argumentation and machine learning. Learning capabilities are especially important in healthcare applications, because human errors and system failures will occasionally occur no matter how well we design our systems. To support learning from experience and corrections to procedures, the agent platform should on all occasions maintain records of what happened, what decisions were taken and why, and what the outcomes were.

## 4. Intelligent Agents

Authors try to define symptoms as intelligent agents which will further help to diagnose disease. Although medical science is not based on linear equation but the methodology which has adopted here is based on knowledge based and expected to secure result with minimal error. Authors categorically have taken three disease and listed corresponding symptoms. It can be seen on following tables as Table 1 disease H1N1, Table 2 Jaundice and in Table 3 Malaria. In each table four attributes have been taken as serial no., symptom code, symptom name and symptom value. The symptom value is assigned randomly to every symptom for uniqueness and it is between 0 to 1.

## 5. The Algorithm

As symptom value is assigned randomly to every symptom for uniqueness and it is between 0 to 1. The reason behind assigning these random values is to ultimately let every disease comes out with another unique numerical value between 0 and 1. The algorithm executes in following steps:

i. **Take symptoms with assigned numerical values one by one.**

ii. **Make clustering according to symptoms as unique symptom index and general symptom index.**

iii. **Prepare matrices for unique symptom index and general symptom index.**

iv. **Assign variables U to unique symptom index and G to general symptom index.**

v. **The dot product of U and G will give us the value of disease.**

The numerical value which comes out would be unique for the disease.

### 5.1. Methodology & Implementation:

At very first stage as a sample three diseases as well as their corresponding symptoms have been taken up in three tables. Authors try to device the cluster for three major diseases viz. H1N1 flu, jaundice and Malaria.

**Disease (H1N1). D1:**

| Sr. no | Symptom Code | Symptom Name | Symptom Value |
|---|---|---|---|
| 1 | S01 | fever | 0.1 |
| 2 | S02 | cough | 0.01 |
| 3 | S03 | sore throat | 0.19 |
| 4 | S04 | runny nose | 0.35 |
| 5 | S05 | body aches | 0.27 |
| 6 | S06 | headache | 0.9 |
| 7 | S07 | chills | 0.33 |

| | | | |
|---|---|---|---|
| 8 | S08 | fatigue | 0.45 |
| 9 | S09 | diarrhea | 0.6 |
| 10 | S10 | vomiting | 0.019 |

**Jaundice Disease D2:**

| Sr. no | Symptom Code | Symptom Name | Symptom Value |
|---|---|---|---|
| 1 | S11 | Yellow skin | 0.72 |
| 2 | S12 | Yellow eyes | 0.81 |
| 3 | S13 | reddish urine | 0.12 |
| 4 | S14 | Bronze skin | 0.39 |
| 5 | S15 | Loss of appetite | 0.01 |
| 6 | S16 | Furry tongue | 0.79 |
| 7 | S17 | Pale feces | 0.43 |
| 8 | S18 | Nausea | 0.075 |
| 9 | S19 | Itching skin | 0.16 |
| 10 | S20 | Lethargy | 0.15 |

**Malaria Disease D3:**

| Sr. no | Symptom Code | Symptom Name | Symptom Value |
|---|---|---|---|
| 1 | S21 | Fever | 0.91 |
| 2 | S22 | Rigors | 0.05 |
| 3 | S23 | Headaches | 0.49 |
| 4 | S24 | Myalgia | 0.55 |
| 5 | S25 | Loss of appetite | 0.12 |
| 6 | S26 | tiredness | 0.59 |
| 7 | S27 | vomiting | 0.23 |
| 8 | S28 | Nausea | 0.045 |
| 9 | S29 | cough | 0.14 |
| 10 | S30 | Enlarged liver/spleen | 0.39 |

Now the equation of the form is given by:

$$\mathbf{D} = \prod_{i=1}^{n} \mathbf{F(SDi,SOi)}$$

SDi = unique symptom index for disease D1
SOi = General symptom index for disease D1

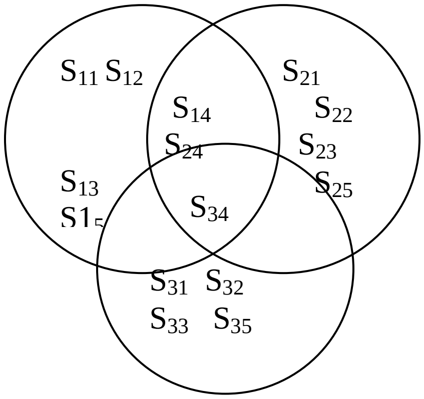
$$\mathbf{D1=} \begin{bmatrix} s11 \\ s12 \\ . \\ . \\ . \\ s1i \end{bmatrix} \mathbf{.\ [Si1\ Si2\ Si3\ .\ .\ .\ .\ .\ .Sin]}$$

$$\mathbf{D1=S11*Si1+ S12*Si2+S13*Si3+\ldots+S1n*Sin.}$$

$$\mathbf{D=}\sum_{i=1}^{n} \mathbf{S1i*Si1}$$

### 5.2. *Clustering:*

We try to train the system with the systems of disease and try to cluster the data. As an example



### 5.3. *Error Analysis:*

The difference of mathematical model of unique symptom index and General symptom index will refer for error. So if only unique symptom index is used for calculation of disease, general symptom index to be ignored hence more accurate and estimated value would come out for analysis.

## 6. Justification for Analysis:

While developing mathematical model first analysis starts with clustering reason is as we have two symptom indexes called U and G. All those symptoms which are core symptoms of corresponding disease are in U and all those symptoms which are also part of other disease are in G. eg Pale nail symptom can help us to diagnose on Jaundice but shivering can lead diagnosis towards malaria and pneumonia as well. So to considering both the indexes we need to develop mathematical model.

## 7. Intelligence through Machine Learning & Advantages

As initially the system is empty and works on knowledge based system. So when ever very first time user will interact with the system it will generate result with error but as many times system will be trained much consistent and efficient knowledge base will be created and system will produce more correct result. Although the implication is limited but adequate training will make it feasible. Following advantages come out with training of the system:

i. Enhancement of knowledge base
ii. Quick updation of u and G indexes.
iii. Domain enhancement of U & G indexes
iv. As system gets trained more accurate result comes out.
v. • Case-based learning
vi. Argumentation-based machine learning
vii. A general domain knowledge repository

## 8. Limitation of LHCMS:

As authors repeatedly saying the fact that medical science does not allow any conclusion which comes out through linear equation or linear analysis but this mathematical model which is designed for the logical health care and monitoring system (LHCMS) produces result with gradually degradation in error. LHCMS is limited upto diagnosis only. The system cannot diagnose disease where the symptom has different degrees of parameter. eg in the symptom shivering one needs to clarify the degree of shivering a patient has. Otherwise it would be considered as standard input shivering. So this kind of behavior of symptoms to be avoided to let LHCMS work within its domain.

## 9. Conclusion:

LHCMS is an IT aid to provide medical expertness within the limit and an enhancement through computing. The system justifies itself with core idea of IT ie connectivity With relevance and it reaches to public health center with complete solution since the data collection and processing is done with more accuracy and hence becomes intelligent system. The challenges and future focus of the project is exhaustive data analysis, where study requires more accurate data representation.

Because more relevant data analysis takes place more intelligent system would be designed. Finally future focus from IT point of view is linking among various attributes to conclude more complex cases. After this diagnosis stage is over authors look forward to design the system for treatment as futuristic approach of this paper.

## References:

[1]. S Rajendran,H Srimati,G Vadivu, 'GIS Based RHCMS', CSI journal,Vol.32, Issue 1,Apr2008.

[2]. Prakash R, Singh S K, "General Awareness & Development of interactive knowledge centers in villages" proceeding pp11-16 of NC technical development for content creation & localization in Indian languages ,Dec 2006, MHRD.

[3]. Health informatics article from www. en.wikipedia.org/wiki/Medical_informatics

[4]. Don de Savigny and Pandu Wijeyaratne. GIS for Health and Envieonment. 1995.

[5]. "Reporting of health statistics from service data:Country based practices". Publication from International workshop, Bankok oct'07.

[6]. Pandey N, Bhatia V, E-Health : A new health care perspective, pp B87-91 in proceeding of NCIT&A, New Delhi 2006.

[7]. Prakash R, Prasad D N, 'Design Model of Intelligent Health Monitoring System for Rural Areas' proceedings of ISTE Day 2008, national conference on Technology for rural India : Challenges and Perspectives.

[8]. D. Castro, "Explaining International Health IT Leadership," Information Technology and Innovation Foundation, Washington (2009).

[9]. "Connecting the Nation's Cancer Community," National Cancer Institute, National Institutes of Health, U.S. Department of Health and Human Services, NIH Publication No. 08-6363, January 2009, http://plan.cancer.gov/pdf/nci_2010_plan.pdf.

[10]. "The caBIG Pilot Phase, Report: 2003-2007," National Cancer Institute, National Institutes of Health, U.S. Department of Health and Human Services, November 2007.

[11]. K. Colbert, "Funding data for the NCICB," Personal communication with author to Office of Budget and Finance, National Cancer Institute. (July 2009).

[12]. D.W. Bates et al., "Effect of Computerized Physician Order Entry and

# Precise Stock Price Prediction System Using Neural Networks Trained by Enhanced and Meticulous Learning

Yogesh Karunakar[1], Abhijeet Gole[2], Priti Jha[3]

[1]University of Mumbai – askyogi@gmail.com
[2]Ruia College,University of Mumbai – abhijeet.mumbai@gmail.com
[3]Ruia College,University of Mumbai

***Abstract:*** In Most of the developing countries, investing in stocks, albeit the risk factor is the most lucrative way of earning quick bucks. This has lead to the development of various models for financial markets and investment. Black-Scholes model opened a new domain for research in the field of stock markets. The model develops partial differential equations whose solution, the Black–Scholes formula, is widely used in the pricing of European-style options. The Aim of "Neural Network Based Stock Price Forecasting Model" is to develop a Model which will be used to Forecast Future Stock Prices. It will be developed by using one of the Concepts in Artificial Intelligence [8], "Neural Networks". Network created for this Model trained and this Trained Network is tested for Prediction of the Future Stocks Prices. One Layer or Two Layer network is created and trained by using Backpropogation Algorithm in Neural Networks. Neural Networks are a class of Pattern Recognition Methods which have been successfully implemented in Data Mining and Prediction in a variety of fields

***Keywords*** – Black-Scholes model, Neural Networks, Stock markets, Backpropogation, Pattern recognition**.**

## 1. Introduction

**Neural Network Based Stock Price Forecasting Model:** It is basically a model which will be used for predicting future stock prices. It has been developed using Neural Network Concepts. Neural Network is one of the areas in Artificial Intelligence. A Neural Network is a powerful data modeling tool that is able to capture and represent complex input/output relationships. The motivation for the development of neural network technology stemmed from the desire to develop an artificial system that could perform "intelligent" tasks similar to those performed by the human brain.

Neural Networks resemble the human brain in the following two ways:

i) A Neural Network acquires knowledge through learning.

ii) A Neural Network's knowledge is stored within inter-neuron connection strengths known as synaptic weights.

The Advantage of Neural Networks lies in their ability to represent both linear and non-linear relationships and in their ability to learn these relationships directly from the data being modeled. With the neural networks' ability to learn nonlinear, chaotic systems, it may be possible to outperform traditional analysis and other computer-based methods.

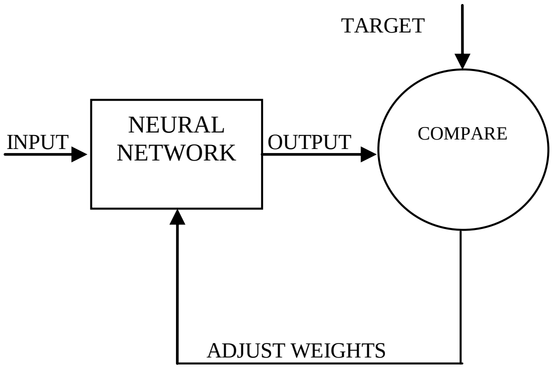


**Figure1. Block representation of neural network**

## 2. Neural Network

An Artificial Neural Network is a system based on the operation of biological neural networks, in other words, is an emulation of biological neural system. There are certain tasks that a program made for a common microprocessor is unable to perform. It's a system composed of a large number of basic elements arranged in layers and that are highly interconnected. The structure has several inputs and outputs, which may be trained to react (O's values) to the inputs stimulus (I's values) in the desired way. An artificial neuron is an element with inputs, output and memory that may be implemented with software or hardware. It has inputs (I) that are weighted added and compared with an Activation Function.

The General Form of Equation:

S =I1 * w1 + I2 * w2 + … + In * wn - b

O= f (S)

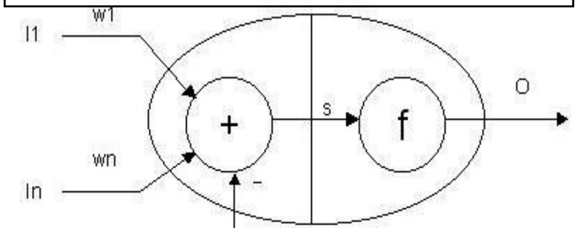
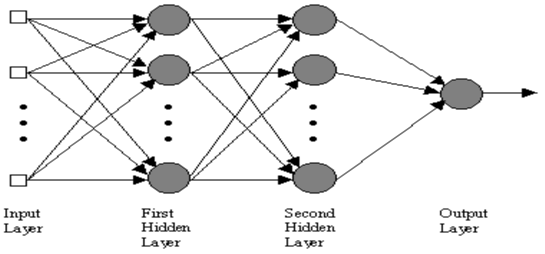


**Fig 2: Structure of Artificial Neuron**

The most common Neural Network Model is the Multilayer Perceptron (MLP). This type of Neural Network is known as a supervised network because it requires a desired output in order to learn. The goal of this type of network is to create a model that correctly maps the input to the output using historical data so that the model can then be used to produce the output when the desired output is unknown. A graphical representation of an MLP is shown below.



## **3.** Training Phase or Learning Phase

A **Neural Network** can be trained in the following ways.

1. Supervised Learning

2. Unsupervised Learning

### 3.1 Supervised Learning:

Supervised learning is the type of learning that takes place when the training instances are labeled with the correct result, which gives feedback about how learning is progressing. This is akin to having a supervisor who can tell the agent whether or not it was correct.

The training data consist of pairs of input objects (typically vectors), and desired outputs.

Supervised Learning is a Stochastic approximation of an unknown average error, It attempts to map an unknown function F: X → Y from observed training samples (x1,y1)….(xn, yn) by minimizing an Least Mean Square error.

### 3.2 Unsupervised Learning:

In Unsupervised learning, we are given some data *x* and the cost function to be minimized, that can be any function of the data *x* and the network's output, *f*. The cost function is dependent on the task (what we are trying to model) and our *a priori* assumptions (the implicit properties of our model, its parameters and the observed variables).Unsupervised Learning use unlabelled Pattern samples i.e., It doesn't use Target data. Other neural network systems use similar types of input data. Simpler systems may use only past share prices[7] or chart information[8]. A system developed by Yoon[5] based its input on the types and frequencies of key phrases used in the president's report to shareholders. Bergerson[6] created a system that traded commodities by training it on human designed chart information rather than raw data. Such directed training has the advantage of focusing the neural network to learn specific features that are already known as well as reducing learning time. Finally, the self-organizing system built by Wilson[9] used a combination of technical, adaptive (based on limited support functions), and statistical indicators as inputs. Determining the proper input data is the first step in training the network. The second step is presenting the input data in a way that allows the network to learn properly without overtraining. Various training procedures have been developed to train these networks.

## 4. Algorithm and Strategies Developed

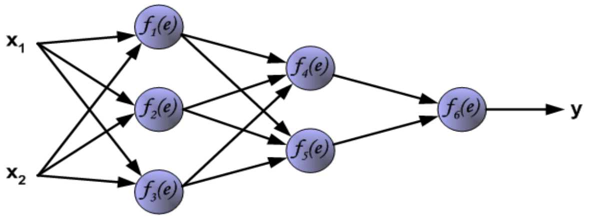
### 4.1 Training Algorithm

### 4.1.1 Back Propagation Algorithm:

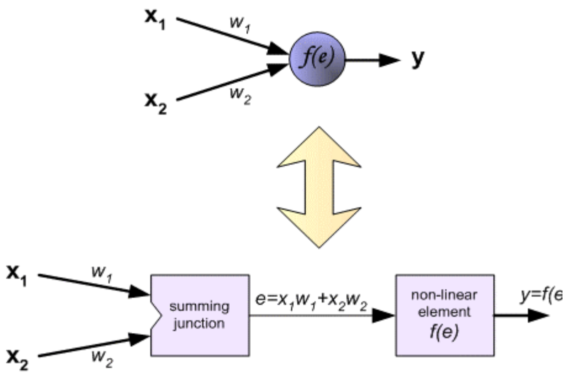
The back propagation algorithm trains a given feed-forward multilayer neural network for a given set of input patterns with known classifications. When each entry of the sample set is presented to the network, the network examines its output response to the sample input pattern. The output response is then compared to the known and desired output and the error value is calculated. Based on the error, the connection weights are adjusted. The back propagation algorithm is based on *Widrow-Hoff delta learning rule* in which the weight adjustment is done through *mean square error* [5]of the output response to the sample input. The set of these sample patterns are repeatedly presented to the network until the error value is minimized. Back Propagation Algorithm trains Multilayer Feed Forward Neural Network using Supervised Learning.

### 4.1.2 Graphical Representation of Steps of BackPropagation Algorithm:

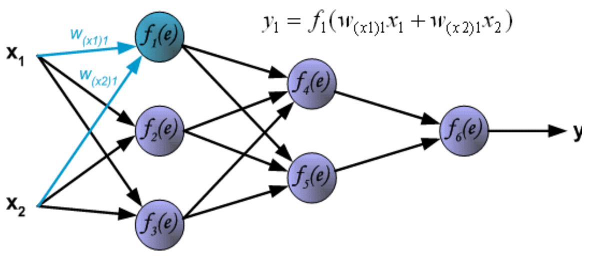
**Step1:**
It describes teaching process of multi-layer neural network employing *Back propagation* Algorithm. It is three layer neural network with two inputs and one Output.



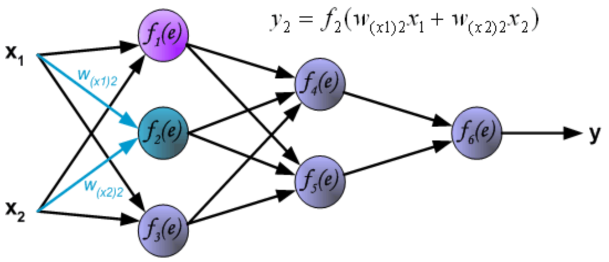
Each neuron is composed of two units. First unit adds products of weights coefficients and input signals. The second unit realizes nonlinear function, called neuron activation function. Signal *e* is adder output signal, and *y = f(e)* is output signal of nonlinear element. Signal *y* is also output signal of neuron.



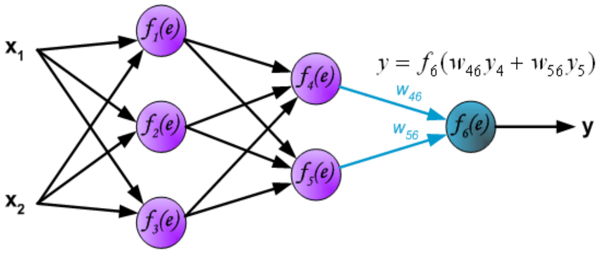
**Step2:**
To teach the Neural Network we need training data set. The training data set consists of input signals ($x_1$ and $x_2$) assigned with corresponding target (desired output) *z*. The network training is an iterative process. In each iteration weights coefficients of nodes are modified using new data from training data set.



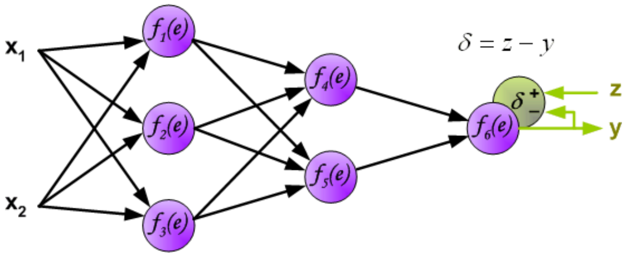
Similarly, for 2nd Neuron, It will be calculated as shown below. Each Neuron's Output will propagate through the Next Layer.



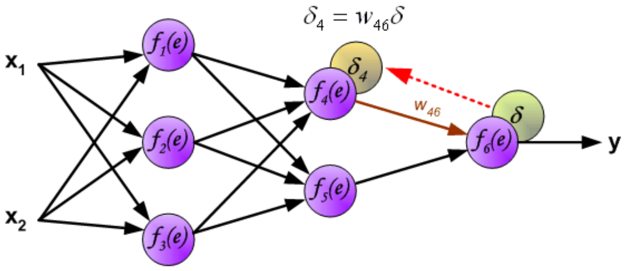
And for the Last Neuron, The final output is calculated as follows:

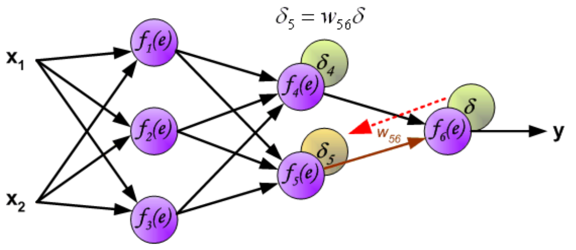


**Step3:**
In the next algorithm step the output signal of the network *y* is compared with the desired output value (the target), which is found in training data set. The difference is called error signal *d(delta)* of output layer neuron.

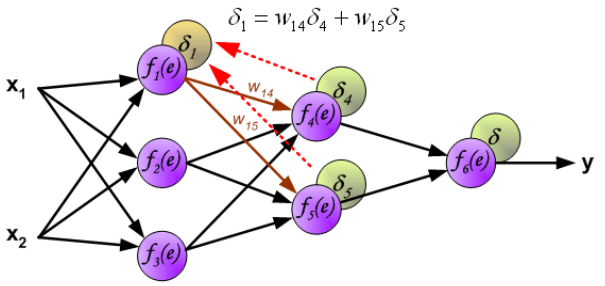


**Step4:**
The idea is to propagate error signal *d* (computed in single teaching step) back to all neurons, which output signals were input for discussed neuron.




$$\delta_5 = w_{56}\delta$$

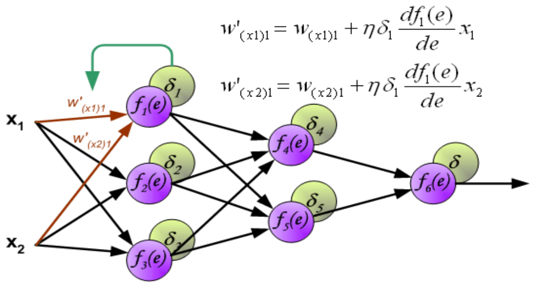
The weights' coefficients $w_{mn}$ used to propagate errors back are equal to this used during computing output value. Only the direction of data flow is changed (signals are propagated from output to inputs one after the other).


$$\delta_1 = w_{14}\delta_4 + w_{15}\delta_5$$

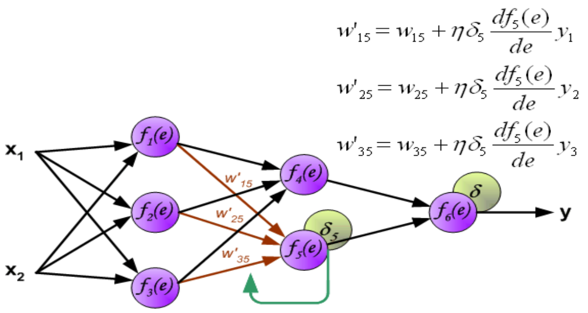
This technique is used for all network layers. If propagated errors came from few neurons they are added.

**Step5:**

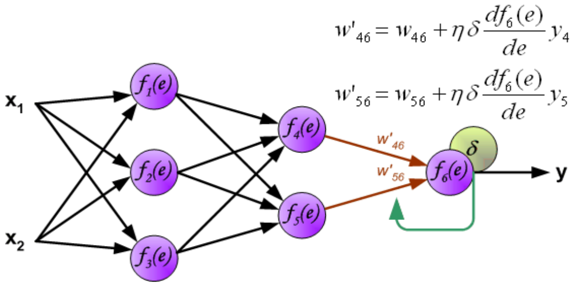
The Modified weight value is calculated for the first neuron as shown in the figure. When the error signal for each neuron is computed, the weights coefficients of each neuron input node may be modified. In Calculation shown below *df(e)/de* represents derivative of neuron activation function (which weights are modified).


$$w'_{(x1)1} = w_{(x1)1} + \eta\delta_1 \frac{df_1(e)}{de} x_1$$
$$w'_{(x2)1} = w_{(x2)1} + \eta\delta_1 \frac{df_1(e)}{de} x_2$$

Similarly, the weight values will be modified for the successive Neurons as Shown in the following figure.


$$w'_{15} = w_{15} + \eta\delta_5 \frac{df_5(e)}{de} y_1$$
$$w'_{25} = w_{25} + \eta\delta_5 \frac{df_5(e)}{de} y_2$$
$$w'_{35} = w_{35} + \eta\delta_5 \frac{df_5(e)}{de} y_3$$

After, all the weight values are modified, the whole procedure from step3 is repeated**.**


$$w'_{46} = w_{46} + \eta\delta \frac{df_6(e)}{de} y_4$$
$$w'_{56} = w_{56} + \eta\delta \frac{df_6(e)}{de} y_5$$

### 4.2 Formal Background

**CreateNetwork:** In this Module, An Artificial Neural Network of different layers is created. One Layer/Two Layer Network is created in Neural Network Toolbox.

Then a feed-forward Backpropagation network is created.

**Load Data:** In this Module, Input Data is loaded from the Database. A Preprocessing is performed on the input data. The inputs and targets are scaled so that they fall in the range [-1, 1].

**Training:** In this module, One Layer or Two Layer Network Created is passed through Training Phase. In Training Phase or Learning Phase, a given network is trained so as to make it learn different kinds of input which may appear in the future when it will be tested.

Finally, when the network is trained, it is simulated or tested with the inputs for which it will predict an output that will have minimum squared error (mse).

## 5. Implementation of the Model:

The Input Data Consists of 865 Tuples Containing Information such as Date, Open Price, High Price, Low Price, Close Price, Volume of Stocks.

**Neural Network Based Forecasting Model is implemented as shown:**

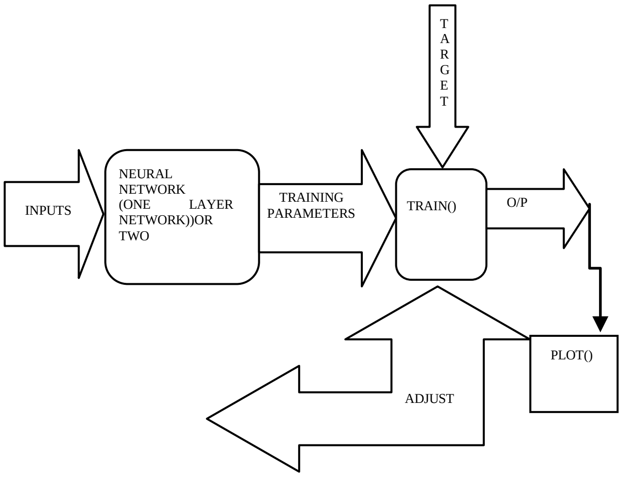
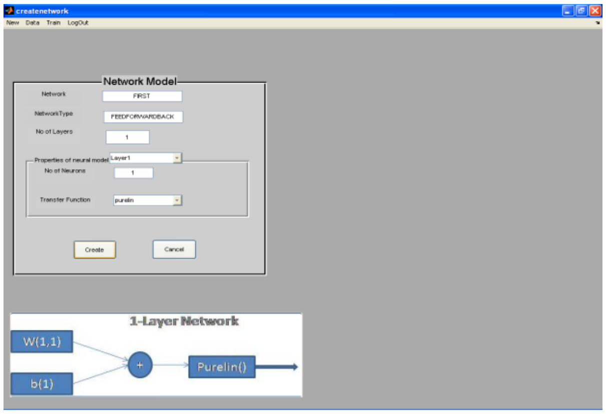


**Figure: Block diagram for system implementation**

## TestReport

| Date | Fall | Level |
|---|---|---|
| Apr 20,1992 | 572.08 | 3896.90 |
| May 6,1992 | 327.00 | 3560.72 |
| May 12,1992 | 336.68 | 3086.37 |
| Mar 31,1997 | 302.64 | 3360.89 |
| Apr 4,2000 | 361.48 | 4691.46 |

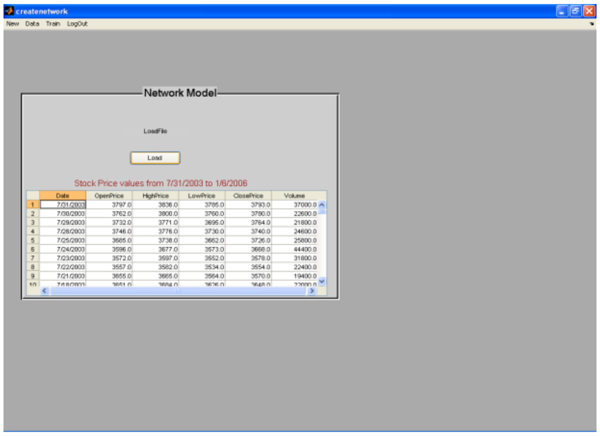
loading continues...



Network Model

Network: FIRST
NetworkType: FEEDFORWARDBACK
No of Layers: 1
Properties of neural model: Layer1
No of Neurons: 1
Transfer Function: purelin

**ONE LAYER NETWORK**



Network Model

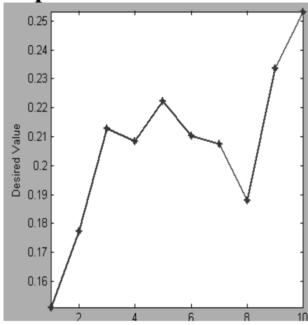
Stock Price values from 7/31/2003 to 1/6/2006

|  | Date | OpenPrice | HighPrice | LowPrice | ClosePrice | Volume |
|---|---|---|---|---|---|---|
| 1 | 7/31/2003 | 3797.0 | 3836.0 | 3785.0 | 3793.0 | 37000.0 |
| 2 | 7/30/2003 | 3762.0 | 3800.0 | 3760.0 | 3780.0 | 22600.0 |
| 3 | 7/29/2003 | 3732.0 | 3771.0 | 3695.0 | 3764.0 | 21800.0 |
| 4 | 7/28/2003 | 3746.0 | 3776.0 | 3730.0 | 3740.0 | 24600.0 |
| 5 | 7/25/2003 | 3685.0 | 3738.0 | 3662.0 | 3726.0 | 25800.0 |
| 6 | 7/24/2003 | 3596.0 | 3677.0 | 3573.0 | 3668.0 | 44400.0 |
| 7 | 7/23/2003 | 3572.0 | 3597.0 | 3552.0 | 3578.0 | 31800.0 |
| 8 | 7/22/2003 | 3557.0 | 3582.0 | 3534.0 | 3554.0 | 22400.0 |
| 9 | 7/21/2003 | 3655.0 | 3665.0 | 3564.0 | 3570.0 | 19400.0 |
| 10 | 7/18/2003 | 3651.0 | 3684.0 | 3626.0 | 3648.0 | 22000.0 |

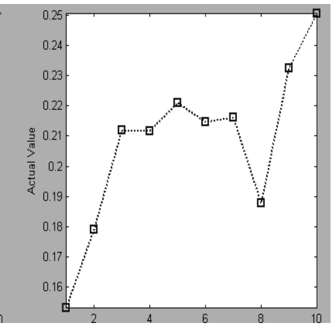
**INPUTDATA**

TestCaseNo: 1
TestUnit: Training Module (One Layer Network)
TestData:
Learning Rate=0.11,
Goal=0,
Momentum Unit=0.12,
TransferFunction='Purelin'.
Epochs=300.

**Expected Result**

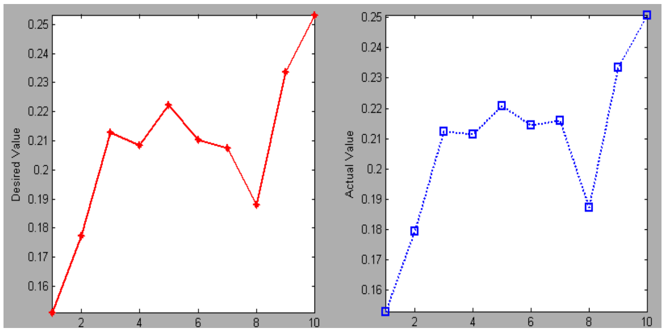


**Actual Result**



**TestPerformance: Good**

TestCaseNo: 2
TestUnit: Training Module(Two Layer Network)
TestData:
Learning Rate=0.10,
Goal=0,
Momentum Unit=0.12,
TransferFunction for 1st layer='Purelin'.
TransferFunction for 2nd layer='Purelin'.
Epochs=300.



**Expected Result** **Actual Result**

**TestPerformance: Very Good.**

## 6. Analysis and Conclusion:

With trainlm(), as one of the training functions , the Actual Output shown by the network is closer to the Desired Output. If one transfer function is 'tansig' and other transfer function is 'logsig' and traning function is 'trainlm', the Actual Output shown by the network shows much variations and thus not appropriate. In one layer Network, with transfer function as purelin and training function as 'trainlm' , the Output given by the Network approximates the Desired Output i.e, Minimum squared Error (MSE) is considerably low. In the Two Layer Network , with one of the transfer functions as 'tansig'/'logsig' and second transfer function as 'purelin', the output by the network falls within the range of the Desired Output , thus considerable.In the Two Layer Network , with one of the transfer functions as 'tansig'/'logsig' and second transfer function as 'purelin',learning rate=0.11 and momentum unit=0.12 ,the output by the network falls within the range of the Desired Output.

In the Two Layer Network , with one of the transfer functions as 'purelin' and second transfer function also as 'purelin', the output by the network falls exactly within the range of the Desired Output , thus making it very appropriate and (MSE) very low. With trainlm(), as one of the training functions, the Output shown by the network is appropriate, it nears the Desired Output, MSE is considerably low.

With traingd(),as one of the training functions ,the Output shows large variation , MSE high , thus not applicable for the problem under consideration.

## References

[1]Addison Wesley, Eugene Charniak, "Introduction to Neural Networks".

[2]Bart Kosko , "Neural Networks and Fuzzy Systesm" A Dynamic System Approach to Machine Intelligence, Prentice-Hall of India Pvt Ltd.,1992

[3]Margaret H.Dunham, "Data mining introductory and advanced topics, Dorling Kindersley(India) Pvt.Ltd.,2006.

[4]James Garson, "Intro to Connectionism" Introduction to science and theory of connectionism, 1997.

[5] Y. Yoon and G. Swales. Predicting stock price performance: A neural network approach. In *Neural Networks in Finance and Investing*, chapter 19, pages 329–342. Probus Publishing Company, 1993

[6] K. Bergerson and D. Wunsch. A commodity trading model based on a neural network-expert system hybrid. In *Neural Networks in Finance and Investing*, chapter 23, pages 403–410. Probus Publishing Company, 1993.

[7] G. Tsibouris and M. Zeidenberg. Testing the Efficient Markets Hypothesis with gradient descent algorithms. In *Neural Networks in the Capital Markets*, chapter 8, pages 127–136. John Wiley andSons, 1995.

[8] K. Kamijo and T. Tanigawa. Stock price pattern recognition: A recurrent neural network approach. In *Neural Networks in Finance and Investing*, chapter 21, pages 357–370. Probus Publishing Company,1993.

[9] C. L. Wilson. Self-organizing neural network system for trading common stocks. In *Proc. ICNN'94,Int. Conf. on Neural Networks*, pages 3651–3654, Piscataway, NJ, 1994. IEEE Service Center.

# Performance Comparison of Electronic Printwheel System by PI and PID Controller Using Genetic Algorithms

*Sobuj Kumar Ray[1] ,Diponkar Paul[2]
[1]International University of Business Agriculture and Technology
[2] World University of Bangladesh
Corresponding Addresses
Sobuj_kumar_ray@yahoo.com, dipo0001@ntu.edu.sg

***Abstract* -** PID controller is employed in every aspect of industrial automation. The application of PID controller extends from small industry to high technology industry. For those who are in heavy industries such as refineries and ship-buildings, working with PID controller is like a routine work. We would optimize the PID controller. The PID controller was tuned by using the classical technique that has been taught to us like Ziegler-Nichols method. We make use of the power of computing world by tuning the PID in a stochastic manner. In this work it is proposed that the controller be tuned using the Genetic Algorithm technique. Genetic Algorithms (GAs) are a stochastic global search method that emulates the process of natural evolution. Genetic Algorithms have been shown to be capable of locating high performance areas in complex domains without experiencing the difficulties associated with high dimensionality or false optima as may occur with gradient decent techniques. Using genetic algorithms to perform the tuning of the controller will result in the optimum controller being evaluated for the system every time. For this study, the model selected is an Electronic Printwheel control system. The PI and PID controller have been initially tuned for printwheel system by using a classical technique Ziegler Nichols (Z-N). The same model optimizes using the GA method. The results of both designs will be compared, analyzed and conclusion will be drawn out of the simulation made.

## 1. Introduction

A control system for operating a single high-speed print wheel is disclosed. The control system includes a print wheel derive motor having a coded disc with transparent portion forming a four-level Gray code for indicating the position of the print wheel. An optical electronic system is used for positioning the print wheel and also for the purpose of providing an electronic defend which uses the motive power of the print wheel drive for holding the print wheel in selected printing position. A similar optical and electronic system is used for controlling a paper advance mechanism. The disclosed system also includes a print hummer with voice clock coil type drive to permit positive, bidirectional hammer control. A recent development in typewriters and word processors is the print wheel printer, a machine that has become a significant factor in current office machine production. With this unit typefaces are mounted on spokes projecting from a central wheel. These machines produce high quality work with limited individuality. The mode of identification is significantly different from the work of a type ball machine and in some respects from the earlier type bar typewriters. Thus it is important to distinguish between the work of print wheel machines and other classes of typewriters. Consideration is given to individual identifying characteristics; the effects of the machine's ability to produce justified or flush right margins; and the general limitations on their identification. The aim of this project is to create a PID and PI controller for the Electronic Printwheel systems that is tuned using genetic algorithms. Most system is notoriously difficult to control optimally using a PID and PI controller because the system parameters are constantly changing. It is for this reason that genetic algorithms tuning strategy was applied. Genetic Algorithms are effective at finding high performance areas in large domains and are the ideal choice to tune the PID and PI controller. Genetic Algorithms were examined in detail, it was decided to create an objective function that evaluated the optimum PID and PI gains based on the controlled systems overall error. GA's outperformed standard tuning practices, e.g. Ziegler Nichols, at designing PID and PI controllers, in the tests carried out. According to a survey for process control systems conducted in 1989, more than 90 of the control loops were of the PID type. PID control has been an active research topic for many years. Since many process plants controlled by PID controllers have similar dynamics it has been found possible to set satisfactory controller parameters from less plant information than a complete mathematical model. These techniques came about because of the desire to adjust controller parameters in situ with a minimum of effort, and also because of the possible difficulty and poor cost bent of obtaining mathematical models. The two most popular PID techniques were the step reaction curve experiment, and a closed-loop "cycling" experiment under proportional control around the nominal operating point. It has been pointed out that PID controllers can be used only for plants with relatively small time delay. When the delay constant increases; the PID controller cannot guarantee good responses. In fact, apart from the traditional PID control structure, other control strategies may also be used to deal with such case. PID control consists of three types of control, Proportional, Integral and Derivative control [1].

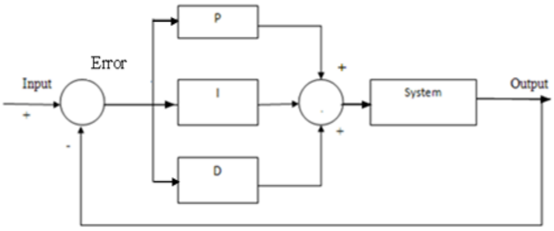


**Figure 1**. Schematic of PID Controller

The proportional controller output uses a 'proportion' of the system error to control the system. However, this introduces an offset error into the system.

$$P_{term} = K_p \times Error \quad (1)$$

The integral controller output is proportional to the amount of time there is an error present in the system. The integral action removes the offset introduced by the proportional control but introduces a phase lag into the system.

$$I_{term} = K_I \times \int Error\ dt \quad (2)$$

The derivative controller output is proportional to the rate of change of the error. Derivative control is used to reduce/eliminate overshoot and introduces a phase lead action that removes the phase lag introduced by the integral action.

$$d_{term} = K_D \times \frac{d(Error)}{dt} \quad (3)$$

The three types of control are combined together to form a PID controller with the transfer function:

$$C_{PID}(s) = \frac{K_D s^2 + K_P s + K_I}{s} \quad (4)$$

The PID controller is a "three mode" controller. That is, its activity and performance is based on the values chosen for three tuning parameters, one each nominally associated with the proportional, integral and derivative terms.
The block diagram of a closed–loop system with a PID controller in direct path [2] as shown in figure 2

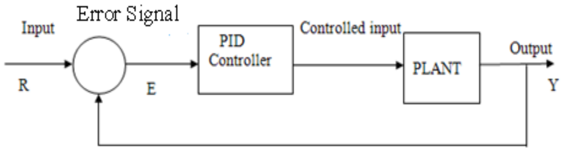


**Figure 2**. Block diagram of PID controller.

As the name suggests, the PI algorithm consists of two basic modes, the Proportional mode, and the Integral mode .When utilizing this algorithm it is necessary to decide which modes are to be used (Por I) and then specify the parameters (or settings) for each mode used. Generally, two basic algorithms are used P or PI.

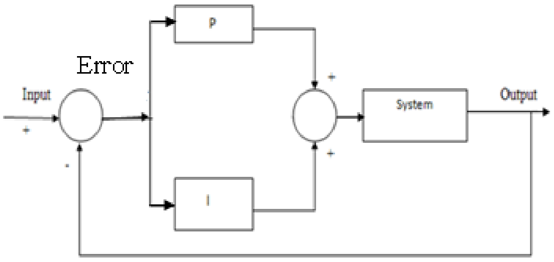


**Figure 3**. Schematic of PI Controller

The mathematical representation is,

$$\frac{mv(s)}{e(s)} = K_c \quad (5)$$

(Laplace domain) or $mv(t) = mv_{ss} + K_c(t)$ (time domain) the proportional mode adjusts the output signal in direct proportion to the controller input (which is the error signal, e). The adjustable parameter to be specified is the controller gain, $k_c$ .This is not be confused with the process gain, $K_p$ . The larger $k_c$ the more the controller output will change for a given error. For instance, with a gain of 1 an error of 10% of scale will change the controller output by 10% of scale. Many instrument manufacturers use Proportional Band (PB) instead of $k_c$ the time domain expression also indicates that the controller requires calibration around the steady-state operating point. This is indicated by the constant term $mv_{ss}$ . This represents the 'steady-state' signals for the mv and is used to ensure that at zero error the cv is at set point. In the Laplace domain this term disappears, because of the 'deviation variable' representation. A proportional controller reduces error but does not eliminate it (unless the process has naturally integrating properties), i.e. an offset between the actual and desired value will normally exist
The mathematical representation is,

$$\frac{mv(s)}{mv(t)} = K_C[1 + \frac{1}{T_I s}] \quad (6)$$

$$mv(t) = mv_{ss} + K_C[e(t) + \tfrac{1}{T_I s}\int e(t)dt] \quad (7)$$

The additional integral mode (often referred to as reset) corrects for any offset (error) that may occur between the desired value (set point) and the process output automatically over time. The adjustable parameter to be specified is the integral time ($T_i$) of the controller.

- Drive a certain distance in a straight line using encoders.
- Maintain position by maintaining a certain encoder tick count, this causes the motors to fight back against being pushed with the precise power output required.
- Driving tracking/shooting gimbals using the camera in the 2006 game.
- Maintain rotational velocity of impeller or collector wheels for balls by adjusting speed based on a target number of encoder ticks over time.
- Precision position of manipulators using encoders or potentiometers.

A method and apparatus for providing variable print hammer energy information and variable character spacing information for every character of every font in an electronic typing system having interchangeable print wheels, wherein the individual print wheels carry self-descriptive information in coded form on a read-only memory. The descriptive information is encoded on a portion of the print wheel and contains, in high density, machine readable, permanent form, and sufficient coded data to instruct the electronic typing system as to the optimal use of the particular font of characters contained on the type wheel. In a preferred embodiment, a single initializing revolution of the print wheel upon each insertion or machine start-up cycle serves to load the encoded data into a read/write memory for subsequent call-up and use by the electronic typing system during the print-out of each character. (Dingyu Xue,YangQuan Chem,and Derek P.Athenton, 2007) The data is serially encoded on the print wheel for reading by single track optional sensing apparatus. Alternate embodiments using parallel data tracks and magnetic sensing apparatus are discussed.

In control system position-control is quite common, that has variable load inertia. For example, the load inertia seen by the motor in an electronic printer wheel change when different printwheels are used. To illustrate the design of a robust system that is insensitive to the variation of the load inertia, consider that the forward path transfer function of a unity feedback control system [2]

$$G_P(s) = \frac{K_i K_b}{s[(Js+B)(Ls+B)+K_i K_b]} \tag{8}$$

The system parameters are

$K_i$ = motor torque constant =1N-m/A

$K_b$ = motor back emf constant =1v/rad/sec

R= motor resistant =1 Ω

L=motor inductance =0.01H

B= motor and load viscous-friction coefficient =0

J =motor and load inertia, varies between 0.02 and 0.02 N-m/rad/ sec $^2$

K =amplifier gain

Substituting these parameters into (8) we get

For $J = 0.01$

$$G_p(s) = \frac{10000K}{s(s^2+100s+10000)} \tag{9}$$

For $J = 0.02$;

$$G_p(s) = \frac{5000K}{s(s^2+100s+5000)} \tag{10}$$

## 2. Method

Genetic Algorithms (GA's) are a stochastic global search method that mimics the process of natural evolution. The genetic algorithm starts with no knowledge of the correct solution and depends entirely on responses from its environment and Evolution operators (i.e. reproduction, crossover and mutation) to arrive at the best solution. By starting at several independent points and searching in parallel, the algorithm avoids local minima and converging to sub optimal solutions. In this way, GAs have been shown to be capable of locating high performance areas in complex domains without experiencing the difficulties associated with high dimensionality, as may occur with gradient decent techniques or methods that rely on derivative information [5]. A genetic algorithm is typically initialized with a random population consisting of between 20-100 individuals. This population (mating pool) is usually represented by a real-valued number or a binary string called a chromosome. For illustrative purposes, the rest of this section represents each chromosome as a binary string. How well an individual performs a task is measured is assessed by the objective function (D. E. Goldberg, 1989). The objective function assigns each individual a corresponding number called its fitness. The fitness of each chromosome is assessed and a survival of the fittest strategy is applied. In this project, the magnitude of the error will be used to assess the fitness of each chromosome. There are three main stages of a genetic algorithm; these are known as reproduction, crossover and mutation. Genetic Algorithms provides an adaptive searching mechanism inspired on Darwin's principle of the fittest. It is invented by John Holland of the university of Michigan, after David Goldberg gave a basic idea of GA in his book '' Genetic Algorithm Search, Optimization and Matching Learning'' .GA is a search and optimization techniques inspired by biological process namely. 'natural' selection' and natural genetics'. GA starts with no knowledge of correct solution and depends on responses from its environment. GA manipulates not just one potential solution to a problem, but a collection of potential solution, called a population. The potential solution in population is called 'individuals' or 'chromosomes', each of them is associated to a fitness value. The chromosomes are subjected to an evolutionary process which takes several cycles. Basic operations are selection, reproduction, crossover, and mutation. During crossover some reproduced individuals cross and exchange their genetic characteristics and such crossover create new chromosomes from the existing one s in the population. The selection mechanism for parent chromosomes takes the fitness of parent into account, ensuring that the better solution have a higher chance to procreate and donate their beneficial characteristics to their offspring. Newly generated individuals in time replace the existing ones. Through this process after a while the population will converge to a 'best' solution.

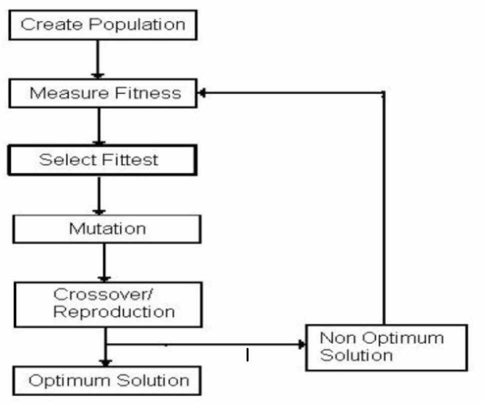


**Figure 4**. Graphical Illustration the Genetic Algorithm Outline

Whom it points is selected. This continues until the selection criterion has been met. The probability of an individual being selected is thus related to its fitness, ensuring that fitter individuals are more likely to leave offspring. Multiple copies of the same string may be selected for reproduction and the fitter strings should begin

More complex crossover techniques exist in the form of Multi-point and Uniform Crossover Algorithms. Multi-point crossover is an extension of the single point crossover algorithm and operates on the principle that the parts of a chromosome that contribute most to its fitness might not be adjacent. There are three main stages involved in a Multi-point crossover.

1. Members of the newly reproduced strings in the mating pool are 'mated' (paired) at random.
2. Multiple positions are selected randomly with no duplicates and sorted into ascending order.
3. The bits between successive crossover points are exchanged to produce new offspring.

Example: If the string 11111 and 00000 were selected for crossover and the multipoint crossover positions were selected to be 2 and 4 then the newly created strings will be 11001 and 00110 as shown in Figure 5.

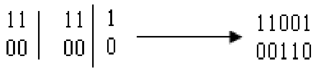


**Figure 5.** Illustration of a Multi-Point Crossover

In uniform crossover, a random mask of ones and zeros of the same length as the parent strings is used in a procedure as follows.
1. Members of the newly reproduced strings in the mating pool are 'mated' (Paired) at random.
2. A mask is placed over each string. If the mask bit is a one, the underlying bit is kept. If the mask bit is a zero then the corresponding bit from the other string is placed in this position.

Example: If the string 10101 and 01010 were selected for crossover with the mask
10101 then newly created strings would be 11111 and 00000 as shown in Fig. 6.

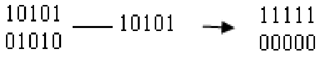


**Figure 6**. Illustration of a Uniform Crossover

Uniform crossover is the most disruptive of the crossover algorithms [4] and has the capability to completely dismantle a fit string, rendering it useless in the next generation. Because of this Uniform Crossover will not be used in this project.
The following schematic diagram illustrates the three types of children.

### 2.1 Mutation

Using selection and crossover on their own will generate a large amount of different strings however there are two main problems with this:

1) Depending on the initial population chosen, there may not be enough diversity in the initial string to ensure the GA searches the entire problem space.
2) The GA may converge on sub-optimum string due to a bad choice of initial population.

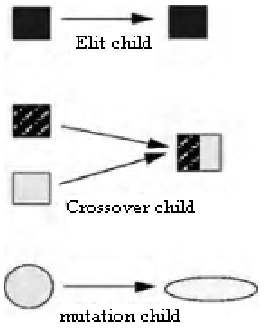


**Figure 7**: Schematic diagram illustrates the three types of children

This problem may be overcome by the introduction of a mutation operator into the GA. Mutation is the occasional alteration of a value of a string position. It is considered a back ground operator in the genetic algorithm.
The probability of mutation is normally low because a high mutation rate would destroy fit of
Mutation and degenerate the genetic algorithm into a random search. Mutation probability values of around 0.1% to 0.01% are common, these values represent the probability that a certain string will be selected for mutation, that is for a probability of 0.1%: one string in one thousand will be selected for mutation ( C. R. Houck, J. Joines. and M.Kay., 1996). Once a string is selected for mutation, a randomly chosen element of the string is changed or 'mutated'.
For example, if the GA chooses bit position 4 for mutation in the binary string 10000, the resulting string is 10010 as the fourth bit in the string is flipped as shown in Figure 8

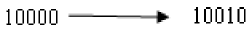
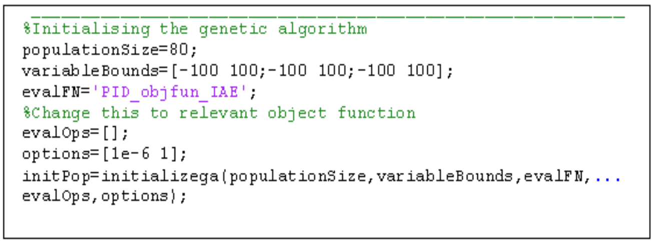


**Figure 8.** Illustration of Mutation Operation

### 2.2 Elitism

With crossover and mutation taking place, there is a high risk that the optimum solution could be lost as there is no guarantee that these operators will preserve the fittest string. To counteract this, elitist models are often used. In an elitist model, the best individual from a population is saved before any of these operations take place. After the new population is formed and evaluated, it is examined to see if this best structure has been preserved. If not, the saved copy is reinserted back into the population. The GA then continues on as normal [6] Fig 9 Initializing the Population of the Genetic Algorithm
The following code is based on the Genetic Algorithm
Optimization Toolbox
(GAOT) [3].

- ***PopulationSize*** **-** The first stage of writing a Genetic Algorithm is to create a population. This command defines the population size.



```
%Initialising the genetic algorithm
populationSize=80;
variableBounds=[-100 100;-100 100;-100 100];
evalFN='PID_objfun_IAE';
%Change this to relevant object function
evalOps=[];
options=[1e-6 1];
initPop=initializega(populationSize,variableBounds,evalFN,...
evalOps,options);
```

**Figure 9**. Codes Initializing the Population of a Genetic Algorithm

- ***VariableBounds* -** Since this project is using genetic algorithms to optimize the gains of a PID controller there are going to be three strings assigned to each member of the population, these members will be comprised of a P, I and a D string that will be evaluated throughout the course of the GA. The three terms are entered into the genetic algorithm via the declaration of a three-row variablebounds matrix. The number of rows in the variablebounds matrix represents the number of terms in each member of the population. Figure 9 illustrates a population of eighty members being initialized with values randomly selected between -100 and 100.

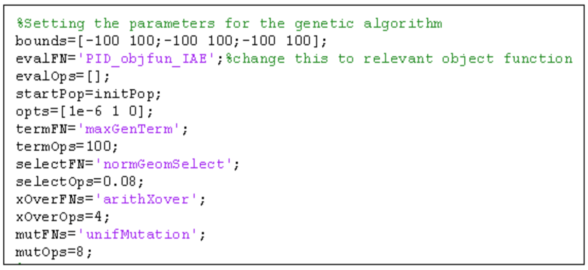
· ***EvalFN* -** The evaluation function is the declaration of the file name containing the objective function.

· ***Options* -** Although the previous examples in this section were all binary encoded,
this was just for illustrative purposes. Binary strings have two main drawbacks:
1. They take longer to evaluate due to the fact they have to be converted to/from binary.
2. Binary strings lose precision during conversion.
As a result of this and the fact that they use less memory, real (floating point) numbers will be used to encode the population. This is signified in the options command in Figure 9, where the '1e-6' term is the floating point precision and the '1' term indicates that real numbers are being used (0 indicates binary encoding is being used).

· **Initialisega** - This command combines all the previously described terms and creates an initial population of 80 real valued members between –100 and 100 with 6 decimal place precision.

## 2.3 Initializing the Population Genetic Algorithm

A genetic algorithm is initialized as shown in Figure 10.



```
%Setting the parameters for the genetic algorithm
bounds=[-100 100;-100 100;-100 100];
evalFN='PID_objfun_IAE';%change this to relevant object function
evalOps=[];
startPop=initPop;
opts=[1e-6 1 0];
termFN='maxGenTerm';
termOps=100;
selectFN='normGeomSelect';
selectOps=0.08;
xOverFNs='arithXover';
xOverOps=4;
mutFNs='unifMutation';
mutOps=8;
```

**Figure 10**. Initializing the Genetic Algorithm

***Bounds*** - The bounds for the genetic algorithm to search within are set using this command. These bounds may be different from the ones used to initials the population and they define the entire search space for the genet algorithm. **startPop -** The starting population of the GA, 'startPop', is defined as the population described in the previous section, i.e. 'initPop', see Figure 4.12.opts - The options for the Genetic Algorithm consist of the precision of the string values i.e. 1e-6, the declaration of real coded values, 1, and a request for the progress of the GA to be displayed, 1, or suppressed, 0.· ***TermFN*** - This is the declaration of the termination function for the genetic algorithm. This is used to terminate the genetic algorithm once certain criterion has been met. In this project, every GA will be terminated when it reaches a certain number of generations using the 'maxGenTerm' function. This termination method allows for more control over the compile time (i.e. the amount of time it takes for the genetic algorithm to reach its termination criterion) of the genetic algorithm when compared with other termination criteria e.g. convergence termination criterion. · ***TermOps* -** This command defines the options, if any, for the termination function. In this example the termination options are set to 100, which mean that the GA will reproduce one hundred generations before terminating. This number may be altered to best suit the convergence criteria of the genetic algorithm i.e. if the GA converges quickly then the termination options should be reduced. · ***SelectFN* -** Normalized geometric selection ('normGeomSelect') is the primary selection process to be used in this project. The GAOT toolbox provides two other selection functions, Tournament selection and Roulette wheel selection. Tournament selection has a longer compilation time than the rest and as the overall run time of the genetic algorithm is an issue, tournament selection will not be used. The roulette wheel option is inappropriate due to the reasons mentioned in section. · ***SelectOps* -** When using the 'normGeomSelect' option, the only parameter that has to be declared is the probability of selecting the fittest chromosome of each generation, in this example this probability is set to 0.08. · *XOverFN* - Arithmetic crossover was chosen as the crossover procedure. Single point crossover is too simplistic to work effectively on a chromosome with three alleles, a more uniform crossover procedure throughout the chromosome is required. Heuristic crossover was discarded because it performs the crossover procedure a number of times and then picks the best one. This increases the compilation time of the program and is undesirable. The Arithmetic crossover procedure is specifically used for floating point numbers and is the ideal crossover option for use in this project. · *XOverOptions* -This is where the number of crossover points is specified. In the example shown in Figure 5.8, the number of crossovers points is set to four.. The 'multiNonUnifMutation', or multi non-uniformly distributed mutation operator, was chosen as the mutation operator as it is considered to function well with multiple variables.. The mutation operator takes in three options when using the'multiNonUnifMutation' function. The first is the total number of mutations, normally set with a probability of around 0.1%. The second parameter is the maximum number of generations and the third parameter is the shape of the distribution. This last parameter is set to a value of two, three or four where the number reflects the variance of the distribution.5.9 performing the Genetic Algorithm The genetic algorithm is compiled using the command shown in Figure 4.13. Once this command is entered, the genetic algorithm will iterate until it fulfills the criteria described by its termination function. Writing an objective function is the most difficult part of creating a genetic Algorithm. An objective function could be created to find a PID controller that gives the smallest overshoot, fastest rise time or quickest settling time but in order to combine all of these objectives it was decided to design an objective function that will minimize the error of the controlled system. Each chromosome in the population is passed into the objective function one at a time. The chromosome is then evaluated and assigned a number to represent its fitness, the bigger its number the better its fitness. The genetic algorithm uses the chromosome's fitness value to create a new population consisting of the fittest members.

## 3. Simulation Result

Print wheel control system for resetting and setting a plurality of print wheels, said control system having means for generating timing pulses in accordance with the rotation of the print wheels, a pulse signal distributor for generating control signals associated with successive characters on each of the print wheels, and a selector network and controls responsive to the control signals for arresting the movement of each of the print wheels, depending upon the character of each of the print wheels to be selected.

### 3.1 Development Electronic Print Wheel System

To aid with the development of this project a system was chosen at random and a PID and PI controller was designed for it using conventional methods. A genetic algorithm was then created to evaluate the PID and PI coefficients of the same system and the results of the two techniques were compared. The system was selected as position control (Electronic print wheel system) [2] system is of order three.
The system chosen was:

$$G(s) = \frac{5000}{s^3 + 100s^2 + 5000s} \qquad (11)$$

### 3.2 Ziegler-Nichols Designed PID and PI Controller

The Ziegler-Nichols tuning method using root-locus was the 'conventional' method used to evaluate the PID and PI gains for the system. Using the 'rlocus' command in Matlab,The crossover point and gain of the system were found to be j71.1 and 101 respectively, as shown in Figure 11

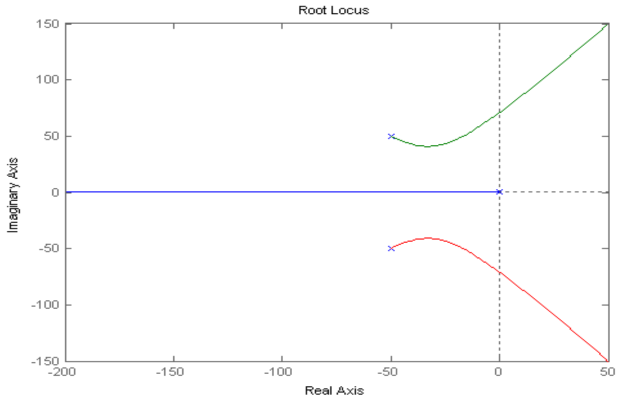


**Figure 11.** Plot of root locus for $G(s)$

With a frequency ( $\omega_c$ ) of 1rad/s the period $T_c$ is calculated as,

$$T_c = \frac{2\pi}{\omega_c} \text{ Sec} \qquad (12)$$

**Table 1.** Ziegler-Nichols PID and PI Tuning Parameters Gives

| Controller | $K_P$ | $T_I$ | $T_D$ |
|---|---|---|---|
| PID | $0.6K_c$ | $\frac{T_c}{2}$ | $\frac{T_c}{8}$ |
| PI | $0.45K_c$ | $\frac{T_c}{1.2}$ | |

**Table 2.** Ziegler-Nichols PID Tuning Values

| Controller | $K_P$ | $T_I$ | $T_D$ |
|---|---|---|---|
| PID | 60.60 | 0.044 | 0.011 |
| PI | 31.99 | 0.073 | |

Using the relationship $K_I = \frac{K_p}{T_I}$ and $K_D = K_P T_D$, the PID and PI gain can be evaluated.

**Table 3.** Ziegler-Nichols PID and PI Gain values

| Controller | $K_P$ | $K_I$ | $K_D$ |
|---|---|---|---|
| PID | 60.60 | 1377.27 | 0.66 |
| PI | 31.99 | 434.39 | |

Table 3 shows the PID and PI gain values for the system G(s). A genetic algorithm, Initial PID GA. Initial PI GA. was created to evaluate the optimum PID and PI gain values for the system G(s). A number of objective functions were created in order to evaluate the PID and PI values chosen by the Genetic Algorithm.

Comparisons of Steady Step Response of Integral of Time Multiplied by Absolute Error (ITAE)

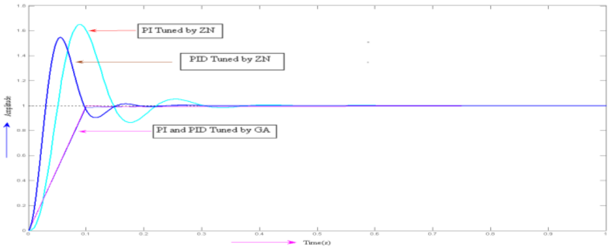


**Figure 12.** Step Response of Integral of Time Multiplied by Absolute Error (ITAE)

Comparisons of Steady Step Response Integral of Absolute Magnitude of the Error (IAE)

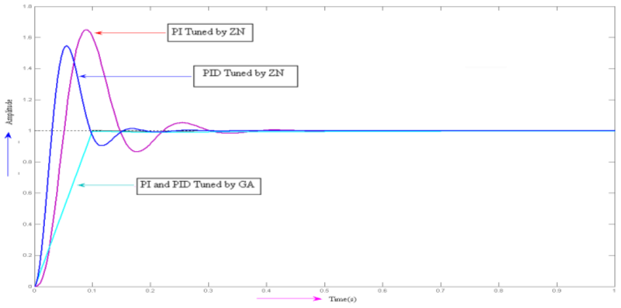


**Figure 13.** Step Response Integral of Absolute Magnitude of the Error (IAE)

Comparisons of Steady Step Response Integral of the Square of the Error (ISE)

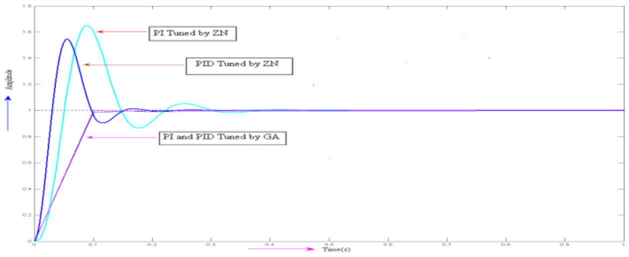


**Figure 14**. Step Response Integral of the Square of the Error (ISE)

Comparisons of Steady Step Response Mean of the Square of the Error (MSE)

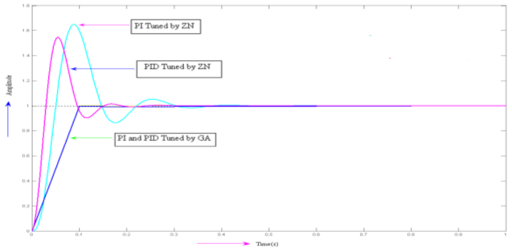


**Figure15.** Step Response Mean of the Square of the Error (MSE)

## 4. Discussion

Figure 12 to 15 shows Ziegler-Nichols designed PID and PI controller Vs GA designed PID and PI Controller using ITAE, IAE, ISE and MSE as performance Criterion. Under the conditions of this experiment, it can be seen that the IAE Objective functions performs having a smaller rise time, smaller Overshoot and smaller settling time than the other PI controllers (Ms Jennifer Bruton, 2003). Again the ITAE Objective functions performs having a smaller rise time, smaller Overshoot and smaller settling time than other PID controllers. Each of the genetic algorithm-tuned PID and PI controllers outperforms the Ziegler-Nichols tuned controller in terms of rise time, overshoot and settling time. Each of the PID controller performance is more than PI in terms of rise time, overshoot and settling time. The ITAE objective function was chosen as the primary performance criterion for the remainder of this project due to its smaller settling time and smaller overshoot than any other method in conjunction with a slightly faster compile time due to there being just one multiplication to be carried after the error has been calculated. This is coupled with the fact that ITAE has been a proven measure.

## 5. Conclusion and discussion

It was established that the steady state characteristics of GA's outperformed standard tuning practices when designing a PID and PI controller. It was determined that the Steady Step Response of Integral of Absolute Magnitude of the Error (**IAE)** performance criterion based objective function produced the most effective PI controllers when compared with other performance criterion i.e. MSE, ITAE and ISE. Again Integral of Time Multiplied by Absolute Error **(ITAE)** performance criterion based objective function produced the most effective PID controllers when compared with other performance criterion i.e. IAE, MSE, and ISE. It was proved by comparison of their steady state characteristics that PID outperformed standard tuning practices than PI controller. When testing the genetic algorithm component, it was discovered to frequently produce controllers that made the overall controlled system unstable. The exact cause for this could not be determined. To rectify this problem, the genetic algorithm was modified so it would analyze the controller it evaluates. If the controller produces an unstable system, it is replaced with the last stable controller evaluated by the genetic algorithm. The genetic algorithm online tuned PID controller proved to be a capable controller. Adequate testing of the controller could not be performed due to the simulation difficulties mentioned previously. It is illustrative of the fact that we have worked on a transfer function which is fixed for printwheel system parameter. But in real life its parameter is not constant. Again we have used only step response but ramp response, impulse etc could be used as the remarkable scope for the future work on it. This process can be applied to many other control systems such as ball and hoop control system, sun seeker system etc.

## References

[1] Ms Jennifer Bruton , Ian Griffin" On-line PID Controller Tuning using Genetic

Algorithms" 2003.

[2] Linear Feedback Control by Dingyu Xue,YangQuan

Chem,and Derek P.Athenton

Chapter 6,Copyright @ 2007

[3] U.S Patent March 14,1978 Sheet 7 of 7 4,078,485

Print wheel control John Gilkeson Guthrie

http://patents?id=hjIrAAAAEBAJ&printsec=abstract&zoom=4&source=gbs_overview_r&cad=0#v=onepage&q=&f=false

[4] Automatic Control systems, 7th Ed.

By Benjamin C.Kuo

[5] O' Mahony, T., Downing,C.J. and Klaudiuz, F.,

'Genetic Algorithms for PID

Parameter Optimisation: Minimising Error Criteria',

[online], URL:

http://www.pwr.wroc.pl/~i-8zas/kf_glas00.pdf

[6] D. E. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley Publishing Co., Inc., 1989

[7] C. R. Houck, J. Joines. and M.Kay. A genetic algotithm for function optimisation: A Matlab implementation. ACM Transactions on Mathematical Software, 1996, [Online], URL: http://www.eos.ncsu.edu/eos/service/ie/research/kay_res/GAToolBox/gaot

## First Author's Biography

Mr. Sobuj Kumar Ray was born in Bogra, Bangladesh in 1987. Mr. Ray received his Bachelor degree in Electrical and Electronic Engineering from the Rajshahi University of Engineering and Technology (RUET), Rajshahi, Bangladesh in April 2010. Now he is a faculty in the department of Electrical and Electronic Engineering, Internal University of Business Agriculture and Technology (IUBAT), Uttara,Dhaka, Bangladesh(www.iubat.edu). The major fields of study of Mr. Ray comprise control system and power system.

## Second author's Biography

Mr. Diponkar Paul is currently working as Assistant Professor in the department of Electrical and Electronic engineering at World University Bangladesh. After passing his master degree from March 2008 he was serving as Assistant Professor, EEE at Bangladesh University upto July 2010. He is having qualifications: B.Sc. Engg., DISM (software engineering), M.Sc. Engg. His research interests are in the area of energy conversions, power system modeling and advanced control theories covering the application of IT. From 0ct 2004 to July 2006, he was working as Lecturer in department of computer science and engineering at Pundra University of science & technology, Bogra. In Singapore during his master degree at Nanyang technological university, he was involved in financial service operation integrated to IT system administration jobs from Dec 2006 to February 2008.

# Architecture of CORBA based P2P-Netpay Micro-payment System in Peer to Peer Networks

Kaylash Chaudhary and Xiaoling Dai

School of Computing, Information and Mathematics Science
The University of the South Pacific, Laucala Campus, Suva, Fiji
chaudhary_k@usp.ac.fj[1] , dai_s@usp.ac.fj[2]

***Abstract:*** Micro-payment systems have the potential to provide non-intrusive, high-volume and low-cost pay-as-you-use services for a wide variety of web-based applications. We proposed a new model, P2P-NetPay, a micro-payment protocol characterized by off-line processing, suitable for peer-to-peer network service charging. P2P micro-payment systems must provide a secure, highly efficient, flexible, usable and reliable environment, the key issues in P2P micro-payment systems development. Therefore, in order to assist in the design and implementation of an efficient micro-payment system suitable for P2P networks, we describe prototype architecture for a new CORBA based micro-payment model. We present an object-oriented design and describe a prototype implementation of P2P-NetPay for file-sharing P2P system. We compare socket, CORBA, RMI and web-service-based P2P-NetPay prototypes and outline directions for future research in P2P micro-payment implementations.

***Keywords***: Micro-payment system, P2P-NetPay, CORBA, Web services

## 1. Introduction

File sharing in peer to peer networks has been an important issue in past years. Many file sharing application were developed for peers to exchange content but this introduced the problem of “free-rider” within the network where peers only consume services rather than provide [1]. Micro-payment system was introduced to liberate “free-rider” problems and escalate efficient use of the peer to peer network. Various micropayment system were established which suffered from the problems of lack of scalability, dependence on online brokers, communication overhead and security [9, 10, 11, 12]. We proposed a new micropayment system, P2P-Netpay [6], which addressed afore mentioned distress for payment system.

Software architecture is the structure of the system, which comprises of software components, the externally visible properties of those components, and the relationship between those components. It is a path for communication among the elements which captures early design decisions for systems. Software architecture plays an important role in the system development. P2P-Netpay software architecture design should be scalable, reliable, secure and flexible.

This paper focuses on architecture of CORBA based P2P-Netpay micropayment system implementing Broker and peers using CORBA. Earlier implementation of P2P-Netpay used java sockets for peer communication [5]. The design and implementation of CORBA system will be discussed.

Web Service is an emerging standard for business to business (B2B) communication with different platforms to handle wide variety of clients [2]. We have also developed our Broker using web service which provides payment service to peers. The peers are implemented using either CORBA or socket for communication between them but they consume services offered by Broker using web service.

The four architectural styles (CORBA, socket, Remote Method Invocation (RMI) and web service) for P2P-Netpay are discussed in this paper with the comparison amongst them. We outline our plans for further research and development in P2P micro-payment system.

## 2. Motivation

In any software development, architecture plays an important role in achieving design and business goals, quality solution and reusable or extendable solution [3]. One such situation is based on P2P-Netpay protocol implemented via CORBA for peer to broker communication and sockets for peer to peer communication [4, 5].

The Application Programming Interface (API) for socket programming is low-level since it creates communication overheads while two peer applications communicate for transferring of files. It was noted during design/implementation of P2P-Netpay that the programmer was responsible for method of communication. For example, in P2P-Netpay if a peer user wants to send ecoins to peer vendor to buy files, a peer user must establish a socket connection with peer vendor and write to socket while peer vendor reads from a socket. Thus, sockets require more instructions to be executed each time a message is sent or received.

When a file is transferred from peer vendor to peer user, it is read in portions and those small portions are sent to peer user while others are being read. Therefore, a series of messages are sent and received during file transfer. Series of messages sometimes leads to download being aborted. This is due to connection timeout. Each time a peer needs to communicate with other peer, a connection must be established and it must be terminated when finished.

Socket has certain advantages in network programming. Due to P2P-Netpay [6] protocol requirements, every file download deals with ecoins therefore a proper architecture is needed for providing service to peer user.

To unfetter, above concerns using socket, three other methods were considered for development which are RMI, CORBA and Web Services.

## 3. Overview of P2P-Netpay

P2P-Netpay, which is a basic offline protocol suitable for micro-payments in peer to peer networks is discussed in [4, 5, 6]. Below are some of the briefly discussed micropayment terminologies:

- ***Payword Chain*** – A payword chain is represents a set of E-coins in the P2P-Netpay system.
- ***E-coin*** – An “e-coin” is a payword element and the value of a payword e-coin might be one-cent but could be some other value.
- ***E-wallet*** – An “e-wallet” is used to store e-coins and send e-coins to a vendor paying for information goods.
- ***Touchstone*** – A “touchstone” is used to verify the e-coins whether the ecoins are valid or not.
- ***Index*** – An “index” is used to indicate the current spent amount of each e-coin (payword) chain. For example if you have spent 2cs to buy an information goods, the current index value is 3.

The customer registers and purchases some e-coins (using macro-payment such as credit card) with the Broker/CIS site. Peer-users buy e-coins from Broker/CIS which is sent to peer-user’s e-wallet. When buying items from peer-vendor, the peer-user sends e-coins from e-wallet. The peer-vendor verifies the e-coin and allows peer-user to download file. At the end of the day, peer-vendor can redeem e-coins with Broker/CIS for real money. When a peer-user first tries to spend an e-coin the peer-vendor communicates with the Broker/CIS to obtain a validating touchstone for the coin.

Each e-coin encodes a “payword chain” which utilizes a fast hashing function to provide the next valid coin in the chain each time a coin is spent. When a peer-user downloads a file from another peer-vendor, the new peer-vendor obtains the touchstone and index from previous peer-vendor. If the previous peer-vendor is offline then the new peer-vendor contacts Broker/CIS for touchstone and index.

The transfer of e-coins from Broker/CIS to peers is secured by public key encryption. An index is used to indicate the amount of e-coin spent so far which prevents peer-users from double spending and peer-vendors from over debiting [7]. The peer-user and peer-vendor does not reveal identities to any third party or each other. Only the secure Broker/CIS can identify the participants in a particular transaction. In P2P-Netpay, the peer-user needs to contact the Broker/CIS to buy e-coins when e-coins run out.

## 4. P2P-Netpay Architecture

We have developed a revised software architecture for implementing CORBA based P2P-Netpay micro-payment systems for content sharing in peer-to-peer networks. The interaction between Broker/Central Index Server (CIS), Peer User and Peer Vendor is illustrated in Figure 1.

CORBA is a standard developed by the Object Management Group (OMG) to support distributed object computing [13] which uses Broker/CIS as an intermediary to handle request in a system and separates components interface from its implementation. CORBA is not a language, it is middleware platform. It provides infrastructure for the programming of distributed systems using C, C++, Smalltalk, Ada and Java. Object Request Broker (ORB) forwards operations on objects to the desired object and returns results to the client. These inter-ORB interactions authorize a peer user to communicate with peer vendor or peer user/vendor to communicate with Broker or vice versa.

The CORBA Interface Definition Language (IDL) describes the objects together with their methods and attributes which is independent from the languages in which these interfaces will actually be implemented. There are two IDL files in P2P-Netpay, one for implementation of Broker/CIS operations and other for peer operation. Peers use the Broker/CIS operation to buy ecoins, request Touchstone and redeem ecoins for real money whereas peer operation IDL is used for downloading files from peers. Broker/CIS operation IDL is compiled to generate code for both the peer and the Broker/CIS. The generated code for peers is in the form of object stubs. From the peer's perspective these stubs are function calls directly into the object. In actuality the stubs forward the request to the remote object via the ORB. The compiler also generates skeleton code for the Broker/CIS (typically referred to as a servant in CORBA), which needs to be fleshed-out with the implementation of the requested operations. The same happens for peer operation IDL file.

The P2P-Netpay Broker/CIS system is built on top of the multi-tier web-based architecture presented as follows:

Client tier (HTML Browser): The browser communicates with the web server which runs the JSPs to register peers.

- Web tier (Broker/CIS Web Server and JSPs): In the web tier in the systems, Java Server Pages (JSPs) and JavaBeans are used to service the web browser clients, process request from the clients and generate dynamic content from them. After receiving the client request, the JSPs request information from a JavaBean. The JavaBean can in turn request information from an application server (CORBA). Once the JavaBean generates content the JSPs can query and display the Bean’s content. The Broker/CIS keeps track of online peers. It also stores the file names with the host and port of peers. Broker/CIS is designed using CORBA in Java so that it can serve multiple clients at one time.

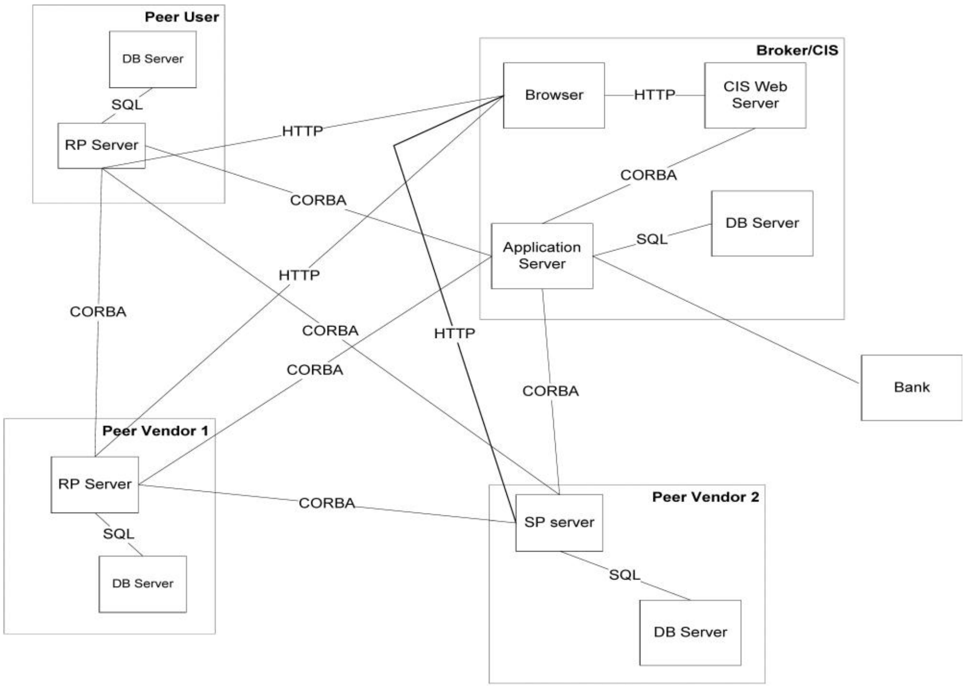


Figure 1: P2P-Netpay interactions

The peer user registers with Broker/CIS site using Java Server Pages (JSP) in web server which communicates through application server using CORBA to register and allow software download. When buying a file, ecoins are transferred using CORBA from peer user to Peer Vendor 1 which in turn requests Touchstone from either Peer Vendor 2 or Broker/CIS depending on where peer user has spent that ecoin before this transaction.

- Application Server tier (CORBA): In the P2P-Netpay Broker/CIS, CORBA is used as the middleware for the application server, which is implemented in the Java language that has a CORBA IDL mapping.
- Database Server tier: On the back-end of the system we use Ms Access to implement the databases accessed via a Java Database Connectivity (JDBC) interface. JDBC, which is a multi-database application programming

interface, provides Java applications with a way to connect to and use relational databases. When a Java application interacts with a database, JDBC can be used to open a connection to the database and SQL code is sent to the database.

The peer system which is either user or vendor is a three tier architecture which includes:

- Client tier (Client Application): This client application communicates with peer server to get requests from other connected peers or CIS/Broker.
- Peer Server tier (Requesting peer server/ Supplying peer server): It is implemented in Java to handle the functionalities such as sharing files, communicating with server, checking balance, redeeming, searching Broker/CIS and browsing peers. This server also uses CORBA IDL mapping for communication between peers.
- Database Server tier: We use Ms Access to implement the database accessed via JDBC. Only the Java application can interact with the database. Peers cannot open the database manually and edit any data since this database is password protected.

## 5. P2P-Netpay Implementation

The three main functionality of P2P-Netpay are buying ecoins, downloading file and redeem spending.

- ***Buy Ecoins***

Figure 2 shows how a peer buys e-coin in the P2P-Netpay system. The peer checks the amount left in the e-wallet and if wishing to buy e-coin, peer enters the amount and clicks buy button. The client application requests the e-coins through CORBA to Broker/CIS application server. The Broker/CIS application server debits from the peer's credit card, stores e-coins in the database and sends an e-wallet to java application (client application) in peer's computer.

- ***Downloading File***

Figure 3 shows how a peer-user downloads file using P2P-Netpay micro-payment system. After browsing peers or searching the Broker/CIS, peer-user clicks on download popup menu on the title of the file name. The client application sends the request including file name and e-coins through CORBA interface to peer-vendor server (PVS). If the touchstone and index does not exist in its database, the server contacts other Peer Vendor Server (PVS) or Broker/CIS in order to obtain touchstone and index. If the e-coins are valid, the PVS stores in redeem database and sends the file to peer-user. Peer-user server (PUS) than debits the e-coin.

- ***Redeem Spending***

Figure 4 illustrates how a peer-vendor redeems spent e-coins with Broker/CIS. When redeeming spent e-coins, peer-user clicks on redeem button on the client application where the PVS aggregates all payments and sends to Broker/CIS application server using CORBA interface where it verifies spent e-coins and sends the balance to peer-user.

The basic user interfaces are presented in Figure 5. A peer can be a user or a vendor. To use the services of P2P-Netpay, a peer must be registered with the broker and download the application. A peer must remember the Peer ID generated by the system to login as described in Figure 5 (1). For the initial login the peer server connects to Broker/CIS server to get the Peer ID and password using CORBA after which it is stored in the local database. If the login is successful, the main interface of the application appears as in Figure 5 (2). The IP address and port is listed for the peers who are currently online. These IP and port is not to which peer is listening to as in sockets but its the *RemotePeerManagerServer* which provides peers to invoke methods remotely. CORBA has its own servers running to which peers can connect and use the services. There are much functionality provided such as checking balance, searching file, uploading file and redeeming.

Suppose a peer user would like to search for file named “ewallet”, user enters the name in (3) and the results are displayed in (4). The results are obtained from Broker/CIS application server. It will only display results of the peer who is currently online. To download a file, right click on the file name and select download from popup menu. The application will connect to peer hosting that particular file using CORBA and download starts after ecoin verification.

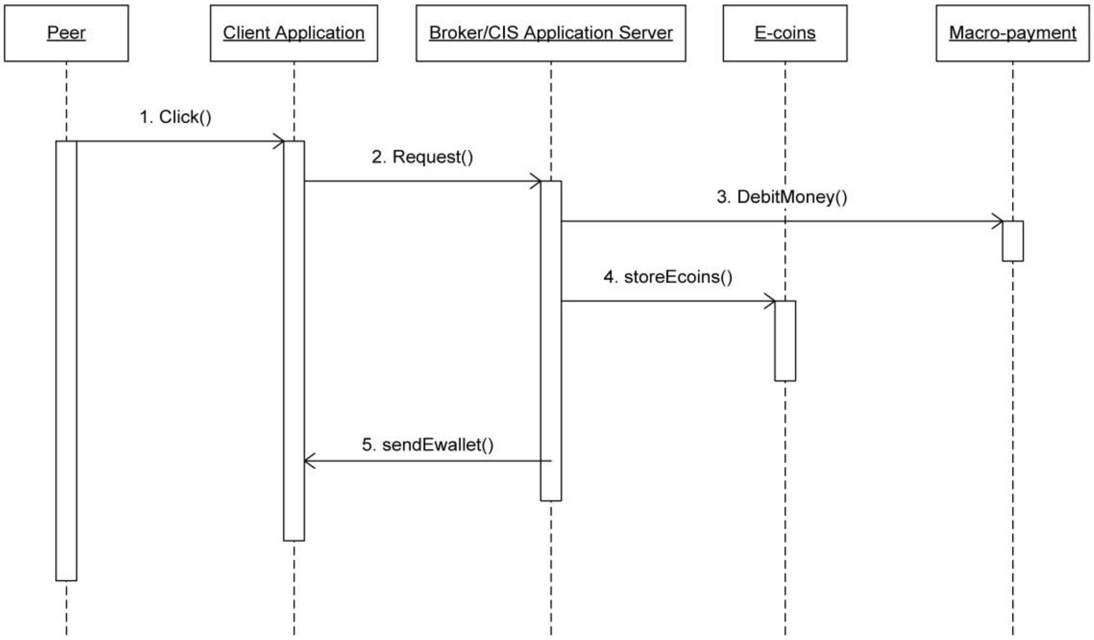


Figure 2: Buy ecoin sequence diagram

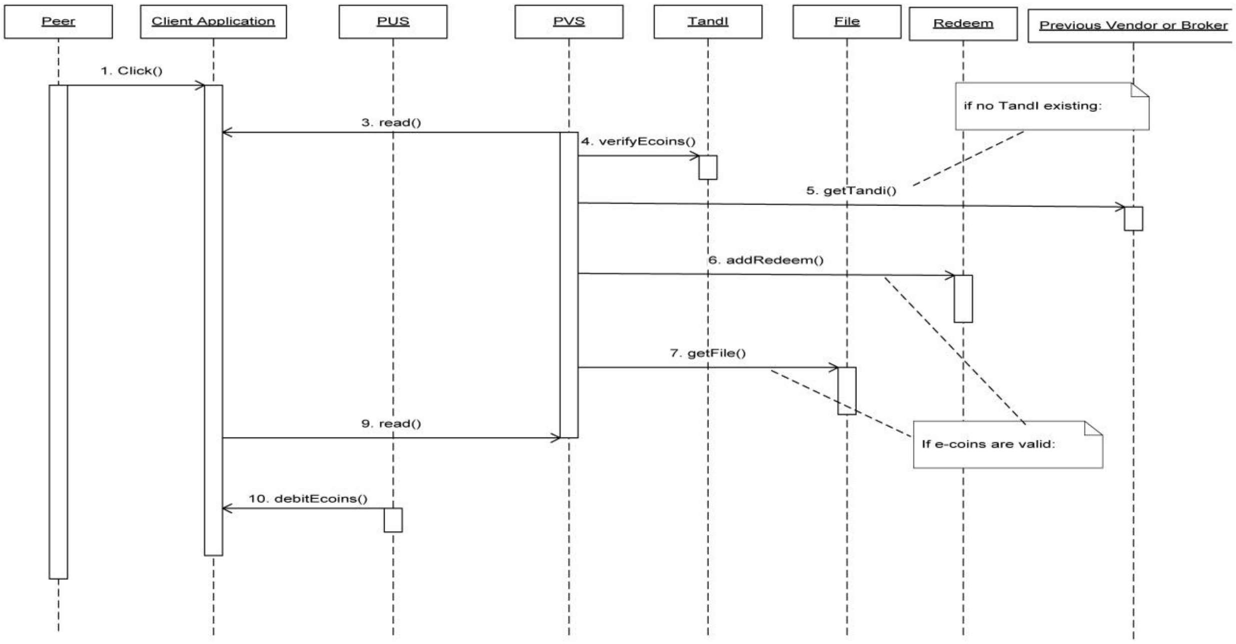


Figure 3: Download file sequence diagram

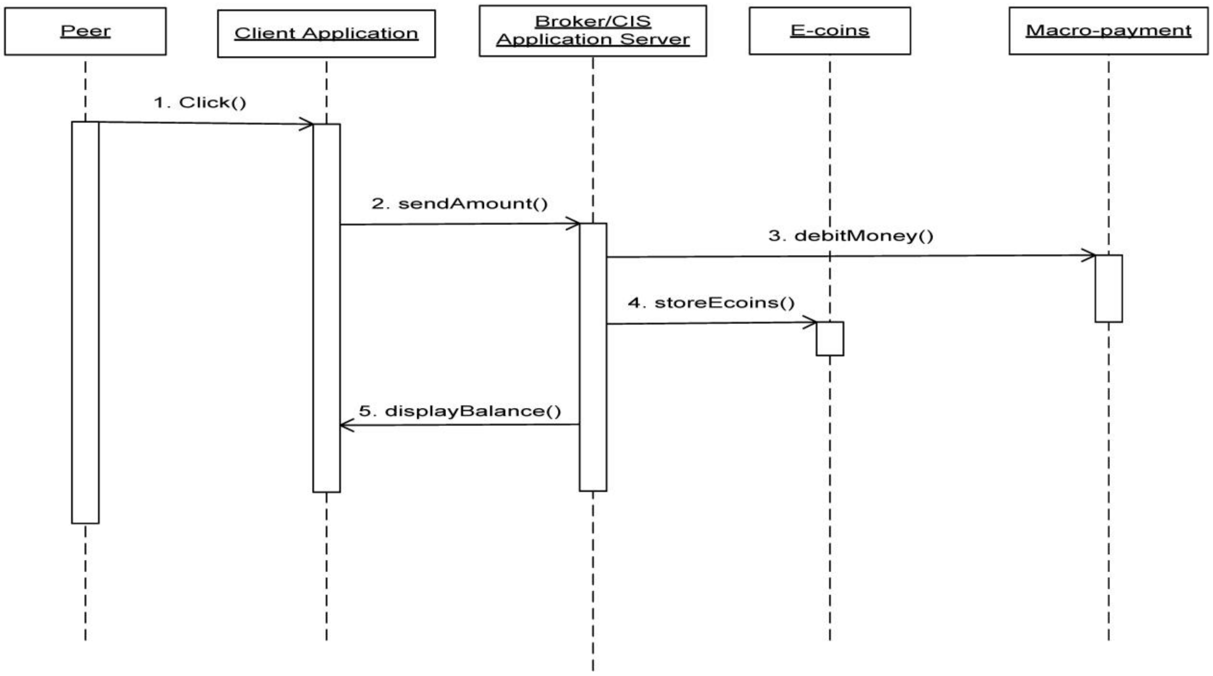
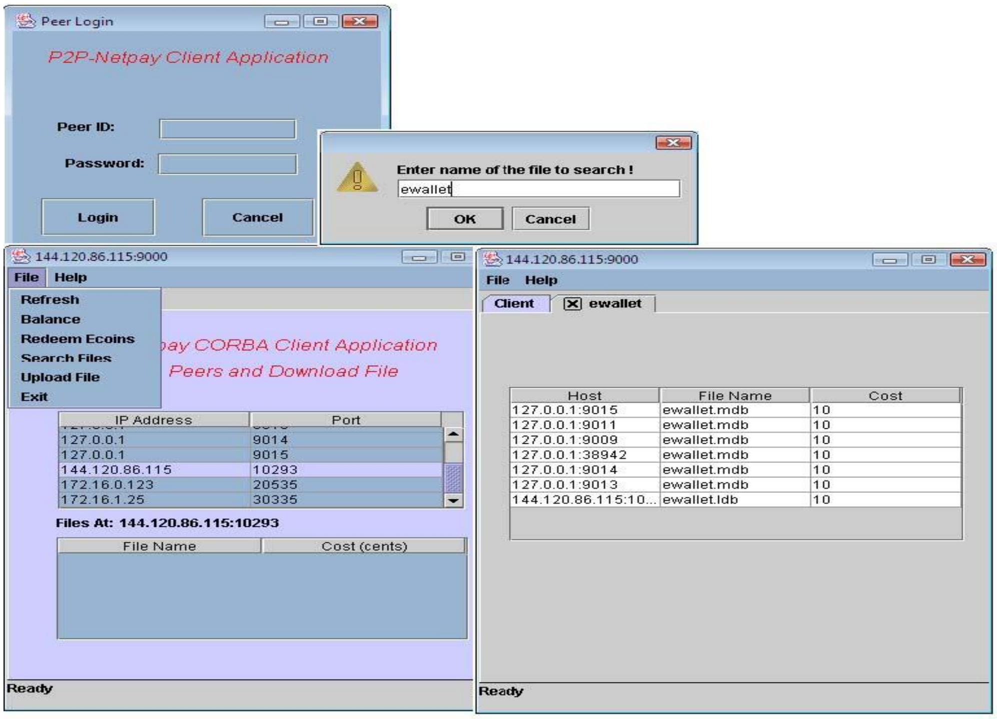


Figure 4: Redeem spending sequence diagram



Peer ID:

Password:

Enter name of the file to search !

ewallet

| IP Address | Port |
|---|---|
| 127.0.0.1 | 9014 |
| 127.0.0.1 | 9015 |
| 144.120.86.115 | 10293 |
| 172.16.0.123 | 20535 |
| 172.16.1.25 | 30335 |

Files At: 144.120.86.115:10293

| File Name | Cost (cents) |
|---|---|

| Host | File Name | Cost |
|---|---|---|
| 127.0.0.1:9015 | ewallet.mdb | 10 |
| 127.0.0.1:9011 | ewallet.mdb | 10 |
| 127.0.0.1:9009 | ewallet.mdb | 10 |
| 127.0.0.1:38942 | ewallet.mdb | 10 |
| 127.0.0.1:9014 | ewallet.mdb | 10 |
| 127.0.0.1:9013 | ewallet.mdb | 10 |
| 144.120.86.115:10... | ewallet.ldb | 10 |

Figure 5: P2P-Netpay user interfaces

## 6. Comparison of P2P-Netpay Architectures

Different architectures have been investigated for P2P-Netpay peers. Those are CORBA, socket, Web Services and RMI. Architectures have certain advantages and disadvantages. The advantages/disadvantages are in context of P2P-Netpay and not in general. Currently, Web Service is another option for distributed computing infrastructure [3] but the benefits of CORBA outweigh web services such as it supports multiple programming languages, it is a platform middleware, it interoperates with other middleware, it is highly flexible and it supports Remote Procedure Call and message-passing paradigms [2].

Web Services when compared to CORBA also supports multiple programming languages, is designed for web only, supports operation in heterogeneous connection at ends and uses XML to define interfaces and format messages [2].

Table 1 summarizes architecture comparison using five criteria's described below:

- ***Easy to use***: less coding required in achieving goals.
- ***Programming language*****:** Some architectures support multiple programming languages while others not. The benefit is different systems implemented via different programming languages can communicate with each other. Why implementation using different languages and not the same? This is to suite other users comfort ability in coding or using a particular language.
- ***Platform*****:** Much architecture doesn't support multiple platforms. There may be a situation that a server may be running on different machines and clients on another. To make these clients and servers to communicate, architecture must be platform independent.
- ***File downloading*****:** files broken into parts and then sent to peers for reassembling using previous architecture.
- ***Processing time*****:** time taken to process requests. The processing times are summarized in [8] and the results presented in the following table for time is based on [8].

Table 1: P2P-Netpay architecture comparison

| Property | ***Socket*** | ***Web Services*** | ***RMI*** | ***CORBA*** |
|---|---|---|---|---|
| *Easy to use* | **Low**, more codes required to handle series of messages for downloading file | **Medium**, messages are sent as SOAP; codes are required to build and read SOAP message | **High,** less to code | **High**, functionality is achieved with less coding |
| *Programming language* | **No** support for multiple programming language | **High**, supports multiple programming language; | **No** support for multiple programming language | **High,** implementation for either Broker/CIS or peer may differ. Broker/CIS may be implemented in Java whereas client can be implemented as C++ |
| *Platform* | **No** support for multiple platform | **No** support for multiple platform | **No** support for multiple platform | **High**, currently P2P-Netpay has only windows based application; there is a possibility for different platform applications. |
| *File downloading* | **Low**, breaking down of file into parts and sent as series of messages which results in download being aborted | - Very difficult impossible to implement peers as Web services<br>- uses HTTP for communication<br>- Each peer cannot host files on web servers and keep on updating since peers needs to have knowledge about HTML, JSP etc. | **High**, a remote method is invoked – no breaking down of file into parts | **High**, a remote method is invoked – no breaking down of file into parts |
| *Processing time* | **Less** processing time | **Medium** | **Less** processing time | **Medium**, since it supports distributed system. |

The above comparison shows that CORBA architecture is recommended for P2P-Netpay. Socket is very simple to understand and program. Peer/server must listen to a port in order to communicate and have to cater for each message sent and received. Sometimes the connection times out which aborts downloading requiring establishing connection again. Web services, on the other hand, is a new technology which uses XML to define messages sent to peers by Broker/CIS. Only the Broker/CIS can be implemented as Web service which has three-tier architecture when compared to multi-tier architecture for CORBA.

RMI is very similar to CORBA but it doesn't support multiple programming languages and platforms. The processing time for CORBA is higher when compared to RMI and sockets as in [8]. This is because the number of servers increase the processing time increases. In P2P-Netpay, only two servers will communicate so the processing time will be equivalent to socket and RMI.

## 7. Summary

We have developed a revised prototype architecture to support an efficient, secure and anonymous micro-payment system for high-volume, low-cost file sharing system. This incorporates a Broker/CIS which is used to generate, verify and redeem e-coins, a peer e-wallet stored on peer machine and peer application server components. All communication between peers and Broker/CIS – peer is through CORBA interface. Our initial prototype used CORBA architecture for Broker/CIS and socket for peers. Due to disadvantages of socket and in turn advantages of CORBA resulted in a new architecture for P2P-Netpay. We also compared four architectures for P2P-Netpay and decided that CORBA is more suitable architecture since P2P-Netpay is more likely to expand in future. Currently, we are investigating on web based file sharing application using P2P-Netpay payment service.

## 8. References

[1] Adar, E. and Huberman, B.: Free Riding on Gnutella, First Monday, 5(10), (2000)

[2] Baker, S., "Web Services and CORBA" *Lecture Notes in Computing Science*, vol. 2519, 2010, pp. 618-632. Publisher: Springer Berlin / Heidelberg

[3] CA, B., Barai, M. and Caselli, V., "Service Oriented Architecture with Java", Packt Publishing Ltd, 2008

[4] Chaudhary, K., Dai, X. & Grundy, J., "Experiences in Developing a Micro-payment System for Peer-to-Peer Networks", *International Journal of Information Technology and Web Engineering (IJITWE)*, vol. 5, no.1, March 2010, pp. 23 – 42. Publisher: IGI Global

[5] Chaudhary, K. & Dai, X., "P2P-NetPay: An Off-line Micro-payment System for Content Sharing in P2P-Networks", *Journal of Emerging Technologies in Web Intelligence (JETWI)*, vol.1, no.1, August 2009, pp. 46 - 54. Publisher: Academy Publisher.

[6] Dai, X., Chaudhary, K. and Grundy, J.: "Comparing and Contrasting Micro-payment Models for Content Sharing in P2P Networks", *Third International IEEE Conference on Signal-Image technologies and Internet-Based System (SITIS'07)*, 16 - 19 December 2007, Published by IEEE Computer Society, pp. 347-354

[7] Dai, X. and Grundy, J.: "Off-line Micro-payment System for Content Sharing in P2P Networks", *2nd International Conference on Distributed Computing & Internet Technology (ICDCIT 2005)* ,December 22-24, 2005, Lecture Notes in Computer Science Vol. 3816, pp 297 –307

[8] Eggen, R. and Eggen, M., "Effciency of Distributed Parallel Proceeing using Java RMI, Sockets, and CORBA", www.imamu.edu.sa/dcontent/IT_Topics/java/paper3.pdf

[9] Wei, K., Smith, A. J., Chen, Y. R. and Vo, B.(2006), "WhoPay: A scalable and anonymous payment system for peer-to-peer environments", in *Proc. 26th IEEE Intl. Conf. on Distributed Computing Systems, Los Alamitos, CA*, Computer Society Press, 2006, pp. 13-23.

[10] Yang, B. and Garcia-Molina, H.(2003), "PPay: micropayments for peer-to-peer systems", in *proc. Of the 10th ACM conference on computer and communication security*, ACM press, 2003, pp. 300- 310.

[11] Zghaibeh, M. and Harmantzis, F.C.(2006), "Lottery-based Pricing Scheme for Peer to Peer Networks**"**, *ICC apos;06. IEEE International Conference on Communications, 2006*, Volume 2, June 2006, pp. 903 – 908.

[12] Zou, E. J., Si, T. , Huang, L. and Dai, Y. (2005), "A New Micro-payment Protocol Based on P2P Networks", *Proceedings of the 2005 IEEE International Conference on e-Business Engineering (ICEBE'05),*IEEE Computer Society Press, 2005, pp. 449 – 455.

[13] OMG's CORBA : http://www.corba.org/

## Author Biographies

**Kaylash Chaudhary** received his MSc degree in Computing Science from University of the South Pacific in Fiji in 2009. He is now an assistant lecturer in the School of Computing, Information & Mathematical Sciences, University of the South Pacific. His research interests include software engineering, distributed system design and implementation, software architecture, electronic micro-payment systems for file-sharing, in peer-to-peer networks, mash-ups, and service-orientated architecture.

**Xiaoling Dai** received her B. S. degree in Mathematics with first class honors from Hebei University in China in 1984. In 2004, she received the Ph. D. degree in Computing Science from University of Auckland in New Zealand. She is now a Senior Lecturer in the School of Computing, Information & Mathematical Sciences, University of the South Pacific from 2005. Her research interests include component-based software engineering, distributed system design and implementation, software architecture, electronic micro-payment systems for e-commerce, file-sharing, or m-commerce in client-server, peer-to-peer, and mobile networks, web service security and service-oriented software engineering.

# A Study on Adjacency Matrix for Zero-Divisor Graphs over Finite Ring of Gaussian Integer

**Pranjali, Amit Sharma and R.K.Vats,**
**Department of Mathematics, National Institute of Technology,**
**Hamirpur, 177005, INDIA**
**Email: pranjalishrma11@rediffmail.com, apsharmanit@gmail.com, ramesh_vats@rediffmail.com**

**ABSTRACT :** The paper studies the characterization of adjacency matrix corresponding to zero-divisor graphs of finite commutative ring of Gaussian integer under modulo 'n'. For each positive integer we calculate number of zero-divisors & examine nature of the matrix, and then we generalized the order of matrix in each case. Firstly, we have started with some example, which motivates the later results. The study is useful in computer science application such as: coding theory, network communication, museum guard problems, etc.

**Keywords:** Gaussian Integer, Zero-divisor, Adjacency Matrix, Commutative ring.

**AMS Classification:** 05Cxx; 14C20; 15Axx,13Axx.

## 1. INTRODUCTION

Let R be finite commutative ring of Gaussian integer, Gaussian integer contains set of all complex numbers a+ib, where a and b are integer. It is denoted by Z[i]. It forms Euclidian domain under usual complex operations, with norm $N(a+ib) = a^2+b^2$. It is clear that a+ib is unit in Z[i] iff N(a+ib) =1, which implies that 1, -1, i, -i are only units.

Let <n> be the principal ideal generated by n in Z[i], where n is a natural number and let $Z_n$={0,1,2,3,4….n-1} be ring of integer modulo n. The factor ring Z[i]/<n> is isomorphic to $Z_n[i]$ = {a+ib : a,b $Z_n$} which implies that $Z_n[i]$ is Principal ideal ring. Therefore $Z_n[i]$ is ring of Gaussian integer under modulo n. Consider $Z_n[i]$, and let Z(R) be set of zero divisors of $Z_n[i]$ and G be zero divisor graph of $Z_n[i]$. Consider a, b $Z_n[i]$, then a and b are said to be adjacent if a.b = b.a = 0. The ring of Gaussian integer $Z_p[i]$ forms field if p≡3(mod4) [1], therefore the graph G has no edge if p≡3(mod4).

The concept of zero divisors graph was given by I. Beck [2] but his motive was in coloring of graphs. In [2] Anderson and Livingston associate to a commutative ring with unity a graph $\Gamma R$, whose vertex was $Z(R)^{\bullet}$ = Z(R) –{0}. The Zero divisor graphs also have discussed and studied for semi groups by De Meyer [4]. Redmond has generalized the notation of zero divisor graphs. On studying this article it is found that now considerable work has been done in this direction. Some time the zero divisor graph for R is allowed to have '0' as a vertex, in such case '0' has an edge to every other vertex in graphs. For simplification, we have used the definition excluding '0' as a vertex. In the first instance of this paper we consider some finite rings of Gaussian integer and discuses nature of adjacency matrix in each case depending on 'n' we also investigate that what can be the order of matrix in each case, we have started with some example which illustrates the general results. The definition of adjacency matrix for zero divisor graphs is as follows:

$$d_{ij} = \begin{cases} 1, & \text{If } v_i \ \& \ v_j \text{ represent zero} - \text{divisor} \\ 0, & \text{Otherwise} \end{cases},$$

where $v_i$ and $v_j$ are vertices of graph G.

One more advantage of the graph that it also detects the nilpotent element of index 2, when self loop found. We will take basic definition from graph theory [5]-[6] for commutative ring with unity [7]. To avoid trivialities when G has no edges, we will assume when necessary R is not Integral domain, (i.e., we left case p≡3(mod4) [4]). We study for the following rings $Z_p[i]$, $Z_{pn}[i]$, $Z_{pq}[i]$. Some examples are giving below for each case of $Z_p[i]$ :

I. p≡2 (mod4)

II. p≡1 (mod4)

And for ring $Z_{pn}[i]$, n>1 the cases are such as:

I. p≡1(mod 4)

II. p≡3(mod 4)

III. p≡2(mod 4) and

Last case for $Z_n[i]$, when n = p.q, n≡2(mod 4)

## 2. The ring $Z_p[i]$,

**Case 2.1: When p≡2(mod4), i.e., R = $Z_2[i]$**

Firstly we move to discuss about R = $Z_2[i]$

The set of zero divisor Z(R) = {1+i}, the possible edge is self loop.

Graph for Z(R) is given as:

Fig. 1: Zero divisor graph of the $Z_2[i]$

and the adjacency matrix is given as:

$$[1]_{1\cdot 1}$$

**Observation from matrix:**

i. Ring has only one zero divisor

ii. Matrix is of order '1' and having trace and determinant equal.

**Case 2.2: When p≡1(mod4), i.e., R = $Z_5[i]$ and R=$Z_{13}[i]$.**

**Case 2.2.1:** For R = $Z_5[i]$.

The set of zero divisor Z(R) = {2+i, 3+i, 4+2i, 2+4i, 1+3i, 1+2i, 4+3i, 3+4i}

The graph is shown below:

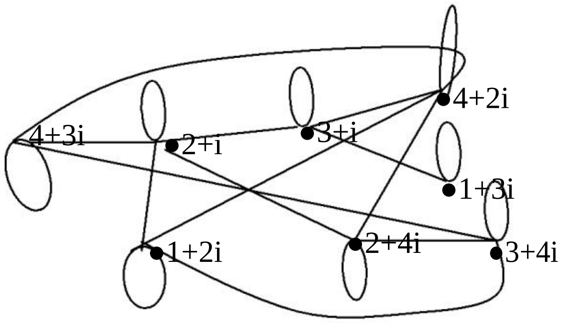


Fig. 2: Zero divisor graph of the $Z_5[i]$

The matrix corresponding to zero divisor graph of $Z_5[i]$

$$\begin{matrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{matrix}_{8\cdot 8}$$

**Observation from matrix:**

i. The matrix corresponding to zero divisor graph of $Z_5[i]$ is non singular.

ii. All vertices have self loop so trace of the matrix is 8.

iii. The rank of the matrix is 8, therefore zero is not the eigen value of the above matrix.

**Case 2.2.2:** For R = $Z_{13}[i]$.

The set of zero divisor = {1+5i, 1+8i, 2+3i, 2+10i, 3+2i, 3+11i, 4+6i, 4+7i , 5+12i, 6+4i, 6+9i, 7+4i, 7+9i, 8+i, 8+12i, 9+6i, 9+7i, 10+2i, 10+11i, 11+3i, 11+10i, 12+5i, 12+8i}

The graph is shown below:

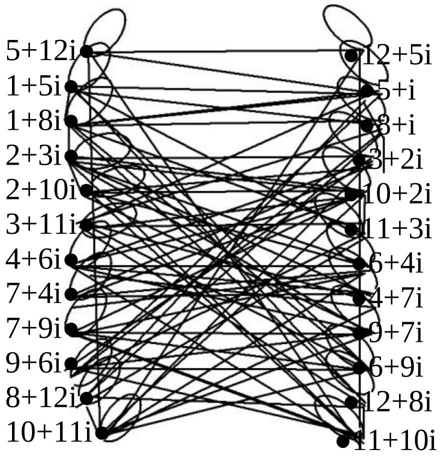


Fig. 3: Zero divisor graph of the $Z_{13}[i]$

The adjacency matrix will be of order 24×24 and can be constructed by using definition.

**Observation from matrix:**

i. The matrix corresponding to zero divisor graph of $Z_5[i]$ is non singular.

ii. Trace of the matrix will be 24.

iii. Zero will not be the eigenvalue of matrix as rank of adjacency matrix is 24.

**Theorem 2.1:** *Let $Z_n[i]$ be ring of Gaussian integer under modulo 'n'. Consider $Z_p[i]$, p≡1(mod4), p be prime. Let M be adjacency matrix corresponding to zero divisor graph of $Z_p[i]$. The order of adjacency matrix in such case is always 2(p-1) ×2(p-1).*

**Proof**: The ring of Gaussian integer Zn[i] is finite commutative ring under modulo 'n'. It is known by theorem [1] that in a finite commutative ring each non-zero element is either a unit or a zero divisor. To obtain number of zero divisor we subtract units from non zero element. It is found that units in $Z_p[i]$ are φ(p) × φ(p), therefore no. of zero divisor = $p^2$-(p-1) × (p-1)-1 = 2(p-1)

Thus it have proved that order of matrix is 2(p-1) × 2(p-1).

**Theorem 2.2:** *Let $Z_n[i]$ be ring of Gaussian integer under modulo 'n'. Consider $Z_p[i]$, p≡1(mod4), p be prime. Let M be adjacency matrix corresponding to zero divisor graph of $Z_p[i]$. Then adjacency matrix will always be non singular.*

**Proof:** Let us consider $Z_p[i]$, p≡1(mod 4) as above discussed example it has observed that if a+ib represent zero divisor then b+ia also represent zero divisor. From the graph it is found there is no such vertex at which a+ib and b+ia both are connected, thus in adjacency matrix no two rows are identical. Therefore we have shown that matrix will be non singular.

**Theorem 2.3:** *Let $Z_n[i]$ be ring of Gaussian integer under modulo 'n'. Consider $Z_p[i]$, p≡1(mod4), p be prime. Let M be adjacency matrix corresponding to zero divisor graph of $Z_p[i]$. Then trace of adjacency matrix is always equal to number of zero divisor.*

**Proof:** Let us consider $Z_p[i]$, p≡1(mod 4) as above discussed example it has observed that if a+ib represent zero divisor then b+ia also represent zero divisor. From the graph it is found that all the vertices has self loop, therefore trace is equal to number of zero divisor.

## 3. The ring $Z_{pn}[i]$

**Case 3.1: When p≡3(mod4), i.e., R = $Z_9[i]$, $Z_{27}[i]$**

Let R = $Z_9[i]$, the set of the zero divisor is Z(R) = {3, 6, 3i, 6i, 3+3i, 6+3i, 6+6i, 3+6i}

The corresponding zero divisor graph is represented as:

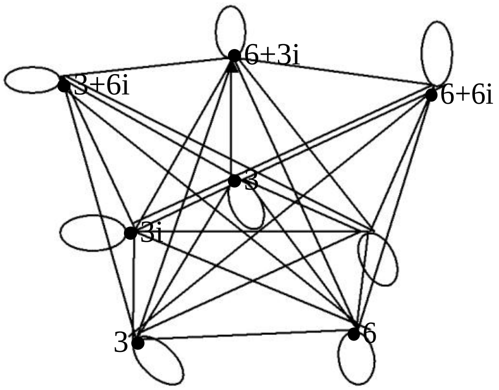


Fig. 4: Zero divisor graph of the $Z_9[i]$

Matrix for the above figure is

$$\begin{matrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{matrix}_{8\cdot 8}$$

**Observation from matrix:**

i. The matrix for the zero divisor of $Z_9[i]$ is singular.

ii. Trace of the adjacency matrix is 8, as all the vertices have self loop.

iii. Rank of the matrix is 1.

iv. Eigen values of above matrix are 0 and 8.

v. Adjacency matrix corresponding to Z(R) is diagonalizable.

**Theorem 3.1:** *Let $Z_n[i]$ be ring of Gaussian integer under modulo 'n'. Consider $Z_{pn}[i]$ , p≡3(mod4), p be prime. Let M be adjacency matrix corresponding to zero divisor graph of $Z_{pn}[i]$ , the order of adjacency matrix in such case is always* $p^{2n}$-8$p^{2n-2}$-1× $p^{2n}$-8$p^{2n-2}$-1.

**Proof:** Similar as theorem 1.

**Theorem 3.2:** *Let $Z_n[i]$ be ring of Gaussian integer under modulo 'n'. Consider $Z_{p^n}[i]$, p≡3(mod4), p be prime. Let M be adjacency matrix corresponding to zero divisor graph of $Z_{p^n}[i]$. Then trace of adjacency matrix is always natural number n>1.*

**Proof:** Let us consider $Z_{p^n}[i]$, p≡3(mod 4), here p, $p^2$, $p^3$… or $p^{n-1}$ as well as ip, $ip^2$, $ip^{n-1}$ represent zero divisor with itself, i.e., graph must have self loop at least two pairs which are conjugate. Therefore, at least the diagonal entry, i.e., $a_{ii}$ and $a_{jj}$ of the adjacency matrix contain 1. Thus the trace of matrix is natural number 'n', n>1.

**Case 3.2: When p≡2(mod4), i.e., R = $Z_4$[i], $Z_8$[i]**

**Case 3.2.1**for R = $Z_4$[i], the set of zero divisors for $Z_4$[i] = {2, 2i, 1+i, 3+i, 2+2i, 1+3i, 3+3i}

The possible edges for the graph are {2,2}, {2i,2i}, {2+2i,2+2i}, {2+2i,1+i}, {2+2i,3+i}, {2,2i}, {2,2+2i}, {2i, 2+2i}, {3+3i,2+2i}, {2+2i,1+3i}

Graph is given as

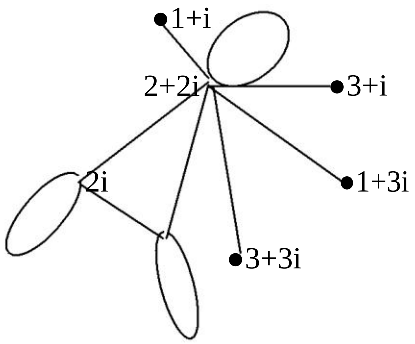


Fig. 5: Zero divisor graph of the $Z_4$[i]

The adjacency matrix for the graph:

$$\begin{matrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{matrix}_{7\cdot 7}$$

**Observation from matrix:**

i. The Adjacency matrix with respect to zero divisor graph of $Z_4$[i] is singular.
ii. The trace of above matrix is 3.
iii. Rank of adjacency matrix is 3, which is less than seven so zero must be eigenvalue.

**Case 3.2.2** for R = $Z_8$[i]

The set of zero divisor Z(R) = {2, 4, 2i, 4i, 6, 6i, 1+i, 4+4i, 1+5i, 1+7i, 5+i, 3+i, 2+2i, 3+3i, 2+6i, 6+2i, 7+7i, 4+6i, 6+4i, 7+i, 5+7i, 7+5i, 5+5i, 4+2i, 2+4i, 3+5i, 5+3i, 4+4i, 1+3i, 3+7i, 7+3i} and graph of the zero divisor is given as:

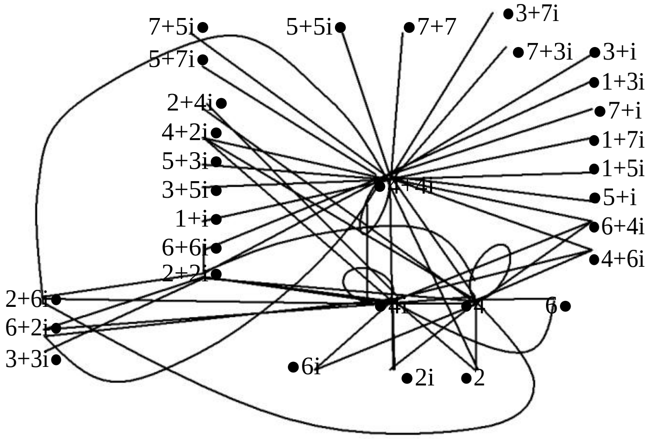


Fig.6: Zero divisor graph of the $Z_8$[i]

and the matrix for the zero divisor graph will be of order 31×31 and can be constructed in similar manner.

**Observation from matrix:**

i. The Adjacency matrix with respect to zero divisor graph of $Z_8$[i] is singular.
ii. The trace of above matrix is 3.
iii. Rank of adjacency matrix is 10.

**Theorem 3.3:** *Let $Z_n[i]$ be ring of Gaussian integer under modulo 'n'. Consider $Z_{p^n}[i]$, p≡2 (mod4), p be prime. Let M be adjacency matrix corresponding to zero divisor graph of $Z_{p^n}[i]$. The order of adjacency matrix in such case is always* $\frac{1}{2} p^{2n-1} \cdot \frac{1}{2} p^{2n-1}$.

**Proof:** Similar as theorem 1 a**s** the units of $Z_{pn}[i]$, are $\frac{1}{2}p^{2n}$ by using the reference [1].

**Theorem 3.4:** *Let $Z_n[i]$ be ring of Gaussian integer under modulo 'n'. Consider $Z_{pn}[i]$, p≡2(mod4), p be prime. Let M be adjacency matrix corresponding to zero divisor graph of $Z_{pn}[i]$. The adjacency matrix in this case is always singular.*

**Proof:** Let us consider $Z_{pn}[i]$, p≡2(mod4) as above discussed example it has observed that if a+ib represent zero divisor then b+ia also represent zero divisor. From the graph, it is found that at least two vertices $(p^n-1)(1+i)$ and (1+i) represent zero divisor with $\frac{1}{2}p^n \cdot (1+i)$. In adjacency matrix at least two rows will be identical, thus determinate of matrix is zero.

**Theorem 3.5:** *Let $Z_n[i]$ be ring of Gaussian integer under modulo 'n'. Consider $Z_{pn}[i]$, p≡2(mod4) p be prime. Let M be adjacency matrix corresponding to zero divisor graph of $Z_{pn}[i]$. Then trace of matrix in such case is always three.*

**Proof:** Let us consider $Z_{pn}[i]$, p≡2(mod4) as above discussed example it has observed that if a+ib represent zero divisor then b+ia also represent zero divisor. From the graph it is found that vertices $\frac{i}{2}p^n$, $\frac{1}{2}p^n$ and $\frac{1}{2}p^n \cdot (1+i)$ which produces self loop always. Thus trace of adjacency matrix is three [1].

## 4. The ring $Z_n[i]$

**When n = p.q, n≡2(mod4),** where p and q are distinct prime numbers, i.e., R = $Z_6[i]$, $Z_{10}[i]$,

**Case 4.1** for R = $Z_6[i]$,

Consider $Z_6[i]$, the set of zero divisors Z(R)= {2, 3, 4, 2i, 3i, 4i, 3+3i, 1+i, 2+2i, 4+4i, 2+4i, 3+i, 5+i, 1+3i, 5+3i, 5+5i, 3+5i, 4+2i, 1+5i}.

The zero divisor graph is shown as:

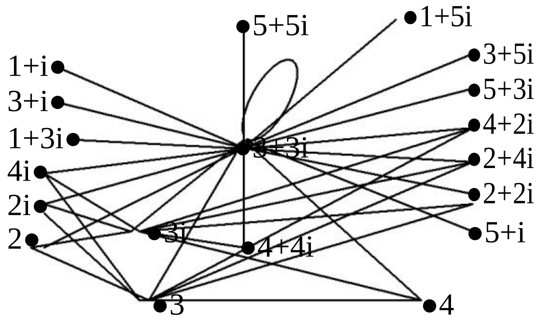


Fig. 7: Zero divisor graph of the $Z_6[i]$

The adjacency matrix is of order 19×19 and can be constructed similarly.

**Observation from matrix:**

i. The determinant of the matrix is zero.
ii. Trace of the adjacency matrix is 1.
iii. Rank of the matrix is 4.
iv. Matrix is singular so, zero must be eigen value.

**Theorem 4.1:** *Let $Z_n[i]$ be ring of Gaussian integer under modulo 'n'. Consider $Z_{pn}[i]$, p≡2(mod4), p be prime. Let M be adjacency matrix corresponding to zero divisor graph of $Z_{pn}[i]$. The adjacency matrix in this case is always singular.*

**Proof:** Let R = $Z_{p^n}[i]$ be ring over Gaussian integer p≡2(mod4). If a+bi represent zero divisor then b+ai also gives zero divisor, and the vertices $(a_1+b_1i)\frac{1}{2}p^n \cdot (1+i) = 0$ and $(b_1+ia_1)\frac{1}{2}p^n \cdot (1+i) = 0$, i.e., some of rows of matrix due to above product will be identical so determinant of matrix is zero.

**Theorem 4.2:** *Let $Z_n[i]$ be ring of Gaussian integer under modulo 'n'. Consider $Z_{pn}[i]$, p≡2(mod4), p be prime. Let M be adjacency matrix corresponding to zero divisor graph of $Z_{pn}[i]$. Then trace of adjacency matrix is always natural number, k>1.*

**Proof:** Let R = $Z_{pn}$[i] , let p≡2(mod4), i.e., form of $2^n$ in this case vertex $\frac{1}{2}p^{n} \cdot (1+i)$ represents zero divisor and in fact, when graph is formed vertex $\frac{1}{2}p^{n} \cdot (1+i)$ always have self loop, when adjacency matrix is constructed then $\frac{1}{2}p^{n} \cdot (1+i)$ vertex have entry 1(diagonal form), so trace is at least k, k>1.

## 5. Conclusion:

In this paper, we study adjacency matrices for zero divisor graph over finite rings of Gaussian integer. Graphs are the most ubiquitous models of both natural and human made structures. In computer science, zero divisor graphs are used to represent networks of communication, network flow, clique problems. Art gallery and museum guard problem are a well-studied visibility problem in computational geometry**.**

## REFERENCES

[1] J. A. Gallian, Abstract Algebra, Narosa Publishing House, ISBN: 81-7319-269-3(1998).

[2] I. Beck, “Coloring of Commutating Ring”, J. Algebra 116, 208-226(1988).

[3] D.F. Anderson, P.S. Livingston, “The Zero-divisor Graph of Commutative Ring”, Journal of Algebra 217, 434-447(1999).

[4] F.R.DeMeyer, T.Mckenzie, K.Schneider, “The Zero-divisor Graph of a Commutative Semi-groups”, Semi Group Forum 65, 206-214(2002).

[5] R.Diestel, Graph Theory, Springer-Verlag, Newyork, 1977.

[6] F. Harary, Graph Theory, Addison-Wesley, Reading, MA, 1972.

[7] I.Kaplansky, Commutative Rings, Univ. of Chicago Press, Chicago, 1974.

[8] V.K Bhat, Ravi Raina, “A Note on Zero-divisor Graph over Rings”, Int. J. Contemp. Math. Sci. 2(14), 667-671(2007).

[9] B.Bollabs, Graph Theory-An Introductory Course, Springer-Verlag, Newyork, 1979.

[10] M.F Atiyah, I.G, Macdonald, “Introduction to Commutative Algebra”, Addison-Wesley, Reading, MA, 1989.

[11] Nafiz Abu Jaradeh Emad Abu Osba and Salah Ai-Addasi, “Zero Divisor graph for the ring of Gaussian integers modulo n”, Taylor & Francis, Communications in algebra, 36: 3865-3877(2008).

# SQL Injection Attacks and Its Counter Measures

Er. Upinder Kaur [1], Er. Navdeep Kochhar[2]

Dept. of Computer Science, Baba Farid college, Deon , Bathinda, Punjab, India.
raj_chawla94@yahoo.com[1], er_navdeepkochhar@rediffmail.com[2]

***Abstract:*** SQL injection is a technique for exploiting web applications that use client-supplied data in SQL queries, but without first stripping potentially harmful characters. Despite being remarkably simple to protect against, there is an astonishing number of production systems connected to the Internet that are vulnerable to this type of attack. To address this problem, we present an extensive review of the different types of SQL injection attacks known to date. For each type of attack, we provide descriptions and examples of how attacks of that type could be performed. We also present and analyze existing detection and prevention techniques against SQL injection attacks. For each technique, we discuss its strengths and weaknesses in addressing the entire range of SQL injection attacks

*Keywords:* SQLIAS, preventions, attacks, SQL injections.

## 1. Introduction:

SQL injection vulnerabilities have been described as one of the most serious threats for Web applications [3][11]. Web applications that are vulnerable to SQL injection may allow an attacker to gain complete access to their underlying databases. Because these databases often contain sensitive consumer or user information, the resulting security violations can include identity theft, loss of confidential information, and fraud. In some cases, attackers can even use an SQL injection vulnerability to take control of and corrupt the system that hosts the Web application. Web applications that are vulnerable to SQL Injection Attacks (SQLIAs) are widespread—a study by Gartner Group on over 300 Internet Web sites has shown that most of them could be vulnerable to SQLIAs. In fact, SQLIAs have successfully targeted high-profile victims such as Travelocity, FTD.com, and Guess Inc. SQL injection refers to a class of code-injection attacks in which data provided by the user is included in an SQL query in such a way that part of the user's input is treated as SQL code. By leveraging these vulnerabilities, an attacker can submit SQL commands directly to the database. These attacks are a serious threat to any Web application that receives input from users and incorporates it into SQL queries to an underlying database. Most web applications used on the Internet or within enterprise systems work this way and could therefore be vulnerable to SQL injection. The cause of SQL injection vulnerabilities is relatively simple and well understood: insufficient validation of user input. To address this problem, developers have proposed a range of coding guidelines (e.g., [18]) that promote defensive coding practices, such as encoding user input and validation. A rigorous and systematic application of these techniques is an effective solution for preventing SQL injection vulnerabilities. However, in practice, the application of such techniques is human-based and, thus, prone to errors. Furthermore, fixing legacy code-bases that might contain SQL injection vulnerabilities can be an extremely labor-intensive task. Although recently there has been a great deal of attention to the problem of SQL injection vulnerabilities, many proposed solutions fail to address the full scope of the problem. There are many types of SQLIAs and countless variations on these basic types. Researchers and practitioners are often unaware of the myriad of different techniques that can be used to perform SQLIAs. Therefore, most of the solutions proposed detect or prevent only a subset of the possible SQLIAs. To address this problem, we present a comprehensive survey of SQL injection attacks known to date. To compile the survey, we used information gathered from various sources, such as papers,Web sites, mailing lists, and experts in the area. For each attack type considered, we give a characterization of the attack, illustrate its effect, and provide examples of how that type of attack could be performed. This set of attack types is then used to evaluate state of the art detection and prevention techniques and compare their strengths and weaknesses. The results of this comparison show the effectiveness of these techniques.

### 1.1 Injection Mechanisms

Malicious SQL statements can be introduced into a vulnerable application using many different input mechanisms. In this section, we explain the most common mechanisms.

**Injection through user input:** In this case, attackers inject SQL commands by providing suitably crafted user input. A Web application can read user input in several ways based on the environment in which the application is deployed. In most SQLIAs that target Web applications, user input typically comes from form submissions that are sent to the Web application via HTTP GET or POST requests [14]. Web applications are generally able to access the user input contained in these requests as they would access any other variable in the environment.

**Injection through cookies:** Cookies are files that contain state information generated byWeb applications and stored on the client machine. When a client returns to a Web application, cookies can be used to restore the client's state information. Since the client has control over the storage of the cookie, a malicious client could tamper with the cookie's contents. If a Web application uses the cookie's contents to build SQL queries, an attacker could easily submit an attack by embedding it in the cookie [8].

**Injection through server variables:** Server variables are a collection of variables that contain HTTP, network headers, and environmental variables. Web applications use these server variables in a variety of ways, such as logging usage statistics and identifying browsing trends. If these variables are logged to a database without sanitization, this could create an SQL injection vulnerability. Because attackers can forge the values that are placed in HTTP and network headers, they can exploit this vulnerability by placing an SQLIA directly into the headers. When the query to log the server variable is issued to the database, the attack in the forged header is then triggered.

**Second-order injection:** In second-order injections, attackers seed malicious inputs into a system or database to indirectly trigger an SQLIA when that input is used at a later time. The objective of this kind of attack differs significantly from a regular (i.e., first order) injection attack. Second-order injections are not trying to cause the attack to occur when the malicious input initially reaches the database. Instead, attackers rely on knowledge of where the input will be subsequently used and craft their attack so that it occurs during that usage. To clarify, we present a classic example of a second order injection attack (taken from [1]). In the example, a user registers on a website using a seeded user name, such as "admin' -- ". The application properly escapes the single quote in the input before storing it in the database, preventing its potentially malicious effect. At this point, the user modifies his or her password, an operation that typically involves (1) checking that the user knows the current password and (2) changing the password if the check is successful. To do this, the Web application might construct an SQL command as follows:

queryString="UPDATE users SET password='" + newPassword +
"' WHERE userName='" + userName + "' AND password='" +
oldPassword + "'" newPassword and oldPassword are the new and old passwords,
respectively, and userName is the name of the user currently
logged-in (i.e., ''admin'--'').
Therefore, the query string that is sent to the database is (assume that newPassword and oldPas-sword are "newpwd" and"oldpwd"):
UPDATE users SET password='newpwd' WHERE userName= 'admin'--' AND password='oldpwd'

Because "--" is the SQL comment operator, everything after it is ignored by the database. Therefore, the result of this query is that the database changes the password of the administrator ("admin") to an attacker-specified value. Second-order injections can be especially difficult to detect and prevent because the point of injection is different from the point where the attack actually manifests itself. A developer may properly escape, type-check, and filter input that comes from the user and assume it is safe. Later on, when that data is used in a different context, or to build a different type of query, the previously sanitized input may result in an injection attack.

## 2. Different types of SQLIA

In this section, we present and discuss the different kinds of SQLIAs known to date. For each attack type, we provide a descriptive name, a description of the attack, an attack example, and a set of references to publications and Web sites that discuss the attack technique and its variations in greater detail. The different types of attacks are generally not performed in isolation; many of them are used together or sequentially, depending on the specific goals of the attacker. Note also that there are countless variations of each attack type. For space reasons, we do not present all of the possible attack variations but instead present a single representative example.

**Tautologies Attack**: Bypassing authentication, identifying injectable parameters, extracting data.

Description: The general goal of a tautology-based attack is to inject code in one or more conditional statements so that they always evaluate to true. The consequences of this attack depend on how the results of the query are used within the application. The most common usages are to bypass authentication pages and extract data. In this type of injection, an attacker exploits an injectable field that is used in a query's WHERE conditional. Transforming the conditional into a tautology causes all of the rows in the database table targeted by the query to be returned. Example: In this example attack, an attacker submits " ' or 1=1 - - " for the login input field (the input submitted for the other fields is irrelevant). The resulting query is:

SELECT accounts FROM users WHERE login='' or 1=1 -- AND pass='' AND pin=

The code injected in the conditional (OR 1=1) transforms the entire WHERE clause into a tautology. The database uses the conditional as the basis for evaluating each row and deciding which ones to return to the application. Because the conditional is a tautology, the query evaluates to true for each row in the table and returns all of them. In our example, the returned set evaluates to a non null value, which causes the application to conclude that the user authentication was successful. Therefore, the application would invoke method displayAccounts() and show all of the accounts in the set returned by the database. [1][28][21][18]

**Illegal/Logically Incorrect Queries Attack**: Identifying injectable parameters, performing database finger-printing, extracting data.

Description: This attack lets an attacker gather important information about the type and structure of the back-end database of a Web application. The attack is considered a preliminary, information gathering step for other attacks. The vulnerability leveraged by this attack is that the default error page returned by application servers is often overly descriptive. In fact, the simple fact that an error messages is generated can often reveal vulnerable/injectable parameters to an attacker. Additional error information, originally intended to help programmers debug their applications, further helps attackers gain information about the schema of the back-end database.

Example: This example attack's goal is to cause a type conversion error that can reveal relevant data. To do this, the attacker injects the following text into input field pin: "convert(int,(select top 1 name from sysobjects where xtype='u'))". The resulting query is:

SELECT accounts FROM users WHERE login='' AND pass='' AND pin= convert (int,(select top 1 name from
sysobjects where xtype='u'))

In the attack string, the injected select query attempts to extract the first user table (xtype='u') from the database's metadata table (assume the application is using Microsoft SQL Server, for which the metadata table is called sysobjects). The query then tries to convert this table name into an integer. Because this is not a legal type conversion, the database throws an error. For Microsoft SQL Server, the error would be: "Microsoft OLE DB Provider for SQL Server (0x80040E07) Error converting nvarchar value 'CreditCards' to a column of data type int." There are two useful pieces of information in this message that aid an attacker. First, the attacker can see that the database is an SQL Server database, as the errormessage explicitly states this fact. Second, the error message reveals the value of the string that caused the type conversion to occur. In this case, this value is also the name of the first user-defined table in the database: "CreditCards." A similar strategy can be used to systematically extract the name and type of each column in the database. Using this information about the schema of the database, an attacker can then create further attacks that target specific pieces of information. [1][22][28]

**Union Query Attack**: Bypassing Authentication, extracting data.

Description: In union-query attacks, an attacker exploits a vulnerable parameter to change the data set returned for a given query. Attackers do this by injecting a statement of the form: UNION SELECT <rest of injected query>. Because the attackers completely control the second/injected query, they can use that query to retrieve information from a specified table. The result of this attack is that the database returns a dataset that is the union of the results of the original first query and the results of the injected second query.

Example: Referring to the running example, an attacker could inject the text "' UNION SELECT cardNo from CreditCards where acctNo=10032 - -" into the login field, which produces the following query:

SELECT accounts FROM users WHERE login=''
UNION SELECT cardNo from CreditCards where
acctNo=10032 -- AND pass='' AND pin=

Assuming that there is no login equal to “”, the original first query returns the null set, whereas the second query returns data from the “CreditCards” table. In this case, the database would return column “cardNo” for account “10032.” The database takes the results of these two queries, unions them, and returns them to the application. In many applications, the effect of this operation is that the value for “cardNo” is displayed along with the account information. [1][ 28][21]

**PiggyBacked Queries Attack**: Extracting data, adding or modifying data, performing denial of service, executing remote commands.
Description: In this attack type, an attacker tries to inject additional queries into the original query. We distinguish this type from others because, in this case, attackers are not trying to modify the original query; instead, they are trying to include new and distinct queries that “piggy-back” on the original query. As a result, the database receives multiple SQL queries. Vulnerability to this type of attack is often dependent on having a database configuration that allows multiple statements to be contained in a single string.
Example: If the attacker inputs “’; drop table users - - ” into the pass field, the application generates the query:

SELECT accounts FROM users WHERE login=’doe’ AND pass=’’; drop table users -- ’ AND pin=123

After completing the first query, the database would recognize the 1 stored procedures are routines stored in the database and run by the database engine. These procedures can be either user-defined procedures or procedures provided by the database by default. query delimiter (“;”) and execute the injected second query. The result of executing the second query would be to drop table users, which would likely destroy valuable information. Other types of queries could insert new users into the database or execute stored procedures. Note that many databases do not require a special character to separate distinct queries, so simply scanning for a query separator is not an effective way to prevent this type of attack. [1][28][18]

**Stored Procedures Attack**: Performing privilege escalation, performing denial of service, executing remote commands.
Description: SQLIAs of this type try to execute stored procedures present in the database. Today, most database vendors ship databases with a standard set of stored procedures that extend the functionality of the database and allow for interaction with the operating system. Therefore, once an attacker determines which backend database is in use, SQLIAs can be crafted to execute stored procedures provided by that specific database, including procedures that interact with the operating system. It is a common misconception that using stored procedures to write Web applications renders them invulnerable to SQLIAs. Developers are often surprised to find that their stored procedures can be just as vulnerable to attacks as their normal applications [18][24]. Additionally, because stored procedures are often written in special scripting languages, they can contain other types of vulnerabilities, such as buffer overflows, that allow attackers to run arbitrary code on the server or escalate their privileges [9].

**Stored procedure for checking credentials.**
CREATE PROCEDURE DBO.isAuthenticated @userName varchar2, @pass varchar2, @pin int
AS EXEC("SELECT accounts FROM users WHERE login=’" +@userName+ "’ and pass=’" +@password+
"’ and pin=" +@pin);
GO

Example: This example demonstrates how a parameterized stored procedure can be exploited via an SQLIA. In the example, we assume that the query string constructed at lines of our example has been replaced by a call to the stored procedure. The stored procedure returns a true/false value to indicate whether the user’s credentials authenticated correctly. To launch an SQLIA, the attacker simply injects “ ’ ; SHUTDOWN; - -” into either the userName or password fields. This injection causes the stored procedure to generate the following query:

SELECT accounts FROM users WHERE login=’doe’ AND pass=’ ’; SHUTDOWN; -- AND pin=

At this point, this attack works like a piggy-back attack. The first query is executed normally, and then the second, malicious query is executed, which results in a database shut down. This example shows that stored procedures can be vulnerable to the same range of attacks as traditional application code. [1][ 4][ 9][10][24][28][21][18]

**Inference Attack**: dentifying injectable parameters, extracting data, determining database schema.
Description: In this attack, the query is modified to recast it in the form of an action that is executed based on the answer to a true/- false question about data values in the database. In this type of injection, attackers are generally trying to attack a site that has been secured enough so that, when an injection has

succeeded, there is no usable feedback via database error messages. In this situation, the attacker injects commands into the site and then observes how the function/response of the website changes. By carefully noting when the site behaves the same and when its behavior changes, the attacker can deduce not only whether certain parameters are vulnerable, but also additional information about the values in the database. There are two well known attack techniques that are based on inference. They allow an attacker to extract data from a database and detect vulnerable parameters. Researchers have reported that with these techniques they have been able to achieve a data extraction rate of 1B/s [2].

**Blind Injection:** In this technique, the information must be inferred from the behavior of the page by asking the server true/- false questions. If the injected statement evaluates to true, the site continues to function normally. If the statement evaluates to false, although there is no descriptive error message, the page differs significantly from the normally-functioning page.

**Timing Attacks:** A timing attack allows an attacker to gain information from a database by observing timing delays in the response of the database. This attack is very similar to blind injection, but uses a different method of inference. To perform a timing attack, attackers structure their injected query in the form of an if/then statement, whose branch predicate corresponds to an unknown about the contents of the database. Along one of the branches, the attacker uses a SQL construct that takes a known amount of time to execute, (e.g. the WAITFOR keyword, which causes the database to delay its response by a specified time). By measuring the increase or decrease in response time of the database, the attacker can infer which branch was taken in his injection and therefore the answer to the injected question.

Example: Using the code from our running example, we illustrate two ways in which Inference based attacks can be used. The first of these is identifying injectable parameters using blind injection. Consider two possible injections into the login field. The first being “legalUser’ and 1=0 - -” and the second, “legalUser’ and 1=1 - -”. These injections result in the following two queries:

SELECT accounts FROM users WHERE login=’legalUser’ and 1=0 -- ’ AND pass=’’ AND pin=0

SELECT accounts FROM users WHERE login=’legalUser’ and 1=1 -- ’ AND pass=’’ AND pin=0

Now, let us consider two scenarios. In the first scenario, we have a secure application, and the input for login is validated correctly. In this case, both injections would return login error messages, and the attacker would know that the login parameter is not vulnerable. In the second scenario, we have an insecure application and the login parameter is vulnerable to injection. The attacker submits the first injection and, because it always evaluates to false, the application returns a login error message. At this point however, the attacker does not know if this is because the application validated the input correctly and blocked the attack attempt or because the attack itself caused the login error. The attacker then submits the second query, which always evaluates to true. If in this case there is no login error message, then the attacker knows that the attack went through and that the login parameter is vulnerable to injection.

The second way inference based attacks can be used is to perform data extraction. Here we illustrate how to use a Timing based inference attack to extract a table name from the database. In this attack, the following is injected into the login parameter: ‘‘legalUser’ and ASCII(SUBSTRING((select top 1 name from sysobjects),1,1)) > X WAITFOR 5 --’’. This produces the following query:

SELECT accounts FROM users WHERE
login=’legalUser’ and
ASCII(SUBSTRING((select top 1 name from sysobjects),1,1))
> X WAITFOR 5 -- ’ AND pass=’’ AND pin=0

In this attack the SUBSTRING function is used to extract the first character of the first table’s name. Using a binary search strategy, the attacker can then ask a series of questions about this character. In this case, the attacker is asking if the ASCII value of the character is greater-than or less-than-or-equal-to the value of X. If the value is greater, the attacker knows this by observing an additional 5 second delay in the response of the database. The attacker can then use a binary search by varying the value of X to identify the value of the first character. [ 2]

**Alternate Encodings Attack** : Evading detection.
Description: In this attack, the injected text is modified so as to avoid detection by defensive coding practices and also many automated prevention techniques. This attack type is used in conjunction

with other attacks. In other words, alternate encodings do not provide any unique way to attack an application; they are simply an enabling technique that allows attackers to evade detection and prevention techniques and exploit vulnerabilities that might not otherwise be exploitable. These evasion techniques are often necessary because a common defensive coding practice is to scan for certain known “bad characters,” such as single quotes and comment operators. To evade this defense, attackers have employed alternate methods of encoding their attack strings (e.g., using hexadecimal, ASCII, and Unicode character encoding). Common scanning and detection techniques do not try to evaluate all specially encoded strings, thus allowing these attacks to go undetected. The application may scan for certain types of escape characters that represent alternate encodings in its language domain. Another layer (e.g., the database) may use different escape characters or even completely different ways of encoding. For example, a database could use the expression char(120) to represent an alternately-encoded character “x”, but char(120) has no special meaning in the application language’s context. An effective code-based defense against alternate encodings is difficult to implement in practice because it requires developers to consider of all of the possible encodings that could affect a given query string as it passes through the different application layers. Therefore, attackers have been very successful in using alternate encodings to conceal their attack strings.
Example: Because every type of attack could be represented using an alternate encoding, here we simply provide an example (see [18]) of how esoteric an alternatively-encoded attack could appear. In this attack, the following text is injected into the login field: “legalUser’; exec(0x73687574646f776e) - - ”. The resulting query generated by the application is:

```
SELECT accounts FROM users WHERE
login=’legalUser’;nexec(char(0x73687574646f776e))
-- AND pass=’’ AND pin=
```

This example makes use of the char() function and of ASCII hexadecimal encoding. The char() function takes as a parameter an integer or hexadecimal encoding of a character and returns an instance of that character. The stream of numbers in the second part of the injection is the ASCII hexadecimal encoding of the string “SHUTDOWN.” Therefore, when the query is interpreted by the database, it would result in the execution, by the database, of the SHUTDOWN command. [1][18]

## 3. PREVENTION OF SQLIAS

Researchers have proposed a wide range of techniques to address the problem of SQL injection. These techniques range from development best practices to fully automated frameworks for detecting and preventing SQLIAs. In this section, we review these proposed techniques and summarize the advantages and disadvantages associated with each technique.

### 3.1 Defensive Coding Practices

The root cause of SQL injection vulnerabilities is insufficient input validation. Therefore, the straightforward solution for eliminating these vulnerabilities is to apply suitable defensive coding practices. Here, we summarize some of the best practices proposed in the literature for preventing SQL injection vulnerabilities.

**Input type checking:** SQLIAs can be performed by injecting commands into either a string or numeric parameter. Even a simple check of such inputs can prevent many attacks. For example, in the case of numeric inputs, the developer can simply reject any input that contains characters other than digits. Many developers omit this kind of check by accident because user input is almost always represented in the form of a string, regardless of its content or intended use.

**Encoding of inputs**: Injection into a string parameter is often accomplished through the use of meta-characters that trick the SQL parser into interpreting user input as SQL tokens. While it is possible to prohibit any usage of these meta-characters, doing so would restrict a non-malicious user’s ability to specify legal inputs that contain such characters. A better solution is to use functions that encode a string in such a way that all meta-characters are specially encoded and interpreted by the database as normal characters.

**Positive pattern matching:** Developers should establish input validation routines that identify good input as opposed to bad input. This approach is generally called positive validation, as opposed to negative validation, which searches input for forbidden patterns or SQL tokens. Because developers might not be able to envision every type of attack that could be launched against their application, but should be able to specify all the forms of legal input, positive validation is a safer way to check inputs.
**Identification of all input sources:** Developers must check all input to their application. There are many possible sources of input to an application. If used to construct a query, these input sources can be a way

for an attacker to introduce an SQLIA. Simply put, all input sources must be checked. Although defensive coding practices remain the best way to prevent SQL injection vulnerabilities, their application is problematic in practice. Defensive coding is prone to human error and is not as rigorously and completely applied as automated techniques. While most developers do make an effort to code safely, it is extremely difficult to apply defensive coding practices rigorously and correctly to all sources of input. In fact, many of the SQL injection vulnerabilities discovered in real applications are due to human errors: developers forgot to add checks or did not perform adequate input validation [20][ 23][33]. In other words, in these applications, developers were making an effort to detect and prevent SQLIAs, but failed to do so adequately and in every needed location. These examples provide further evidence of the problems associated with depending on developer's use of defensive coding. Moreover, approaches based on defensive coding are weakened by the widespread promotion and acceptance of so-called "pseudoremedies" [18]. We discuss two of the most commonly-proposed pseudo-remedies. The first of such remedies consists of checking user input for SQL keywords, such as "FROM," "WHERE," and "SELECT," and SQL operators, such as the single quote or comment operator. The rationale behind this suggestion is that the presence of such keywords and operators may indicate an attempted SQLIA. This approach clearly results in a high rate of false positives because, in many applications, SQL keywords can be part of a normal text entry, and SQL operators can be used to express formulas or even names (e.g., O'Brian). The second commonly suggested pseudo-remedy is to use stored procedures or prepared statements to prevent SQLIAs. Unfortunately, stored procedures and prepared statements can also be vulnerable to SQLIAs unless developers rigorously apply defensive coding guidelines. Interested readers may refer to [1][ 25][ 28][29] for examples of how these pseudo-remedies can be subverted.

### 3.2 Detection and Prevention Techniques

Researchers have proposed a range of techniques to assist developers and compensate for the shortcomings in the application of defensive coding.

**Black Box Testing**: Huang and colleagues [19] proposeWAVES, a black-box technique for testing Web applications for SQL injection vulnerabilities. The technique uses a Web crawler to identify all points in a Web application that can be used to inject SQLIAs. It then builds attacks that target such points based on a specified list of patterns and attack techniques. WAVES then monitors the application's response to the attacks and uses machine learning techniques to improve its attack methodology. This technique improves over most penetration-testing techniques by using machine learning approaches to guide its testing. However, like all black-box and penetration testing techniques, it cannot provide guarantees of completeness.

**Static Code Checkers**: JDBC-Checker is a technique for statically checking the type correctness of dynamically-generated SQL queries [12][13]. This technique was not developed with the intent of detecting and preventing general SQLIAs, but can nevertheless be used to prevent attacks that take advantage of type mismatches in a dynamically-generated query string. JDBC-Checker is able to detect one of the root causes of SQLIA vulnerabilities in code— improper type checking of input. However, this technique would not catch more general forms of SQLIAs because most of these attacks consist of syntactically and type correct queries. Wassermann and Su propose an approach that uses static analysis combined with automated reasoning to verify that the SQL queries generated in the application layer cannot contain a tautology. The primary drawback of this technique is that its scope is limited to detecting and preventing tautologies and cannot detect other types of attacks.

**Combined Static and Dynamic Analysis:** AMNESIA is a model-based technique that combines static analysis and runtime monitoring [17][16]. In its static phase, AMNESIA uses static analysis to build models of the different types of queries an application can legally generate at each point of access to the database. In its dynamic phase, AMNESIA intercepts all queries before they are sent to the database and checks each query against the statically built models. Queries that violate the model are identified as SQLIAs and prevented from executing on the database. In their evaluation, the authors have shown that this technique performs well against SQLIAs. The primary limitation of this technique is that its success is dependent on the accuracy of its static analysis for building query models. Certain types of code obfuscation or query development techniques could make this step less precise and result in both false positives and false negatives. Similarly, two recent related approaches, SQLGuard [6] and SQLCheck also check queries at runtime to see if they conform to a model of expected queries. In these approaches, the model is expressed as a grammar that only accepts legal queries. In SQLGuard,the model is deduced at runtime by examining the structure of the query before and after the addition of user-input. In SQLCheck, the model is specified independently by the developer. Both approaches use a secret key to delimit user input during parsing by the runtime checker, so security of the approach is dependent on attackers not being able to discover the key. Additionally, the use of these two approaches requires the developer to either rewrite code to use a special intermediate library or manually insert special markers into the code where user input is added to a dynamically generated query.

**Taint Based Approaches:** WebSSARI detects input-validation related errors using information flow analysis [20]. In this approach, static analysis is used to check taint flows against preconditions for sensitive functions. The analysis detects the points in which preconditions have not been met and can suggest filters and sanitization functions that can be automatically added to the application to satisfy these preconditions. The WebSSARI system works by considering as sanitized input that has passed through a predefined set of filters. In their evaluation, the authors were able to detect security vulnerabilities in a range of existing applications. The primary drawbacks of this technique are that it assumes that adequate preconditions for sensitive functions can be accurately expressed using their typing system and that having input passing through certain types of filters is sufficient to consider it not tainted. Formany types of functions and applications, this assumption is too strong. Livshits and Lam [23] use static analysis techniques to detect vulnerabilities in software. The basic approach is to use information flow techniques to detect when tainted input has been used to construct an SQL query. These queries are then flagged as SQLIA vulnerabilities. The authors demonstrate the viability of their technique by using this approach to find security vulnerabilities in a benchmark suite. The primary limitation of this approach is that it can detect only known patterns of SQLIAs and, because it uses a conservative analysis and has limited support for untainting operations, can generate a relatively high amount of false positives. Several dynamic taint analysis approaches have been proposed. Two similar approaches by Nguyen-Tuong and colleagues and Pietraszek and Berghe modify a PHP interpreter to track precise per-character taint information. The techniques use a context sensitive analysis to detect and reject queries if untrusted input has been used to create certain types of SQL tokens. A common drawback of these two approaches is that they require modifications to the runtime environment, which affects portability. A technique by Haldar and colleagues [15] and SecuriFly [26] implement a similar approach for Java. However, these techniques do not use the context sensitive analysis employed by the other two approaches and track taint information on a per-string basis (as opposed to percharacter). SecuriFly also attempts to sanitize query strings that have been generated using tainted input. However, this sanitization approach does not help if injection is performed into numeric fields. In general, dynamic taint-based techniques have shown a lot of promise in their ability to detect and prevent SQLIAs. The primary drawback of these approaches is that identifying all sources of tainted user input in highly-modular Web applications and accurately propagating taint information is often a difficult task.

**NewQueryDevelopmentParadigms:** Two recent approaches, SQL DOM [27] and Safe Query Objects [7], use encapsulation of database queries to provide a safe and reliable way to access databases. These techniques offer an effective way to avoid the SQLIA problem by changing the query-building process from an unregulated one that uses string concatenation to a systematic one that uses a type-checked API.Within their API, they are able to systematically apply coding best practices such as input filtering and rigorous type checking of user input. By changing the development paradigm in which SQL queries are created, these techniques eliminate the coding practices that make most SQLIAs possible. Although effective, these techniques have the drawback that they require developers to learn and use a new programming paradigm or query-development process. Furthermore, because they focus on using a new development process, they do not provide any type of protection or improved security for existing legacy systems.

**Intrusion Detection Systems:** Valeur and colleagues [29] propose the use of an Intrusion Detection System(IDS) to detect SQLIAs. Their IDS system is based on a machine learning technique that is trained using a set of typical application queries. The technique builds models of the typical queries and then monitors the application at runtime to identify queries that do not match the model. In their evaluation, Valeur and colleagues have shown that their system is able to detect attacks with a high rate of success. However, the fundamental limitation of learning based techniques is that they can provide no guarantees about their detection abilities because their success is dependent on the quality of the training set used. A poor training set would cause the learning technique to generate a large number of false positives and negatives.

**Proxy Filters:** Security Gateway [28] is a proxy filtering system that enforces input validation rules on the data flowing to a Web application. Using their Security Policy Descriptor Language (SPDL), developers provide constraints and specify transformations to be applied to application parameters as they flow from the Web page to the application server. Because SPDL is highly expressive, it allows developers considerable freedom in expressing their policies. However, this approach is human-based and, like defensive programming, requires developers to know not only which data needs to be filtered, but also what patterns and filters to apply to
the data.

**Instruction Set Randomization:** SQLrand [5] is an approach based on instruction-set randomization. SQLrand provides a framework that allows developers to create queries using randomized instructions instead of normal SQL keywords. A proxy filter intercepts queries to the database and de-randomizes the keywords. SQL code injected by an attacker would not have been constructed using the randomized instruction set. Therefore, injected commands would result in a syntactically incorrect query. While this technique can be very effective, it has several practical drawbacks. First, since it uses a secret key to modify instructions, security of the approach is dependent on attackers not being able to discover the key. Second, the approach imposes a significant infrastructure overhead because it require the integration of a proxy for the database in the system.

## 4. TECHNIQUES EVALUATION

In this section, we evaluate the techniques presented in Section 3 using several different criteria. We first consider which attack types each technique is able to address. For the subset of techniques that are based on code improvement, we look at which defensive coding practices the technique helps enforce. We then identify which injection mechanism each technique is able to handle. Finally, we evaluate the deployment requirements of each technique.

### 4.1 Evaluation with Respect to Attack Types

We evaluated each proposed technique to assess whether it was capable of addressing the different attack types presented in Section 2. For most of the considered techniques, we did not have access to an implementation because either the technique was not implemented or its implementation was not available. Therefore, we evaluated the techniques analytically, as opposed to evaluating them against actual attacks. For developer-based techniques, that is, those that required developer intervention, we assumed that the developers were able to correctly apply all required defensive coding practices. In other words, our assessment of these techniques is optimistic compared to what their performance may be in practice. In our tables, we denote developer-based techniques with the symbol “*”. For the purposes of the comparison, we divide the techniques into two groups: prevention-focused and detection-focused techniques. Prevention-focused techniques are techniques that statically identify vulnerabilities in the code, propose a different development paradigm for applications that generate SQL queries, or add checks to the application to enforce defensive coding best practices. Detection-focused techniques are techniques that detect attacks mostly at runtime. Tables 1 and 2 summarize the results of our evaluation. We use four different types of markings to indicate how a technique performed with respect to a given attack type. We use the symbol “•” to denote that a technique can successfully stop all attacks of that type. Conversely, we use the symbol “×” to denote that a technique is not able to stop attacks of that type. We used two different symbols to classify techniques that are only partially effective. The symbol “◦” denotes a technique that can address the attack type considered, but cannot provide any guarantees of completeness. An example of one such technique would be a black-box testing technique such as WAVES [19] or the IDS based approach from Valeur and colleagues [29]. The symbol “−,” denotes techniques that address the attack type considered only partially because of intrinsic limitations of the underlying approach. For example, JDBCChecker [12][13] detects type-related errors that enable SQL injection vulnerabilities. However, because type-related errors are only one of the many possible causes of SQL injection vulnerabilities, this approach is classified as only partially handling each attack type. Half of the prevention-focused techniques effectively handle all of the attack types considered. Some techniques are only partially effective: JDBC-Checker by definition addresses only a subset of SQLIAs; Security Gateway, because it cannot handle all of the injection sources cannot completely address all of the attack profiles; SecuriFly, because its prevention method is to escape all SQL meta-characters, which still would allow injection into numeric fields; and WAVES, which because it is a testing based technique, cannot provide guarantees as to its completeness. We believe that, overall, the prevention-focused techniques performed well because they incorporate the defensive coding practices in their prevention mechanisms. See Section 4.4 for further discussion on this topic. Most of the detection-focused techniques perform fairly uniformly against the various attack types. The three exceptions are the IDSbased approach by Valeur and colleagues [29], whose effectiveness depends on the quality of the training set used, Java Dynamic Tainting [15], whose performance is negatively affected by the fact that its untainting operations allow input to be used without regard to the quality of the check, and Tautology-checker, which by definition can only address tautology-based attacks. Two attack types, stored procedures and alternate encodings, caused problems for most techniques. With stored procedures, the code that generates the query is stored and executed on the database. Most of the techniques considered focused only on queries generated within the application. Expanding the techniques to also encompass the queries generated and executed on the database is not straightforward and would, in general, require substantial effort. For this reason, attacks based on stored procedures are problematic for many techniques. Attacks based on alternate encoding are also difficult to handle. Only three techniques, AMNESIA, SQLCheck, and SQLGuard explicitly address these types of attacks. The reason why these techniques are successful against such attacks is that they use the database lexer or parser to interpret a query string in the same way that the database would. Other techniques that score well in this category are either developer-based techniques (i.e., Java Static Tainting and WebSSARI) or techniques that address the problem by using a standard API (i.e., SQL DOM and Safe Query Objects). It is important to note that we did not take precision into account in our evaluation. Many of the techniques that we consider are based on some conservative analysis or assumptions that may result in false positives. However, because we do not have an accurate way to classify the accuracy of such techniques, short of implementing all of them and assessing their performance on a large set of legitimate inputs, we have not considered this characteristic in our assessment.

### 4.2 Evaluation with Respect to Injection Mechanisms

We assessed each of the techniques with respect to their handling of the various injection mechanisms that we defined in Section 1.1. Although most of the techniques do not specifically address all of those injection mechanisms, all but two of them could be easily extended to handle all such mechanisms. The two exceptions are Security Gateway and WAVES. Security Gateway can examine only URL parameters and cookie fields. Because it resides on the network between the application and the attacker, it cannot examine server variables and second-order injection sources, which do not pass through the gateway. WAVES can only address injection through user input because it only generates attacks that can be submitted to the application via the Web page forms.

### 4.3 Evaluation with Respect to Deployment Requirements

Each of the techniques have different deployment requirements. To determine the effort and infrastructure required to use the technique, we examined the author's description of the technique and its current implementation. We evaluated each technique with respect to the following criteria: (1) Does the technique require developers to modify their code base? (2) What is the degree of automation of the detection aspect of the approach? (3) What is the degree of automation of the prevention aspect of the approach? (4) What infrastructure (not including the tool itself) is needed to successfully use the technique? The results of this classification are summarized in Table 3.

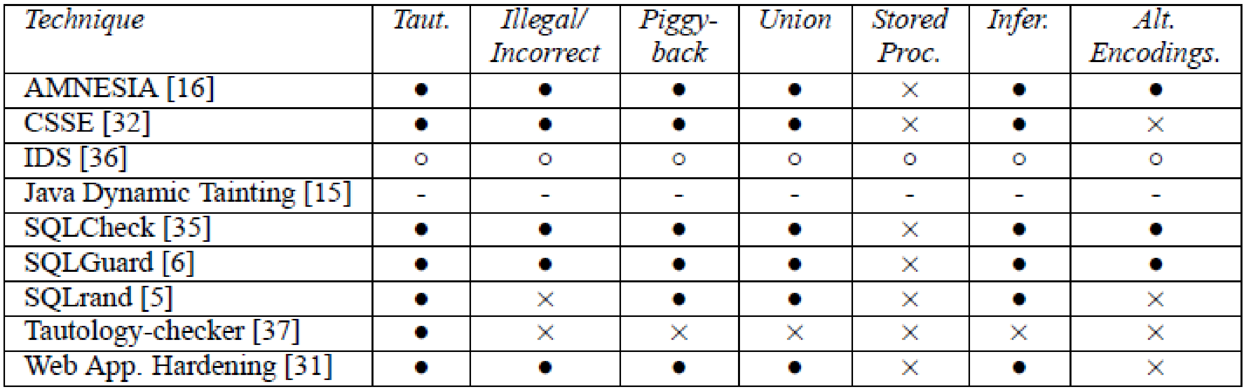
### 4.4 Evaluation of PreventionFocused Techniques with Respect to Defensive Coding Practices

Our initial evaluation of the techniques against the various attack types indicates that the prevention-focused techniques perform very well against most of these attacks. We hypothesize that this result is due to the fact that many of the prevention techniques are actually applying defensive coding best practices to the code base. Therefore, we examine each of the prevention-focused techniques and classify them with respect to the defensive coding practice that they enforce. Not surprisingly, we find that these techniques enforce many of these practices. Table 4 summarizes, for each technique, which of the defensive coding practices it enforces.

## 5. Future work and Conclusion:

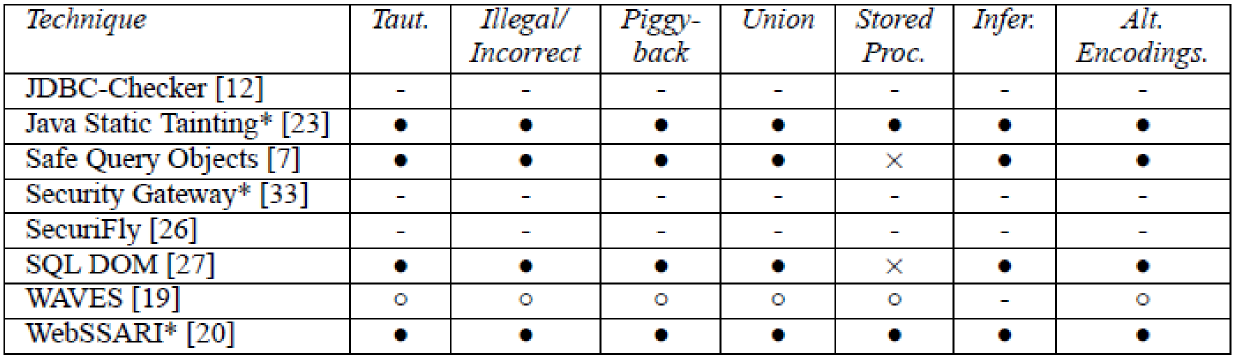
In this paper, we have presented a survey and comparison of current techniques for detecting and preventing SQLIAs. To perform this evaluation, we first identified the various types of SQLIAs known to date. We then evaluated the considered techniques in terms of their ability to detect and/or prevent such attacks. We also studied the different mechanisms through which SQLIAs can be introduced into an application and identified which techniques were able to handle which mechanisms. Lastly, we summarized the deployment requirements of each technique and evaluated to what extent its detection and prevention mechanisms could be fully automated. Our evaluation found several general trends in the results. Many of the techniques have problems handling attacks that take advantage of poorly-coded stored procedures and cannot handle attacks that disguise themselves using alternate encodings. We also found a general distinction in prevention abilities based on the difference between prevention-focused and general detection and prevention techniques. Section 4.4 suggests that this difference could be explained by the fact that prevention-focused techniques try to incorporate defensive coding best practices into their attack prevention mechanisms. Future evaluation work should focus on evaluating the techniques' precision and effectiveness in practice. Empirical evaluations such as those presented in related work (e.g., [17]) would allow for comparing the performance of the different techniques when they are subjected to real-world attacks and legitimate inputs.

**Table 1. Comparison of detection-focused techniques with respect to attack types.**



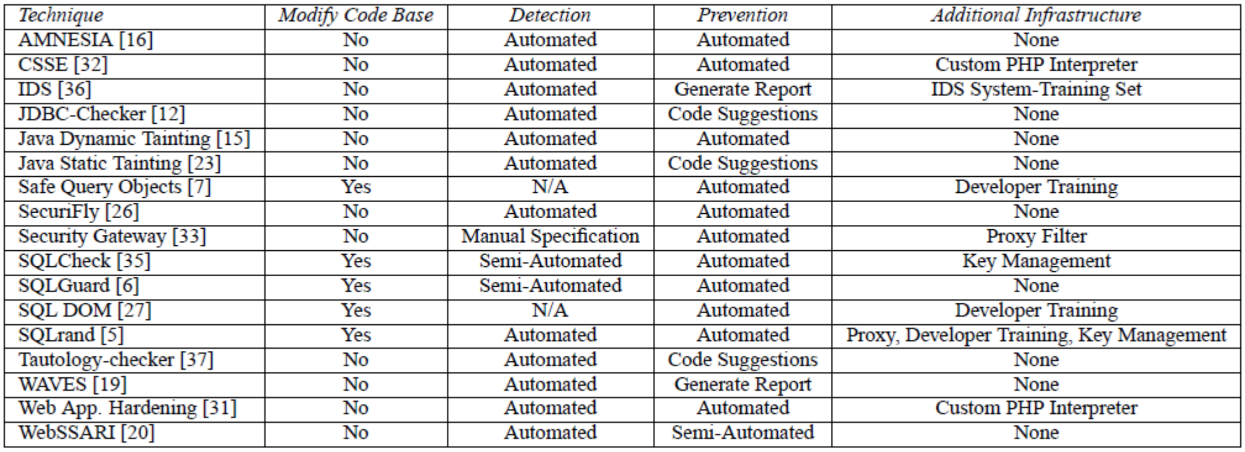
| *Technique* | *Taut.* | *Illegal/ Incorrect* | *Piggy- back* | *Union* | *Stored Proc.* | *Infer.* | *Alt. Encodings.* |
|---|---|---|---|---|---|---|---|
| AMNESIA [16] | ● | ● | ● | ● | × | ● | ● |
| CSSE [32] | ● | ● | ● | ● | × | ● | × |
| IDS [36] | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Java Dynamic Tainting [15] | - | - | - | - | - | - | - |
| SQLCheck [35] | ● | ● | ● | ● | × | ● | ● |
| SQLGuard [6] | ● | ● | ● | ● | × | ● | ● |
| SQLrand [5] | ● | × | ● | ● | × | ● | × |
| Tautology-checker [37] | ● | × | × | × | × | × | × |
| Web App. Hardening [31] | ● | ● | ● | ● | × | ● | × |

**Table 2: Comparison of prevention-focused techniques with respect to attack types.**



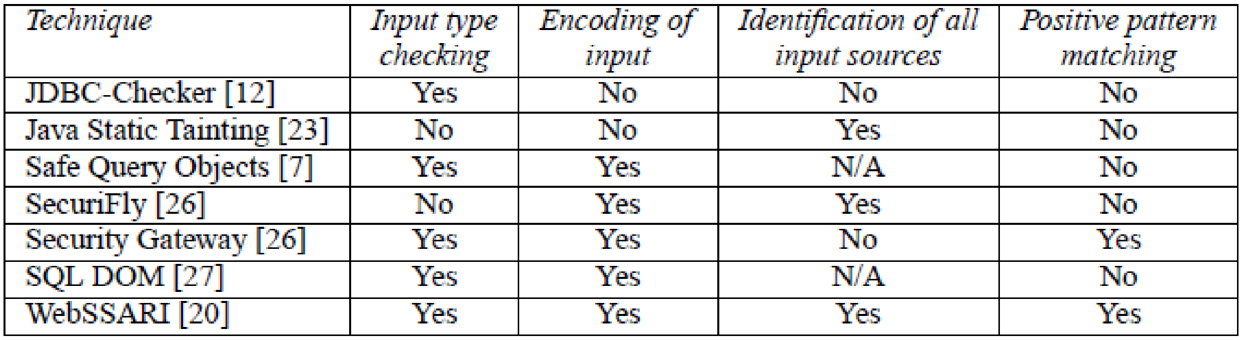
| *Technique* | *Taut.* | *Illegal/ Incorrect* | *Piggy- back* | *Union* | *Stored Proc.* | *Infer.* | *Alt. Encodings.* |
|---|---|---|---|---|---|---|---|
| JDBC-Checker [12] | - | - | - | - | - | - | - |
| Java Static Tainting* [23] | ● | ● | ● | ● | ● | ● | ● |
| Safe Query Objects [7] | ● | ● | ● | ● | × | ● | ● |
| Security Gateway* [33] | - | - | - | - | - | - | - |
| SecuriFly [26] | - | - | - | - | - | - | - |
| SQL DOM [27] | ● | ● | ● | ● | × | ● | ● |
| WAVES [19] | ○ | ○ | ○ | ○ | ○ | - | ○ |
| WebSSARI* [20] | ● | ● | ● | ● | ● | ● | ● |

**Table 3: Comparison of techniques with respect to deployment requirements.**



| *Technique* | *Modify Code Base* | *Detection* | *Prevention* | *Additional Infrastructure* |
|---|---|---|---|---|
| AMNESIA [16] | No | Automated | Automated | None |
| CSSE [32] | No | Automated | Automated | Custom PHP Interpreter |
| IDS [36] | No | Automated | Generate Report | IDS System-Training Set |
| JDBC-Checker [12] | No | Automated | Code Suggestions | None |
| Java Dynamic Tainting [15] | No | Automated | Automated | None |
| Java Static Tainting [23] | No | Automated | Code Suggestions | None |
| Safe Query Objects [7] | Yes | N/A | Automated | Developer Training |
| SecuriFly [26] | No | Automated | Automated | None |
| Security Gateway [33] | No | Manual Specification | Automated | Proxy Filter |
| SQLCheck [35] | Yes | Semi-Automated | Automated | Key Management |
| SQLGuard [6] | Yes | Semi-Automated | Automated | None |
| SQL DOM [27] | Yes | N/A | Automated | Developer Training |
| SQLrand [5] | Yes | Automated | Automated | Proxy, Developer Training, Key Management |
| Tautology-checker [37] | No | Automated | Code Suggestions | None |
| WAVES [19] | No | Automated | Generate Report | None |
| Web App. Hardening [31] | No | Automated | Automated | Custom PHP Interpreter |
| WebSSARI [20] | No | Automated | Semi-Automated | None |

**Table 4: Evaluation of Code Improvement Techniques with Respect to Common Development Errors.**



| *Technique* | *Input type checking* | *Encoding of input* | *Identification of all input sources* | *Positive pattern matching* |
|---|---|---|---|---|
| JDBC-Checker [12] | Yes | No | No | No |
| Java Static Tainting [23] | No | No | Yes | No |
| Safe Query Objects [7] | Yes | Yes | N/A | No |
| SecuriFly [26] | No | Yes | Yes | No |
| Security Gateway [26] | Yes | Yes | No | Yes |
| SQL DOM [27] | Yes | Yes | N/A | No |
| WebSSARI [20] | Yes | Yes | Yes | Yes |

## 6. References

[1] C. Anley. Advanced SQL Injection In SQL Server Applications. White paper, Next Generation Security Software Ltd., 2002.

[2] C. Anley. (more) Advanced SQL Injection. White paper, Next Generation Security Software Ltd., 2002.

[3] D. Aucsmith. Creating and Maintaining Software that Resists Malicious Attack. http://www.gtisc.gatech.edu/bio aucsmith.html, September 2004. Distinguished Lecture Series.

[4] F. Bouma. Stored Procedures are Bad, O'kay? Technical report, Asp.Net Weblogs, November 2003. http://weblogs.asp. net/fbouma/archive/2003/11/18/38178.aspx.

[5] S. W. Boyd and A. D. Keromytis. SQLrand: Preventing SQL Injection Attacks. In Proceedings of the 2nd Applied Cryptography and Network Security (ACNS) Conference, pages 292–302, June 2004.

[6] G. T. Buehrer, B. W. Weide, and P. A. G. Sivilotti. Using Parse Tree Validation to Prevent SQL Injection Attacks. In International Workshop on Software Engineering and Middleware (SEM), 2005.

[7] W. R. Cook and S. Rai. Safe Query Objects: Statically Typed Objects as Remotely Executable Queries. In Proceedings of the 27th International Conference on Software Engineering (ICSE 2005), 2005.

[8] M. Dornseif. Common Failures in Internet Applications, May 2005. http://md.hudora.de/presentations/ 2005-common-failures/ dornseif-common-failures-2005-05-25.pdf.

[9] E. M. Fayo. Advanced SQL Injection in Oracle Databases. Technical report, Argeniss Information Security, Black Hat Briefings, Black Hat USA, 2005.

[10] P. Finnigan. SQL Injection and Oracle - Parts 1 & 2. Technical Report, Security Focus, November 2002. http://securityfocus.com/infocus/1644, http://securityfocus.com/infocus/1646.

[11] T. O. Foundation. Top Ten Most Critical Web Application Vulnerabilities, 2005. http: //www.owasp.org/documentation/topten.html.

[12] C. Gould, Z. Su, and P. Devanbu. JDBC Checker: A Static Analysis Tool for SQL/JDBC Applications. In Proceedings of the 26th International Conference on Software Engineering (ICSE 04) – Formal Demos, pages 697–698, 2004.

[13] C. Gould, Z. Su, and P. Devanbu. Static Checking of Dynamically Generated Queries in Database Applications. In Proceedings of the 26th International Conference on Software Engineering (ICSE 04), pages 645–654, 2004.

[14] N. W. Group. RFC 2616 – Hypertext Transfer Protocol – HTTP/1.1. Request for comments, The Internet Society, 1999.

[15] V. Haldar, D. Chandra, and M. Franz. Dynamic Taint Propagation for Java. In Proceedings 21st Annual Computer Security Applications Conference, Dec. 2005.

[16] W. G. Halfond and A. Orso. AMNESIA: Analysis and Monitoring for NEutralizing SQL-Injection Attacks. In Proceedings of the IEEE and ACM International

Conference on Automated Software Engineering (ASE 2005), Long Beach, CA, USA, Nov 2005.

[17] W. G. Halfond and A. Orso. Combining Static Analysis and Runtime Monitoring to Counter SQL-Injection Attacks. In Proceedings of the Third International ICSE Workshop on Dynamic Analysis (WODA 2005), pages 22–28, St. Louis, MO, USA, May 2005.

[18] M. Howard and D. LeBlanc. Writing Secure Code. Microsoft Press, Redmond, Washington, second edition, 2003.

[19] Y. Huang, S. Huang, T. Lin, and C. Tsai. Web Application Security Assessment by Fault Injection and Behavior Monitoring. In Proceedings of the 11th International World Wide Web Conference (WWW 03), May 2003.

[20] Y. Huang, F. Yu, C. Hang, C. H. Tsai, D. T. Lee, and S. Y. Kuo. Securing Web Application Code by Static Analysis and Runtime Protection. In Proceedings of the 12th International World Wide Web Conference (WWW 04), May 2004.

[21] S. Labs. SQL Injection. White paper, SPI Dynamics, Inc., 2002. http://www.spidynamics.com/assets/documents/ WhitepaperSQLInjection.pdf.

[22] D. Litchfield. Web Application Disassembly with ODBC Error Messages. Technical document, @Stake, Inc., 2002. http://www.nextgenss.com/papers/webappdis.doc.

[23] V. B. Livshits and M. S. Lam. Finding Security Errors in Java Programs with Static Analysis. In Proceedings of the 14th Usenix Security Symposium, pages 271–286, Aug. 2005.

[24] C. A. Mackay. SQL Injection Attacks and Some Tips on How to Prevent Them. Technical report, The Code Project, January 2005. http://www.codeproject.com/cs/database/ SqlInjectionAttacks.asp.

[25] O. Maor and A. Shulman. SQL Injection Signatures Evasion. White paper, Imperva, April 2004. http://www.imperva.com/ application defense center/white papers/ sql injection signatures evasion.html.

[26] M. Martin, B. Livshits, and M. S. Lam. Finding Application Errors and Security Flaws Using PQL: A Program Query Language. In Proceedings of the 20th annual ACM SIGPLAN conference on Object oriented programming systems languages and applications (OOPSLA 2005), pages 365–383, 2005.

[27] R. McClure and I. Kr¨uger. SQL DOM: Compile Time Checking of Dynamic SQL Statements. In Proceedings of the 27th International Conference on Software Engineering (ICSE 05), pages 88–96, 2005.

[28] S. McDonald. SQL Injection: Modes of attack, defense, and why it matters. White paper, GovernmentSecurity.org, April 2002. http://www.governmentsecurity.org/articles/ SQLInjectionModesofAttackDefenceandWhyIt Matters.php.

[29] F. Valeur, D. Mutz, and G. Vigna. A Learning-Based Approach to the Detection of SQL Attacks. In Proceedings of the Conference on Detection of Intrusions and Malware and Vulnerability Assessment (DIMVA), Vienna, Austria, July 2005.

# Removing Redundancy in Dictionary based Compression Techniques

Neha Gupta[1], Ranjit Kumar[2] and Apoorv Gupta[3]

[1]Gateway Institute of Engineering and Technology, MDU Rohtak, India
[2]Gateway Institute of Engineering and Technology, MDU Rohtak, India
[3]Technological Institute of Technology & Sciences, MDU Rohtak, India
[1]guptaneha2006@gmail.com, [2]ranjitpes@gmail.com, [3]apoorv.gupta@infosys.com

***Abstract:*** Many data compression schemes are developed nowadays and they are selected according to the requirements, such as fast encoding, fast decoding, a good compression performance, small amount of required memory etc. In this thesis, the basic dictionary based data compression techniques i.e. LZ77, LZ78 and LZW, have been studied to find their drawbacks, so that they can be improved further. As out of LZ77, LZ78 and LZW, the variants of LZW are widely used in a number of applications. So, the thesis is mainly oriented towards improving on LZW. Based on the study, we have tried to identify the sources of redundancy in these algorithms and have suggested a method which is a simple dictionary-pruning algorithm that removes the irrelevant entries from the dictionary every time the dictionary is out of space; to store new phrases. This ensures that the dictionary is always adaptive.

***Keyword:*** Compression, Decompression, Dictionary-based, LZW, Pruning, Performance.

## 1. Introduction

Data compression is, in the context of computer science, the science (and art) of representing information in a compact form. It has been one of the critical enabling technologies for the ongoing digital multimedia revolution for decades. Most people frequently use data compression software such as zip, gzip and WinZip (and many others) to reduce the file size before storing or transferring it in media. There are two major families of compression techniques when considering the possibility of reconstructing exactly the original source. They are called *lossless* and *lossy* compression. A compression approach is lossless only if it is possible to exactly reconstruct the original data from the compressed version. A compression method is lossy if it is not possible to reconstruct the original exactly from the compressed version. Lossless data compression is generally implemented using one of two different types of modeling: statistical or dictionary-based. Statistical modeling reads in and encodes a single symbol at a time using the probability of that character's appearance. Dictionary-based modeling uses a single code to replace strings of symbols. In dictionary-based modeling, the coding problem is reduced in significance, leaving the model supremely important.

## 2. Dictionary-based modeling (LZW)

The LZW method starts by initializing the dictionary to all the symbols in the alphabet. Then the encoder inputs symbols one by one and accumulates them in a string '*word*'. After each symbol is input and is concatenated to '*word*', the dictionary is searched for string '*word*'. As long as '*word*' is found in the dictionary, the process continues. At a certain point, adding the next symbol '*x*' causes the search to fail; string '*word*' is in the dictionary but string '*word*' + '*x*' (symbol '*x*' concatenated to '*word*') is not. At this point the encoder outputs the dictionary pointer that points to string '*word*', Saves string '*word*' + '*x*' (which is now called a *phrase*) in the next available dictionary entry, and Initializes string '*word*' to symbol '*x*'.

Since the first 256 entries of the dictionary are occupied right from the start, pointers to the dictionary have to be longer than 8 bits. A simple implementation would typically use 16-bit pointers, which allow for a 64K-entry dictionary (where 64K = 216 = 65,536). Such a dictionary will, of course, fill up very quickly in all but the smallest compression jobs. Another interesting fact about LZW is that strings in the dictionary get only one character longer at a time. It therefore takes a long time to get long strings in the dictionary, and thus a chance to achieve really good compression. We can say that LZW adapts slowly to its input data.

The encoding algorithm is:

```
word ← ''
while not EOF do
   x ← read_next_character()
   if word + x is in the dictionary then
      word ← word + x
   else
      output the dictionary index for word
      add word + x to the dictionary
      word ← x
   end if
end while
output the dictionary index for word
```

The decoding algorithm now is:
read a codeword *x* from the compressed file

```
look up dictionary for phrase at x
output phrase
word ← phrase
while not EOF do
  read x
  look up dictionary for phrase at x
  if there is no entry yet for index x then
    phrase ← word + firstCharOfword
  end if
  output phrase
  add word + firstCharOfphrase to the dictionary
  word ← phrase
end while
```

## 3. The Dictionary Pruning Algorithm-LWZ(P)-Proposed Work

In this section a method for dictionary pruning has been proposed. As LZW is a very popular dictionary based data compression technique, modification attributes to include our pruning process.

In LZW, phrases from input string are added to dictionary and corresponding 12 bit codes are sent to the output. So, an LZW dictionary can contain maximum of $2^{12} = 4096$ entries. The basic LZW algorithm is modified in such a way that whenever the dictionary gets full, a function is called that will remove all the entries that have never been used till time, since the creation of dictionary. The main work of the function is to identify these phrases. For this, every entry in dictionary is associated with a flag value. The function checks every phrase for its flag value, and removes it if the flag value matches the deletion condition. Values of flag variable according to specific condition are:

$$\text{dict[i].flag} = \left.\begin{matrix} 0 \\ 1 \\ 2 \end{matrix}\right| \begin{matrix} \text{unused entries} \\ \text{entry used at least once} \\ \text{deleted entry} \end{matrix}$$

### 3.1 Assumptions

**Table 1.** Assumptions table for dictionary pruning algorithm.

| Symbol | Meaning |
|---|---|
| word | string that contains all the characters that have been scanned till time and should be searched in the dictionary |
| x | next character to be scanned from the input file |
| size | number of phrases that are currently present in the dictionary |

### 3.2 Algorithm

The pruning process algorithm work as follows:

a) Scan the input string, character by character until the end of file is reached.

b) After each character '*x*' is input, it is concatenated to '*word*', and the dictionary is searched for string '*word*'.

c) As long as '*word*' is found in the dictionary, the search process continues.

d) At a certain point, adding the next symbol '*x*' causes the search to fail; string '*word*' is in the dictionary but string '*word*' + '*x*' (symbol '*x*' concatenated to '*word*') is not.

e) At this point the encoder outputs the dictionary pointer that points to string '*word*', and saves string '*word*' + '*x*' (which is now called a *phrase*) in the next available dictionary entry, and initializes string '*word*' to symbol '*x*'.

f) This process continues until the dictionary is full i.e. all 4096 locations have been occupied by phrases.

As the dictionary overflows, all the entries having flag = 0 i.e. the phrases whose code has never been used in the output, are searched and removed from the dictionary by setting the corresponding flag value to 2.

Now to insert new entries in the dictionary, the space restored during deletion, is used.

### 3.3 Pseudo-Code

```
LZW(P)
word ← "
while not EOF do
  x ← read_next_character()
  if word + x is in the dictionary then
    word ← word + x
  else
    search for the first occurring available location
    add word + x to the dictionary
    size ← size + 1
    output the dictionary index for word
    word ← x
  end if
  if size = 4096
    dict_prune()
 end while
 output the dictionary index for word
dict_prune()
for all dictionary entries do
  if flag is 0
    set flag ← 2     //marks the entry as deleted
    size ← size – 1
```

### 3.4. Advantages

The dictionary pruning algorithm proposed above has the following advantages:

a) In classic version of LZW, the dictionary becomes static when it reaches its maximum size, but the proposed algorithm remains adaptive, as whenever the dictionary reaches its maximum value, it removes the unused phrases from the dictionary.

b) The proposed algorithm improves the compression by a considerable amount, as it is ensures that dictionary contains only those phrases that will help in compression.
c) The irrelevant entries are always updated, making space for new entries that are more relevant to the input file.

## 4. Performance Evaluation

### 4.1 Comparison of LZW and LZW(P)

This section presents compression and analysis results of a classic and proposed LZW algorithm on a number of different files. Table 2 gives the details of the Datasets on which the above algorithms have been tested.

**Table 2.** Datasets

| S. No. | File Name | File Size(bytes) |
|---|---|---|
| 1 | Book1.txt | 768770 |
| 2 | Book2.txt | 610855 |
| 3 | News.txt | 377108 |
| 4 | Paper.txt | 53155 |

Figure 1 shows the overall compression achieved as the source files were processed using the two algorithms. The two different bars correspond to classic LZW and LZW(P) algorithms. The bars shows that compression using LZW(P) has a consistent advantage over LZW:

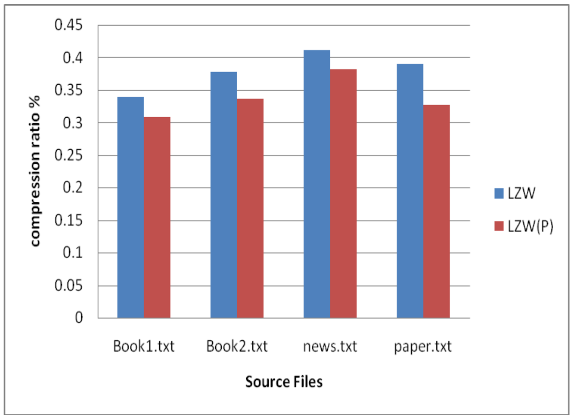


**Figure 1.** LZW (P) performance graph compared to LZW compression.

The graph shows the compression gain that LZW(P) has over LZW. In real scenario, the LZW(P) shows a gain of about 6-8% in compression ratio when tested on different files.
Some experimental results to show the effect of dictionary pruning on compression performance are given in Table 3.

**Table 3.** Performance Analysis of LZW and LZW(P).

| Source Files | Original Size (Bytes) | LZW (Bytes) | LZW(P) (Bytes) |
|---|---|---|---|
| Book1.txt | 768770 | 260536 | 237580 |
| Book2.txt | 610855 | 231020 | 205560 |
| News.txt | 377108 | 155260 | 143902 |
| Paper.txt | 53155 | 20696 | 17398 |

## 5. Conclusion

The various data compression techniques and methods to optimize them were considered. The first algorithm that is proposed takes into account the fact that the dictionary used in LZW becomes static once all the 4096 locations has been occupied. The algorithm adds a process for dictionary pruning to LZW, so that it remains adaptive. It does so by removing the entries that are irrelevant and are not required. These entries take up unnecessary dictionary space that could be utilized by more useful keywords. The proposed algorithm removes these entries whenever the dictionary is full. Waste phrases are found by associating each phrase with a flag value which is 0 for the phrases that were never used during compression. By applying this modification better compression ratios were achieved. So, by adding a little extra overhead, the proposed method achieved about 6%-8% better compression ratios than the classic LZW.

## References

[1] C. L. Yu and J. L. Wu, “Hierarchical dictionary model and dictionary management policies for data compression”, Signal Processing, Sept 1999.

[2] R. N. Horspool, “The Effect of Non-Greedy Parsing in Ziv-Lempel Compression Methods”, IEEE Data Compression Conference, 1995.

[3] S. Subathra, M. Sethuraman and J. V. B. James, “Performance Analysis of Dictionary based Data Compression Algorithms for High Speed Networks”, IEEE Indicon Conference, Dec 2005.

[4] N. Zhang, T. Tao, R. V. Satya and A. Mukherjee, “Modified LZW Algorithm for Efficient Compressed Text Retrieval”, draft, Computer Science Dept., Univ. of Central Florida, 2004.

[5] R. N. Horspool, “Improving LZW”, Research, Dept. of Computer Science, University of Victoria, Victoria, B.C, Canada.

[6] Y. Matias, N. M. Rajboot and S. C. Sahinalp, “The Effect of Flexible Parsing for Dynamic Dictionary Based Data Compression”, Proceedings of the Data Compression Conference, 1999.

[7] D. A. Huffman, “A method for the construction of minimum redundancy codes”, Proceedings IRE 40, Sept 1952.

[8] D. Salomon, “Data Compression - The Complete Reference”, 2nd Ed, Springer-Verlag New York, Inc., New York, 2001.

## Author Biographies

**Neha Gupta** Her Birth place is Punjab and Date of Birth is 2 November 1983.She is B.tech in Information Technology from TITS,Bhiwani (2005 batch) and M.Tech in Computer Science from BITS, Bhiwani (2010 batch). Now she is working as an Asst. Professor in GIET, Sonipat having 3.5 years teaching experience.

**Ranjit Kumar** His Birth place is Bihar and Date of Birth is 20 November 1981.He is B.tech in Computer Science from P.E.S, Maharashtra (2005 batch) and M.E. in Software Engineering from BIT Mesra, Ranchi (2008 batch). Now he is working as an Asst. Professor in GIET, Sonipat having 2.5 years teaching experience.

**Apoorv Gupta** His Birth place is Haryana and Date of Birth is 7 October 1987.He is B.tech in Computer Science from TITS, Bhiwani (2009 batch). Now he is working in Infosys Technology, pune.

# A Review of Secure Routing Protocol in Mobile Ad Hoc Network

Rajender Nath[1], Pankaj Kumar Sehgal[2]
[1]Deptt. of Computer Science and Applications, Kurukshetra University, Kurukshetra, Haryana, India.
[2]Department of Information Technology, MM University,Mullana(Ambala), Haryana, India
email: rnath_2k3@rediffmail.com, pankajkumar.sehgal@gmail.com

***Abstract:*** Mobile ad hoc networks (MANETs) is an emerging research area with commercial and military applications. The fundamental characteristics such as dynamic topology, open medium and distributed cooperation are combined with security threats. The routing protocols plays important and essential role in secure data transmission for entire network. This paper presents the review of some secure routing protocols for MANETs. The paper gives a comparative study of these protocols with respect to various security parameters and attacks.

*Keywords*: routing protocol, security attacks, security threats.

## 1. Introduction

A Mobile Ad hoc Network (MANET) is a collection of wireless mobile nodes forming a temporary network without any established infrastructure or centralized authority. Until now, the main research focus has been on improving the protocols for multi-hop routing, performance and scalability of the ad hoc networks [20]. Though, the performance and scalability have their place in wireless network research, the current and future applications of the ad hoc networks has forced the research community to look at dependability and security aspects of ad hoc networks. Security in an ad hoc network is essential even for basic network functions like routing and packet forwarding, since such network functions are carried out by the nodes themselves rather than specialized routers. Hence, the nodes of an ad hoc network must be trusted for the proper execution of basic network functions. The intruder in the ad hoc network can come from anywhere, along any direction and target any communication channel in the network. Compare this with a wired network where the intruder gains physical access to the wired link or pass through security holes at firewalls and routers. Since the infrastructure-free mobile ad hoc network does not have a clear line of defense, every node must be prepared for the adversary. Hence a centralized or hierarchical network security solution for the existing wired and infrastructure-based cellular wireless networks will not work properly for mobile ad hoc networks. Securing the ad hoc networks, like any other field of computers, is based on the principle of confidentiality and integrity. These principles exist in every field, but the presence of malicious nodes, covert channels and eavesdroppers in the mobile ad hoc network makes this an extremely important and challenging problem [21]. In past several years, there has been a surge of network security research in the field of information assurance that has focused on protecting the data using techniques such as authentication and encryption. These techniques are applicable in a wired and infrastructure based cellular network. In the case of infrastructure-free mobile ad hoc networks these techniques are not applicable [20]. In the infrastructure-free networks, the nodes themselves perform basic network functions like routing and packet forwarding. Therefore, mobile ad hoc network security is a pressing issue which needs immediate research attention [22, 23, 24, 25].

Rest of the paper is structured as follow: Section 2 provides a brief of security mechanism available for secure routing protocols. Section 3 gives a brief summary of some secure routing protocols. Section 4 presents a comparative study of secure routing protocols described in Section 3. Section 5 gives concluding remarks.

## 2. Security Mechanism for Routing Protocols

Message encryption and digital signatures are two important mechanisms for data integrity and user authentication. There are two types of data encryption mechanisms, symmetric and asymmetric (or public key) mechanisms. Symmetric cryptosystems use the same key (the secret key) for encryption and decryption of a message, and asymmetric cryptosystems use one key (the public key) to encrypt a message and another key (the private key) to decrypt it. Public and private keys are related in such a way that only the public key can be used to encrypt messages and only the corresponding private key can be used for decryption purpose. Even if attacker comprises a public key, it is virtually impossible to deduce the private key. Any code attached to an electronically transmitted message that uniquely identifies the sender is known as digital code. Digital signatures are key component of most authentication schemes. To be effective, digital signatures must be non-forgeable. Hash functions are used in creation and verification of a digital signature. It is an algorithm which creates a digital representation or fingerprint in the form of a hash value (or hash result) of a standard length which is usually much smaller than the message and unique to it. Any change to the message will produce a different hash result even when the same hash function is used. In the case of a secure hash function, also known as a one-way hash function, it is computationally infeasible to derive the original message from knowledge of its hash value. In mobile ad hoc networks, the secrecy of the key does not ensure the integrity of the message. For this purpose, message Authentication Code (MAC) [26] is used. It is a hashed representation of a message and even if MAC is known, it is impractical to compute the message that generated it. A MAC, which is a cryptographic checksum, is computed by the message initiator as a function of the secret key and the message being transmitted and it is appended to the message. The recipient re-computes the MAC in the similar fashion upon receiving the message. If the MAC computed by the receiver matches the MAC received with the message then the recipient is assured that the message was not modified. The next section provides some secure routing protocols based on above security mechanism.

## 3. Secure Routing Protocols

### 3.1 Secure efficient ad hoc distance vector routing

The Secure Efficient Ad hoc Distance vector routing (SEAD) [7] protocol is a secure ad hoc network routing protocol which is based on the design of the Destination-Sequenced Distance-Vector (DSDV) [19] routing protocol. In this protocol for the limited CPU processing capability, and to guard against Denial of- Service attacks in which an attacker attempts to cause other nodes to consume excess network bandwidth or processing time, we can use one-way hash function but we can not use the asymmetric cryptographic operations. The key feature of proposed security protocol is the use one- way hash chains, using an one way hash function H. Each node computes a list of hash values $h_o,h_1,\ldots\ldots h_n$, where $h_i = H(h_i\text{-}1)$ and 0<i<=n, based on an initial random value ho. The paper assumes the existence of a mechanism for distributing hn to all intended receivers. If a node knows H and trusted value $h_n$, then it can authenticate any other value hi, o<i<=n by successively applying the hash function H and then computing the result with $h_n$. This protocol provides a robust protocol against attackers trying to create in correct routing state in other node by modifying the sequence number or the routing metric. SEAD does not provide a way to prevent an attacker to use the same metric and sequence number learned from some recent update message, for sending a new routing update to a different destination.

### 3.2 A Secure on demand routing protocol

A Secure OnDemand Routing Protocol for Ad Hoc Networks (ARIADNE) [8] provides security against arbitrary active attackers and relies only on efficient symmetric cryptography This paper present the design and performance evaluation of a new secure on-demand ad hoc network routing protocol, called Ariadne. Ariadne is more general, more efficient or more secure. Ariadne does not require a trusted hardware and does not require powerful processors. This protocol prevents attackers or compromised nodes from tampering with uncompromised routes consisting of uncompromised nodes, and also prevents a large number of types of Denial-of-Service attacks. In case of using only highly efficient symmetric cryptographic primitives Ariadne is efficient. This protocol can authenticate routing messages using one of three schemes: shared secrets between each pair of nodes, shared secrets between communicating nodes combined with broadcast authentication, or digital signatures. The performance of ad hoc network routing protocol has been evaluating by ns-2 simulator.

### 3.3 ENDAIRA

endairA[1] is designed by the inspiration of Ariadne with digital signature. The name endairA is just reverse of Ariadne. The protocol endiarA is based on the various possible attacks on the Ariadne[8]. The protocol focus on the route discovery process of on-demand source routing protocols. The result is based on the simulation paradigm and actual implementation is still pending.

### 3.4 Cooperation of nodes fairness in dynamic ad-hoc networks

Cooperation of nodes fairness in dynamic ad-hoc networks CONFIDANT [6] protocol is designed as an extension to reactive source-routing protocol such as DSR. It is a collection of components which interact with each other for monitoring, reporting, and establishing routes by avoiding misbehaving nodes. CONFIDANT components in each node include a network monitor, reputation system, trust manager, and a path manager. When DSR is fortified with the CONFIDANT protocol extensions, it is very scalable in terms of the total number of nodes in the network and it performs well even if more than 60% of the nodes are misbehaving. The overhead for incorporating different security components is manageable for ad hoc environment. However, detection based reputation system has few limitations and routes are still vulnerable to spoofing and Sybil attacks.

### 3.5 Security-Aware Routing

A Security-Aware Routing (SAR) [14] Protocol is an on demand routing protocol based on AODV. This protocol integrates the trust level of a node and the security attributes of a route to provide the integrated security metric for the requested route. A Quality of Protection (QoP) vector used is a combination of security level and available cryptographic techniques. SAR uses the timestamps and sequence numbers to stop the replay attacks. Interception and subversion threats can be prevented by trust level key authentication. Attacks like modification and fabrication can be stopped by verifying the digital signatures of the transmitted packet. The main drawbacks of using SAR are the excessive encrypting and decrypting required at each hop during the path discovery. By using SAR route discovered may not be the shortest route in the terms of hop-count, but it is secure.

### 3.6 Secure routing protocol

The Secure Routing Protocol (SRP) [15] is extension that can be applied to many of the on demand routing protocols. SRP provide protection against attacks that disrupt the route discovery process and identify the correct topological information. The main purpose of SRP is to provide the security association (SA) between a source and destination node without need of cryptographic validation of the communication data by the intermediate node. This protocol assumes that security association can be achieved though a shared key $K_{st}$ between the source s and target t. The source node s initiates the route discovery by sending a route request packet to the destination t. This protocol uses the additional header called the SRP header. The SRP header contains the following information: the query sequence number $Q_{sec}$, query identifier number $Q_{id}$, and a 96 bit MAC field. If SRP header is missing intermediate nodes discard a route request message otherwise they forward the request towards destination after extracting $Q_{id}$ , Source, and destination address.

## 4. Comparative Study

We summarize the various secure routing protocols that have been explained in section 3. We consider several attributes and comment on these attributes with respect the each of the protocol discussed above. Table 1 presents the comparative study on various security parameters and security attacks.

**Table 1.** Comparative study of various protocols

| **Performance parameters** | **SEAD** | **ARIADNE** | **ENDAIRE** | **CONFIDANT** | **SRP** | **SAR** |
|---|---|---|---|---|---|---|
| Base Protocol | DSDV | DSR | DSR | DSR | DSR/ ZRP | AODV |
| Encryption Algorithm | Symmetric | Symmetric | Symmetric | Symmetric | Symmetric | Symmetric/ Asymmetric |
| Synchronization | Yes | Yes | Yes | No | No | No |
| Integrity | **No** | **Yes** | **Yes** | **Yes** | **Yes** | Yes |
| Nonrepudiation | **No** | **No** | **No** | **Yes** | **No** | Yes |
| Authentication | **Yes** | **Yes** | **Yes** | **Yes** | **Yes** | Yes |
| Confidentiality | **No** | **No** | **Yes** | **No** | **No** | Yes |
| DoS Attacks | Yes | Yes | Yes | Yes | Yes | No |

## 5. Conclusion

The paper discussed six different secure routing protocols and gives a comparative study based on characteristics of these protocols. The comparative study made in this paper shown the different types of approaches in respect to various security parameters and security attacks. The comparative study will help the researcher to focus on specialized methods for making routing protocols more secure for mobile ad hoc networks.

## References

[1] Gergely Acs, Levente Buttyan and Istvan Vajda, " Provably Secure On-demand Source Routing in Mobile Ad Hoc Networks", IEEE transactions on Mobile Computing, Vol.5, No.11, November 2006, pp. 1533-1546.

[2] Jaier Gomez, Andrew T. Campbell, "Variable –Range Transmission Power Control in Wireless Ad Hoc Networks", IEEE transactions on Mobile Computing, Vol.6, No.1, January, 2007, Pg. 87-99.

[3] Ting-Yao Jiang, Qing-hua Li, "A Secure Routing Protocol for Mobile Ad-Hoc Network " Proceeding of the third International conference on Machine Learning and Cybernetics, 26-29 Augest 2004, Pg. 2825-2829.

[4] Rendong Bai and Mukesh Singhal, "DOA: DSR over AODV routing for mobile ad hoc network", IEEE transactions on Mobile Computing, Vol 5, No 10, October 2006, Pg. 1403-1416.

[5] Y.-C. Tseng, S.-Y. Ni, Y.-S. Chen, and J.-P. Sheu, "The Broadcast Storm Problem in a Mobile Ad Hoc Network", ACM Wireless Networks, Vol 8, No 2, Mar. 2002, pp. 153-167.

[6] S. Buchegger and J. L. Boudec, "Performance Analysis of the CONFIDANT Protocol Cooperation Of Nodes Fairness In Dynamic Ad-hoc NeTworks", In Proc. Of IEEE/ACM Symposium on Mobile Ad Hoc Net- working and Computing (MobiHOC), Jun. 2002.

[7] Y. –C. Hu, D. B. Johnson and A. Perrig, "SEAD: Secure Efficient Distance Vector Routing for Mobile Wireless Ad Hoc Networks", Fourth IEEE Workshop on Mobile

[8] Y. –C. Hu, D. B. Johnson, and A. Perrig, "Ariadne: A Secure On-Demand Routing Protocol for Ad Hoc Networks", Mobicom'02, 2002.

[9] A. Perrig, R. Canetti, D. Tygar, and D. Song, "The TESLA Broadcast Authentication Protocol, RSA Cryptobytes (RSA Laboratories)", Vol 5, No 2, Summer/Fall 2002, pp. 2-13.

[10] C. S. R. Murthy and B. S. Manoj, "Ad Hoc Wireless Networks: Architectures and Protocols", Prentice Hall PTR, 2004.

[11] IEEE Std. 802.11, "Wireless LAN Medium Access Control (MAC) and Physical layer (PHY) Specifications," 1997.

[12] Y. Zhang, Wenjing Lou and Yuguang Fang, " Securing Mobile Ad Hoc Networks with Certificate less Public Keys", IEEE transactions on Dependable and Secure Computing, Vol.3, No. 4, Octuber-December 2006, pp. 386-399.

[13] Williams, B. and Camp, T.: "Comparison of Broadcasting Techniques for Mobile Ad Hoc Networks", In: Proceedings of the 3rd ACM International Symposium on Mobile Ad Hoc Networking & Computing (MOBIHOC '02). pp. 194–205. Lausanne, Switzerland. June 9-11 2002.

[14] R. Kravets, S. Yi, and P. Naldurg, "A Security-Aware Routing Protocol for Wireless Ad Hoc Networks", In ACM Symp. on Mobile Ad Hoc Networking and Computing, 2001.

[15] P. Papadimitratos and Z. J. Haas,"Secure Routing for Mobile Ad hoc Networks", In Proc. of the SCS Communication Networks and Distributed Systems Modeling and Simulation Conference (CNDS 2002), Jan. 2002.

[16] J. Newsome, E. Shi, D. Song, and A. Perrig, "The Sybil Attack in Sensor Networks: Analysis & Defenses", Proc. of the 3rd Intl. Symp. on Information Processing in Sensor Networks, 2004.

[17] Charles E. Perkins and Elizabeth M. Royer. "Ad Hoc OnDemand Distance Vector (AODV) algorithm", In Proceedings of the 2nd IEEE Workshop on Mobile Computing Systems and Applications (WMCSA'99), New Orleans, Louisiana, USA, February 1999.

[18] Y. -C. Hu, D. B. Johnson, and A. Perrig, "Rushing Attacks and Defense in Wireless Ad Hoc Network Routing Protocols", WiSe 2003, 2003.

[19] C Perkins and P. Bhagwat, “Highly dynamic destination sequenced distance vector routing (DSDV) for mobile computers”, ACM SIGCOMM, (October 1994).

[20] Hubaux. J., Buttyan, L., and Capkun, S. , “The Quest for Security in Mobile Ad Hoc Networks,” MobiHw 2001.

[21] Stajjano, F. and Anderson, R., “The Resurrecting Duckling: Security Issues for Ad-hoc Wireless Networks,” Proceedings of Security Protocols Workshop, 1999.

[22] Vinayakraj-Jani, P., “Security within Ad hoc Networks,” PAMPAS Workshop, London, Sept. 16/17 2002.

[23] Wrona, K., “Distributed Security: Ad Hoc Networks & Beyond,” PAMPAS Workshop, London, Sept. 16/17 2002.

[24] Buttyan, L., and Hubaux, J.,”Repn on a Working Session on Security,” Wiretess Ad Hoc Networks Mobile Computing and Communications Review, Vol. 6, Number 4,2002.

[25] Michiardi, P., Molva, R., “Simulation-based Analysis of Security Exposures in Mobile Ad Hoc Networks,” European Wireless Conference, 2002.

[26] Zapata, M., “Secure Ad hoc On-Demand Distance Vector Routing.,” ACM Mobile Computing and Communications Review (MCZR), Vol. 6. No. 3, July 2002, pp. 106-107. pp. 1516-1521

# FLC and NN Based Alpha Compensation OF Three Phase Controlled Rectifier FED DC-Motor Drive

P.T.Krishna Sai[1], K.Murali[2] , Dr.G.R.K.Murthy[3]

[1]Vijaya Institute of Technology for Women, Dept of EEE, JNTU Kakinada
[2]Vijaya Institute of Technology for Women, Dept of ECE, JNTU Kakinada
krishh_sai@yahoo.co.in, kalipindimurali@gmail.com, gvenu@yahoo.co.in

***Abstract* -** When a new control strategy of a drive system is formulated, it is often convenient to study the system performance by simulation before building the prototype. The simulation not only validates the system operation, but also permits optimization of the system performance by iteration of its parameters.

Besides control and circuit parameters, the plant parameter variation effect can be studied. Valuable time is thus saved in the development and design of the product and the failure of components of poorly designed systems can be avoided. The simulation program also helps to generate real time controller software codes for downloading to a microprocessor or digital signal processor [2].

Many circuit simulators like PSPICE, EMTP, and MATLAB/SIMULINK incorporated these features. The advantages of simulink over other circuit simulators are the ease in modeling the transients of electrical machines and drives and to include controls in the simulation. To achieve out objectives this efficient simulink software is used. For electrical drives good dynamic performance is mandatory so as to respond to the changes in command speed and torques. So various speed control techniques are being used for real time applications. Here the speed of a dc motor is controlled by using various controllers like PI-controller, Fuzzy controllers. Fuzzy logic and neural network concepts are applied to DC drive system [1].

This project describes application of fuzzy logic controllers for current and speed control loops of DC drive systems. Neural network is employed to linearize the rectifier characteristics in discontinuous conduction mode. Simulation result shows the superiority of proposed controller over fixed parameter PI-controller and best possible fuzzy logic controller can be designed without expert knowledge and extensive tuning of parameters [1].

**Keywords:** cosine wave crossing method, neural compensation, fuzzy controller, conventional PI- controller, armature reaction, GUI (graphical user interface), feed forward back propagation

## 1. Introduction

Motion control is required in large number of industrial and domestic application like transportation systems, rolling mills, paper machines, textile mills, machine tools, fans, pumps, robot, washing machines etc. Systems employed for motion control are called drives and may employ any of the prime movers such as, diesel (or) petrol engines, gas or steam turbines, steam engines, hydraulic motor and electric motors for supplying mechanical energy for motion control. Drives employing electric motors are known as electrical drives. The introduction of variable speed drives increases the automation and productivity and in the process, efficiency. Nearly 65 % of total electric energy input increasing the efficiency of the mechanical transmission and process can reduce the energy consumption. The system efficiency can be increased from 15% to 27% by introduction of variable speed drive operation in place of constant speed operation [10].

Attempts to improve its characteristics when used with power electronic devices are most challenging and will lead to beneficial results. With this in view, a three phase fully controlled converter fed dc drive with fuzzy compensation scheme [1] and neural compensation scheme are analyzed in this work. The results obtained are compared to establish their relative merits with different controllers such as PI and fuzzy based systems to establish the most preferable controller among them. Neural network is used for Δα compensation scheme. By varying various FLC parameters optimally best possible controllers are designed. Simulation results of proposed controller are compared with PI controller results [2].

## 2. System Model

A separately excited DC motor fed from a fully controlled rectifier is taken as a model system. The block diagram of the considered system is shown below in figure. The speed loop has an inner current control loop for fast dynamic response. The FLC for speed loop sets the current reference for the current loop considering the error and change in error in speed. The current loop output $V_s$' is added with the counter emf $V_c$ to get the control signal $V_s$. Then the control voltage is converted into firing angle α by cosine wave crossing technique. The feed forward addition of counter emf gives fasted loop response [1].

An additional error in firing angle called del-alpha (Δα) is added with the firing angle set by the FLC's. The purpose of Δα and method of finding it are done with fuzzy compensation scheme and neural compensation scheme. The field of the motor is considered to be constant and for simplicity, the three-phase fully controlled rectifier is assumed to work in motoring mode. Here we employ Neural Network as Δα compensator in order to show the superiority to establish non-linear mapping between variables as compared to fuzzy logic system [1].

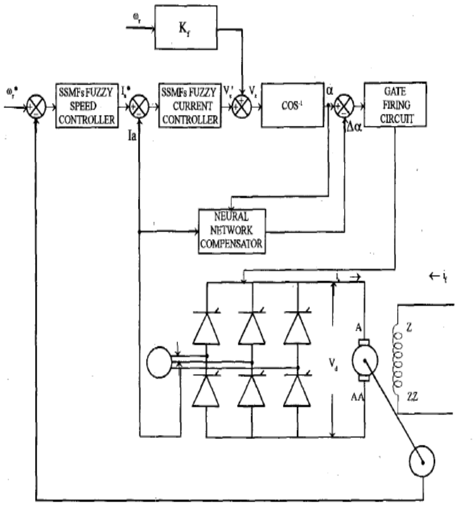


**Fig 2.1: SYSTEM BLOCK DIAGRAM**

**2.2 Mathematical Modeling of a DC Motor:**

The resistance of the field winding and its inductance of the motor used in this study are represented $R_f$ and $L_f$ respectively. The resistance of armature and its inductance are shown by effects are ignored in the description model. Armature reaction effects are ignored in the description of the motor. This negligence is justifiable to minimize the effects of armature reaction since the motor used has either interpoles or compensating winding. The fixed voltage $V_f$ is applied to the field and the field current settles down to a constant value. A linear model of a simple DC motor consists of a mechanical equation and electrical equation as determined in the following equation.

$$J\ \frac{d\omega_m}{dt} = K_m \phi I_a - B_1.\omega_m - T_1 \qquad \ldots\ldots (2.1)$$

$$L_a\ \frac{dI_a}{dt} = V - R_a.I_a - K_b.\phi.\omega_m\ \ldots. \qquad (2.2)$$

Taking Laplace transforms of equations (2.1) ,(2.2) and neglecting initial conditions, we get

$$\mathrm{I}_a(s) = {V(s) - \mathrm{K}_b W_m(s)} \Big/ {(R_a + sL_a)} \quad \ldots. (2.3)$$

$$W_m(s) = \mathrm{K}_b \mathrm{I}_a(s) - {\mathrm{T}_1(s)} \Big/ {(\mathrm{B}_1 + sJ)} \quad .(2.4)\ \ [\text{Ref } 10]$$

Fig 2.2 shows the block diagram representation of the DC motor with Laplace transformed equations

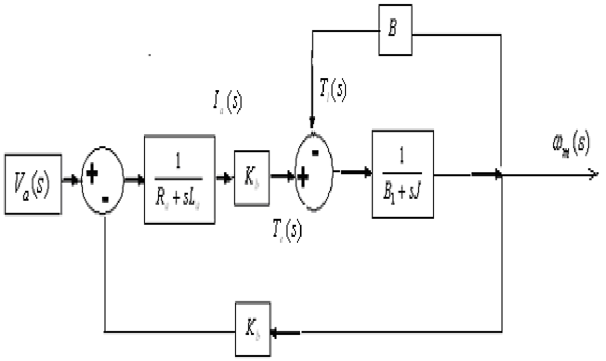


**Fig 2.2: Block diagram representation of the DC motor**

The dynamic model of the system is formed using theses differential equations and MATLAB simulink block as shown in Fig 2.3

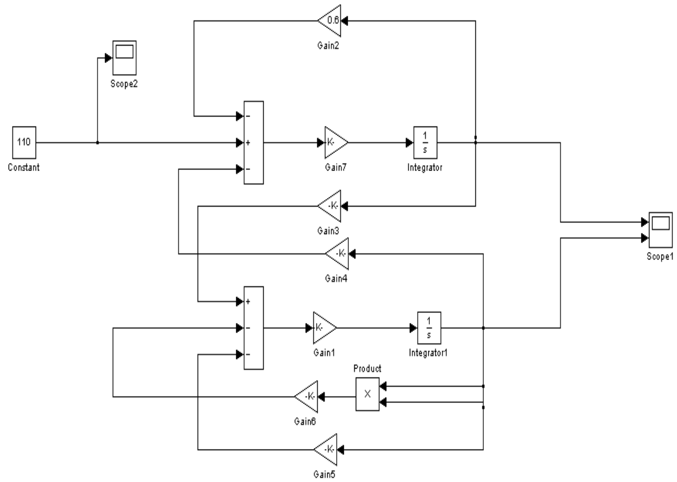


**Fig. 2.3: Simulink Model of dc motor**

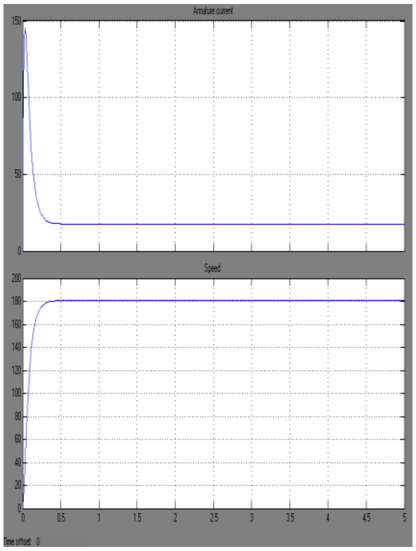


**Fig 2.4: Open loop armature current and speed responses of dc motor**

### 2.3 DC Machine Drive with Phase Controlled Converter

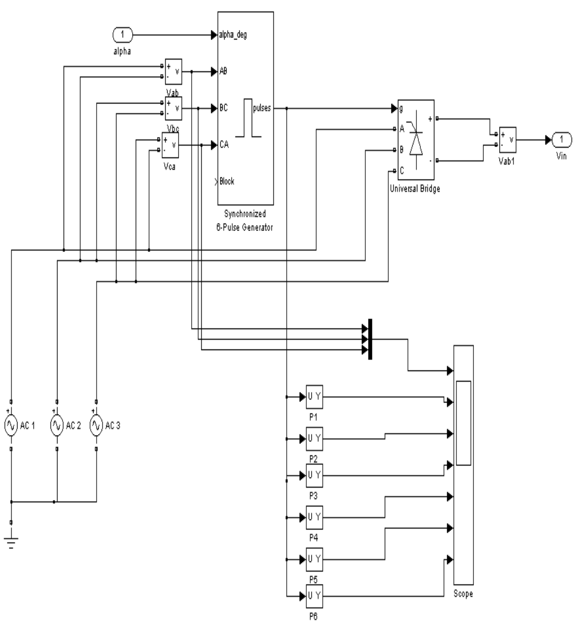
Controlled rectifiers are used to get variable dc voltage from an ac source of fixed voltage. Controlled rectifiers fed dc drives are also known as Static Ward-Leonard system. Controlled rectifiers (Fully-controlled rectifiers) are capable of providing Voltage in two direction and current in one direction which allow the motor control in two quadrant i.e. quadrant I &quadrant IV. For speed control of dc motor there are two types of phase controlled converters are present, they are

I. Three phase full controlled converter

II.Three phase semi controlled converter

### 2.4. Simulink Implementation of Three –Phase Converter

The block diagram of three phase-controlled bridge converters that drives a separately excited dc motor. For simplicity, the converter is used in motoring mode only with fixed field excitation. The firing angle α, required for the rectifier is generated by cosine wave crossing method [5]



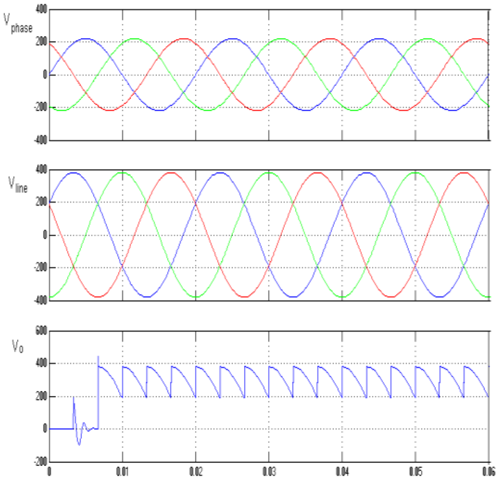


**Fig 2.5: Three phase input and output voltage waveforms**

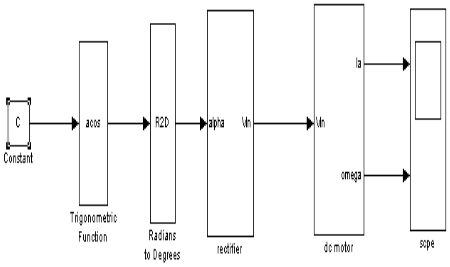


**Fig 2.6: Simulink Block Diagram of phase controlled converter fed dc drive**

### 2.4.1 Converter Equations

The converter may operate in either continuous or discontinuous conduction mode. At low speed when the counter e.m.f is small, the conduction will be continuous. However, at high speed, the conduction will tend to be discontinuous. In continuous and discontinuous conduction mode, the normalized armature circuit equations can be given as follows:

The expressions for Rectifier Current and Voltage in Continuous Conduction Mode when supplied to dc motor are

$$I_a(pu)=\frac{I_a}{3V_m/\pi X}=\frac{X}{R}\left[\cos\alpha-\frac{\pi V_c}{3V_m}\right] \qquad \ldots\ldots\ldots (2.5)$$

$$V_d(pu)=\frac{V_d}{V_m}=\frac{3}{\pi}\cos\alpha$$

Where

$I_a$ = armature current (average)

$V_m$ = peak ac line voltage

X = armature reactance (ωL)

R = armature resistance

A = converter firing angle

$V_c$, = armature counter e.m.f

And $V_d$ = converter output voltage (average)

The expressions for Rectifier Current and Voltage in Discontinuous Conduction Mode when supplied to dc motor are

$$I_a(pu)=\frac{I_a}{3V_m/\pi X}=\frac{X}{R}\left[\cos\left(\pi/3+\alpha\right)-\cos\left(\pi/3+\alpha-\theta_1\right)-\frac{V_c}{V_m}\theta_1\right]$$

$$V_d(pu)=\frac{V_d}{V_m}=\frac{3}{\pi}\left[\cos\left(\pi/3+\alpha\right)-\cos\left(\pi/3+\alpha-\theta_1\right)-\frac{V_c}{V_m}\theta_1\right]+\frac{V_c}{V_m}$$

when

$$\frac{V_0}{V_m}=\frac{\sqrt{1+\left(X/R\right)^2}}{1-\exp\left(-R\theta_1/X\right)}\left[\sin\left(\pi/3+\alpha+\theta_1-\varnothing\right)-\sin\left(\pi/3+\alpha-\varnothing_1\right)\exp\left(-R\theta_1/X\right)\right]$$

……….(2.6)

Where

$\theta_1$ = conduction angle of current pulse ( $0<\theta<\pi/3$ ) and

$\tan\varnothing=X/R$ .

For a fixed **X/R** parameter, the equations above are plotted in Fig below for different α angles, which also indicate the boundary between continuous and discontinuous conduction modes[2].

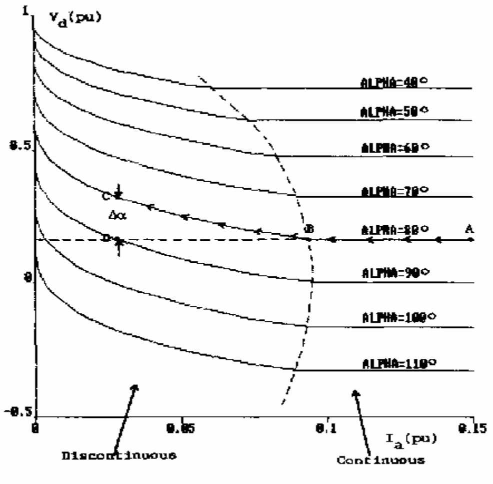


**Fig 2.7: $V_d$-$I_d$ Transfer Characteristics of phase controlled converter**

For example at α = 80, the conduction

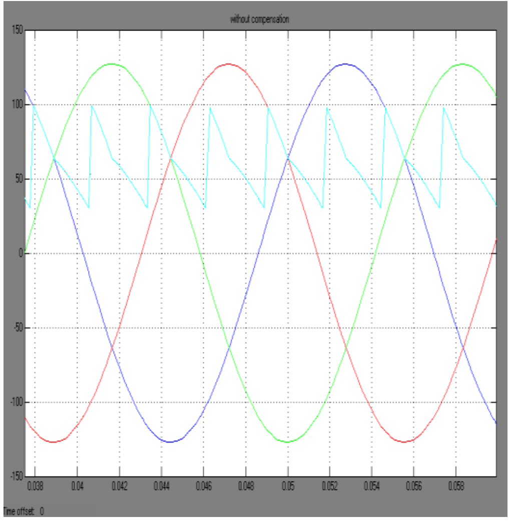


**Fig 2.8: Response of the converter for firing angle of α = 80 degrees**

## 3. Fuzzy speed and current controller

The fuzzy linearization of converter at discontinuous conduction mode the fuzzy control is well suited in a nonlinear system especially where parameter variation problem exist. In addition to converter linearization, fuzzy logic control was applied to the speed and current loops. The objective was to explore the control robustness in the presence of parameter variation and load disturbance effect. However, both loops must satisfy the needs of fast transient response with minimum overshoot with converter linearization, both speed and current loops have essentially fist order characteristics. Therefore, intuitively the same fuzzy control strategy should be valid for both loops .The fuzzy speed and current controllers are equally effective in ac drives with vector control, since the transient response is similar to that of a dc machine [4] .

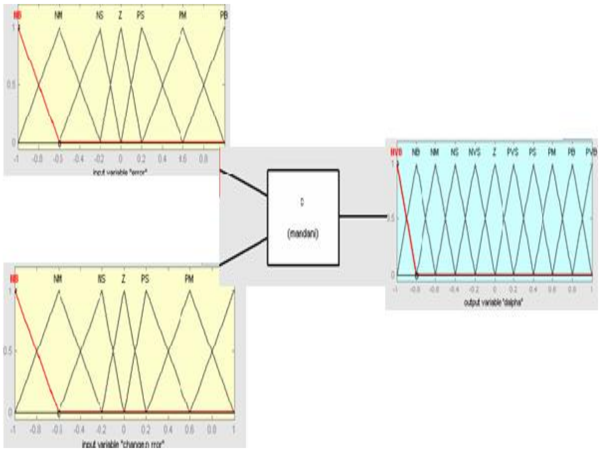


**Fig 3.1: Membership functions for error, change in error and dalpha**

**Table1**

| Ia \ α | NB | NS | Z | PS | PB |
|---|---|---|---|---|---|
| NVB | NVB | PB | PB | PB | PB |
| NB | NVB | Z | Z | Z | Z |
| NM | NVB | NS | NVS | NVS | NVS |
| NS | NVB | NM | NS | NS | NS |
| Z | NVB | NB | NM | NM | NS |
| PS | NVB | NVB | NB | NM | NM |
| PM | NVB | NVB | NB | NB | NB |
| PB | NVB | NVB | NVB | NB | NB |
| PVB | NVB | NVB | NVB | NVB | NB |

Rule base matrix [2]

## 3.1 Fuzzy compensation

The converter which drives the separately excited dc motor may operate in either continuous (or) discontinuous mode. At low speed when the counter emf is small, the conduction will be continuous. However, at high speed, the conduction will tend to be discontinuous. Here the line voltage ($V_m$) can essentially be considered at constant, and therefore, $V_d$ can be controlled linearly by V with cosine wave crossing technique [3]

Fuzzy linearization of converter at discontinuous conduction mode the fuzzy compensation was implemented in mamdani type and it is a two input and single output fuzzy system The input of the fuzzy system are Ia and α and the output of the network is Δα.. It involves three steps (i) Fuzzification (ii) Inference engine (iii) Defuzzification

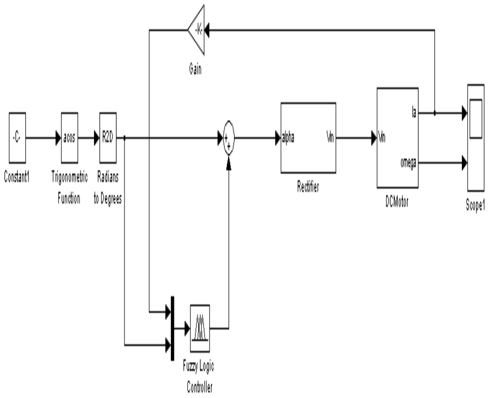


**Fig 3.2: Block diagram of Fuzzy Compensation**

## 3.2 SIMULINK IMPLEMENTATION OF FUZZY COMPENSATION

### 3.2.1 Neural compensation

The required data to train the neural network are obtained by plotting the relationship between rectifier output voltage and current given by the expressions (equations 3.1 and 3.2). The input of the neural network are $I_a$ and α and the output of the network is Δα[1 ]. A two layer network is chosen to do this operation. The Neural Network is trained using back-propagation by neural network training tool GUI. A GUI Network/Data Manager window has its own work area, separate from the more familiar command-line workspace. Thus, when running a GUI Network, you can create a network, view it, train it, simulate it, and export the final results to the workspace. Similarly you can import data from the work space for use in the GUI.Generally here the problem is the compensation of alpha of phase controlled rectifier fed dc drive. Here in this compensation of alpha the GUI tool used is nn tool.

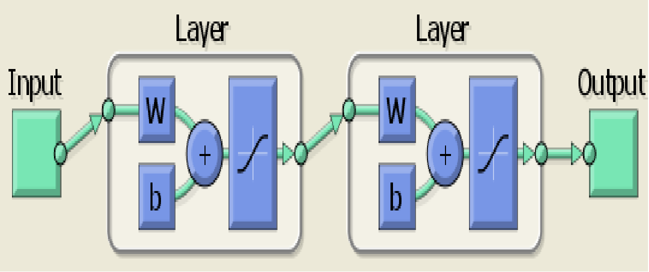


**Fig 3.3: Neural block diagram**

### 3.2.2 Simulation study

Here we have to create a network and perform compensation of del-alpha. Now give the input vector containing the different values of $I_a$ , alpha. Once the network is created, train it and can save the network, its output, etc., by exporting it to workspace. To start type nn tool in the command window. Now import data from the workspace and select the given data for input and target . Now create a network and call it delalpha. And set the network type feed forward back propagation and press creates. Network is created and delalpha is added to network manager. Now select the network delalpha and train the network. After the training finished generate a simulink block by giving the network descriptions place that block used for the compensation.

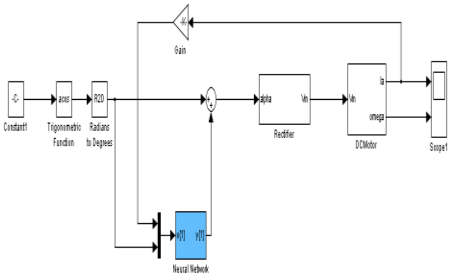


**Fig 3.4: Block diagram for neural compensation**

## 4 SIMULINK IMPLEMENTATION OF NEURAL COMPENSATION

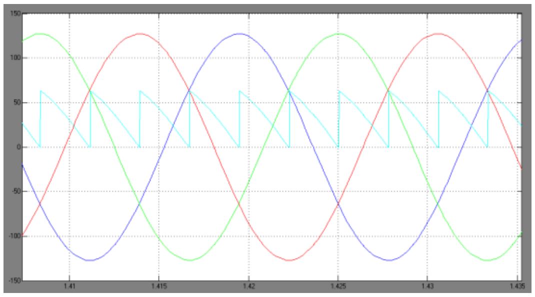
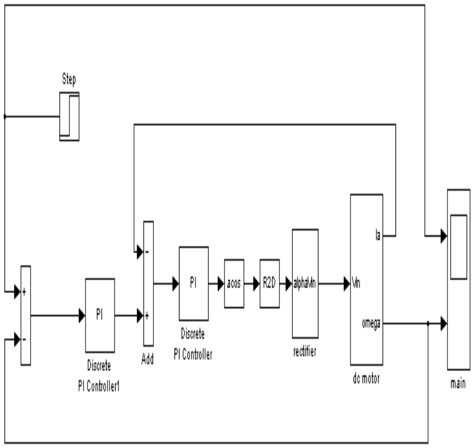


Fig 4 .1: Linearization of rectifier characteristics

### 4.1 SIMULINK IMPLEMENTATION OF CLOSED LOOP SPEED CONTROL

The PI- controller was designed for controlling the speed of the dc motor. The three phase controlled converter fed with PI controller was implemented in simulink and output waveforms was observed for a command speed of step change in speed [12] .

### 4.2 Speed control of separately excited dc motor drive with Proportional and Integral (PI)control



**Output waveform**

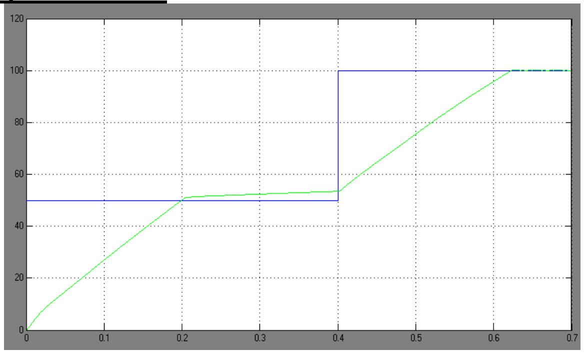
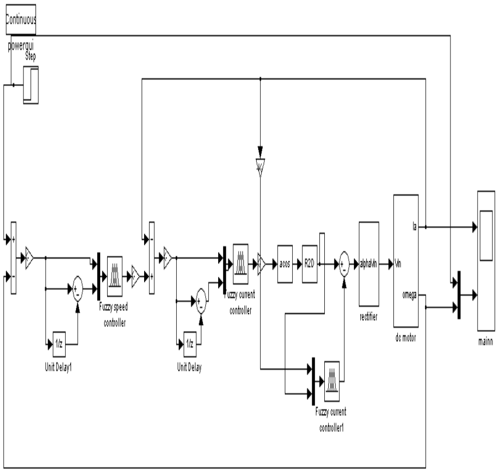


**Fig 4.2 output characteristics of neural compensation of closed loop speed control**

## 4.3 Simulink model of fuzzy logic speed control of dc motor with fuzzy compensation



**Output waveform**

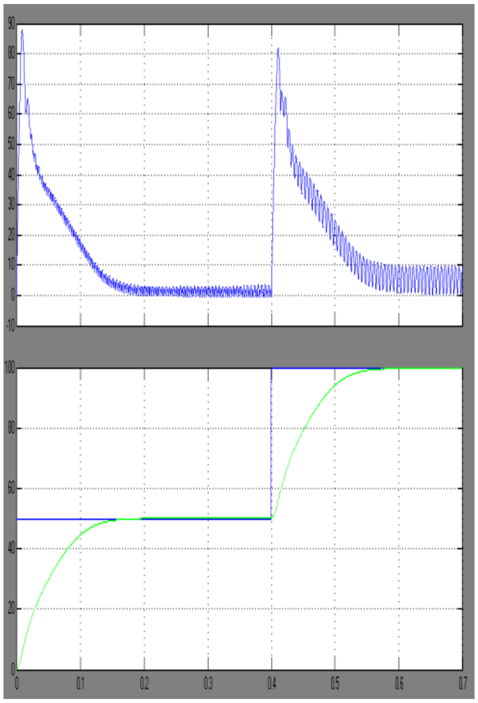
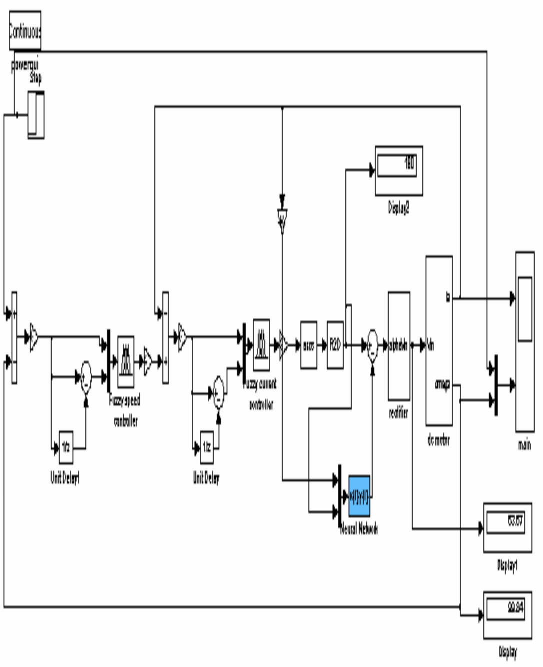


**Fig 4.3 output characteristics of fuzzy logic speed control of dc motor with fuzzy compensation**

## Simulink model of fuzzy logic speed control of dc motor with neural compensation



**Output waveform**

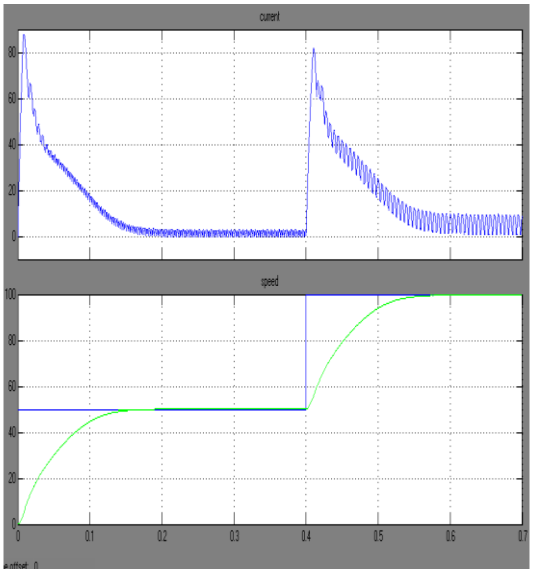


**Fig 4.4 output characteristics of fuzzy logic speed control of dc motor with neural compensation**

## Table-2 : Drive Parameters

Supply voltage (V) -110 volts
Rated motor current ($I_a$) - 20A
Speed of the motor (r.p.s) – 1800
Armature resistance (Ra) – 0.6 ohms
Armature Inductance (La) – 8mH
Moment of Inertia (J) – 0.0465 Kg-m^2
Friction Coefficient (B) – 0.004N.m sec/rad
Line voltage (VL) – 90 volts
Shaft power – 2.5hp
Load Torque (TL) – 2.78*10^(-4) *W^2

## 5 Results:

In this project the neural and fuzzy compensation has been implemented to linearize the transfer characteristics of dc motor at discontinuous mode which occurs at light load(or) high speed. The fuzzy control is then extended to the current and speed control loops, replacing the conventional PI- controller method. The simulation study indicates the superiority of fuzzy control over conventional control methods. The controllers designed have been simulated for command speed of step change form 50 rad/sec top 100 rad/sec. Then percentage overshoot (%Mp) and steady state error ($e_{ss}$) and rise time have been measured

**Table-3**

| **Parameter** | **PI Controller** | **Fuzzy Logic Controller and Compensation** | **FLC and Neural Network Compensation** |
|---|---|---|---|
| Setting Time Ts | 0.2sec | 0.17sec | 0.15sec |
| Over shoot(%) | 6% | 0 rad/sec | 0rad/sec |
| Rise time tr | 0.17 sec | 0.1 sec | 0.08sec |
| Steady State Error | 0 rad/sec | 0 rad/sec | 0 rad/sec |

## 6 Conclusions:

- Neural network compensation has been implemented and compared with fuzzy logic compensation implementation.
- Separately excited dc motor speed has been controlled with classical PI and FLC.
- Simulation results shows that neural network compensation with FLC gives better performance in respect of rise time, percentage overshoot and steady state error in comparison of PI and Fuzzy logic compensation.
- It would be possible to get better PI controller performance by increasing proportional and integral coefficient but it will be a problem in real time applications of PI controller.

## 7 Future Scope

- Till now the speed of dc motor is controlled using PI, Fuzzy controllers.
- The next approach is to control the speed using a neural network controller and neuro fuzzy control.
- Another new approach is using a DSP controller. With the DSP controller an intelligent control approach is possible to reduce the overall system cost and to improve the reliability of the drive performance.

## 8 References:

1. S.Baskar, P.Subbaraj, N.M.Prakasah Kumar'' SSMFS Fuzzy logic control and neural network based alpha compensation of phase controlled rectifier fed DC drives'', IEEE Trans.of IND.Appl.s p p.403-408, 1998

2. **Gilbert** C.D. **Sousa** and B.K. Bose, **"A Fuzzy Set Theory** Based Control **of** a Phase-Controlled DC Machine Drive", **lEEE** Trans,Ind. Appl, vol. 30, Jan **1994.**

**3.** Mathworks, Neural networks tool box users guide (version 7.6) Jan 1998

**4.** C.L. Chen and C.T. Hsieh, "User friendlydesign of Fuzzy **Logic** Controller", **IEE.** Control **Theory** and applications.

**5. J.S.** Mapes and B.K. Bose, "Linearization of the transfer characteristics of a phase- controlled converter under discontinuous conduction", **IEEE Trans,** Ind. Appl, vol. **1A-**i4, pp, **559-564, 1978.**

**6.** J.M. Zurada, "Introduction to Artificial Neural **Systems",** Jaico Publishing House, First Edition, **1994.**

**7.** Drainkov.D, Hel1endoorw.H & Reinfiank.M -**"An** Introduction to **Fuzzy** Control", Narosa.publishing house, I Reprint, **1996.**

**8.** Muhammad H. Rashid, **"Power** Electronics Circuits, Devices and Applications", Prentice Hall of India, **I1** edition, June **1997.**

9.Gopal K.Dubey,"Fundamentals of Electric Drives"

10.P.SBimbra "Generalized Theory of Electrical Machines"

11. Digital control systems 2$^{nd}$ edition, KUO

# Author Biographies

**P.T.Krishna Sai** received his B.Tech degree from the Nimra College of Engineering and Technology affiliated to JNTU Hyderabad, India, in 2008 and his M.Tech degree from JNTU Kakinada, India in 2010. Currently he is working as an Assistant. Professor in the Dept of Electrical and Electronics Engineering at Vijaya Institute of Technology for Women, Enikepadu, Vijayawada, affiliated by JNTU Kakinada .A.P, India His research interest includes Neural Network and Fuzzy Logic applications to power electronics and electrical drives.
Email: krishh_sai@yahoo.co.in

**K.Murali** obtained his M.Sc in electronics from P.B.Siddhartha Post Graduate center affiliated to Acharaya Nagarjuna University Guntur, India, in 2005 and his M.Tech degree in the specialization of Communications and Signal Processing Engineering from V.R.Siddhartha Engineering College affiliated to Acharaya Nagarjuna University, Guntur, India, in 2009.He has 5 years of teaching experience. He has published 4 papers in international and national journals. Currently he is engaged in research on wireless communication under the esteemed guidance of Professor Dr.S.Sri Gowri. Currently he is working as an Assistant. Professor in the Dept of Electronics and Communication Engineering at Vijaya Institute of Technology for Womene, Enikepadu, Vijayawada, affiliated by JNTU Kakinada .A.P, India. His research interest includes wireless communications (G3G, 4G, and LTE),Space communications, and fiber optical communications. Also he is a life member of IAENG
Email: kalipindimurali@gmail.com

**Dr. G.R.K. Murthy** obtained his BE in Electrical Engineering from Andhra University, India in 1960 and his MTech and PhD from the Indian Institute of Technology, Kharagpur, India in 1971 and 1979 respectively. He worked for 38 years in different capacities in the Department of Electrical Engineering at JNT University, India. He is currently working at Vignan University as a Professor, Director of Library and Information Systems and Head of the Department of Electrical Engineering. His research interests include energy conservation and electric drives.

Email: gvenu@yahoo.co.in

# Analysis of a Cournot Duopoly Model's Stability

Hong-xing Yao， Jia-xiu Zu

（Faculty of Science，Jiangsu University , Zhenjiang 212013 , China)

***Abstract*:** In this paper, the feedback control methods are applied to a duopoly model based on heterogeneous expectations. This is the time-delayed feedback control of the production system. This control aims to bring this system into instability equilibrium by using delay of state variables. The validity of the control method is proved through theoretical analysis and numerical simulations. Moreover，scope of convergent condition is given. The production model can quickly reach Nash equilibrium after control, providing theoretical reference and production conditions to enterprises.

***Key words*:** dynamical Cournot model, delayed feedback control, the Pareto optimal, Nash equilibrium, game, bounded rationality

## 1 Introduction

Oligopolistic market is a universal market mechanism, in which a trade is completely controlled by several firms. The firms manufacture the same or homogeneous products and they must consider not only the demand of marker, but also the actions of their competitors [1]. Game theory has been widely applied to oligopolistic markets thank to its ability to consider strategic interactions among firms. Oligopolist is competitive, and the basic solution which refers to competitive equilibrium in Cournot game is Nash equilibrium or Cournot equilibrium. The adjust dynamics to get the Nash equilibrium and the stability are studied by many works [2–9]. But just as what Nash equilibrium reveals, Nash equilibrium reflects individual rationality, but it violates collective rationality – Nash equilibrium of the duopoly game is not Pareto optimal. The prisoners' dilemma shows that, there is a contradiction between individual rationality and collective rationality, and the correct choice based on individual rationality will reduce everybody's welfare. In other words, Pareto improvement cannot be carried on and Pareto optimal cannot be realized by personal interest's maximization. The main question which the prisoners' dilemma poses is whether a cooperative behaviour can emerge among rational and self-interested players whenever there is no formal agreement [10]. In real economical markets we truly can observe that competitors are often able to achieve the cooperation.

Although the duopoly game with output competition (Cournot game) is faced with prisoners' dilemma (Nash equilibrium is not Pareto optimal), it cannot be studied in standard game model with prisoners' dilemma. Because the collection of strategies in this model is a finite set, and in the output competition it is an infinite set. Cafagna [10] has built a strategy with output adjustment (the 'good' strategy), and makes the firms reach a cooperative equilibrium finally. The prisoners' dilemma can be explained based on that. However, the 'good' strategy is based on the premise that producers completely know about their competitors' output and profit. In fact, the producers with mutual competition, or even the producers who have achieved certain cooperation, keep the output, the profit and the related things as the business secrets for their own benefit. So the supposition of incomplete information is more rational. For example, in the model with two producers, as long as one producer does not know the other's cost of production, it is impossible for the first producer to know about the other's profit under different combination of bilateral outputs. That is to say, the first producer cannot have complete information. Then under the premise that each producer incompletely knows about the competitor's information (output, profit and so on), is there a strategy of output adjustment for the producers to use to achieved a cooperative equilibrium?

In this paper, we study that how firms get bigger profits by adjusting their own outputs. It is different from the paper [10-12] that the producers do not know about the market information of the competitor's output and profit, and the cooperative behaviour in duopoly competition is considered with the ‘‘tit-for-tat” conduct.

## 2. The model

There are two firms produce a homogeneous good in a market .Taking production decisions at discrete time periods $t = 1,2,3,\ldots$. Denoting the quantity of output by each firm at time $t$ is $q_{i,t}\left(i = 1,2\right)$. We have that cost function has the linear form:

$$c_{i,t} = c_i q_{i,t} \qquad (1)$$

Let $p(Q)$ denote the inverse demand function:

$$p_t = a - b\sqrt{Q} \qquad (2)$$

Where $a,b > 0$, $a > c_i$ and $Q_t = \sum_i q_{i,t}$

Then the profit of player $i$ at time $t$ is given by:

$$\pi_{i,t} = \left(a - b\sqrt{Q} - c_i\right)q_{i,t} \qquad (3)$$

This paper is about cooperation under the incomplete information, and the following models are based on the assumption that the firms compare their own profits with the cooperative profit. The solving of the cooperative profit has been introduced in duopoly game theory. The cooperative profit means the profit which is solved by maximizing the sum of all firms’ profit. We consider the symmetrical case: $c_1 = c_2 = c$ ,then can get the cooperative profit, $\pi_c = \dfrac{2(a - c)^3}{27b^2}$ ,and the cooperative output, $q_c = \dfrac{2(a - c)^2}{9b^2}$

## 3. The tit-for-tat dynamic strategy

The tit-for-tat strategy is the best behaviour allowing the achievement of cooperation in repeated games [10]. Its characteristic is that every player consists in doing what the opponent did in the previous move. In the paper, we study the Cournot model with the tit-for-tat conduct. And the dynamic equations are based on the incomplete information. Each producer cannot obtain the competitor’s complete information, but he completely knows about his own output and profit. The firm $i$ can compare his profit $\pi_{i,t}$ at time $t$ with the cooperative profit $\pi_c$ which is Pareto optimal. If the cooperative profit $\pi_c - \pi_{i,t} < 0$, then his own profit is more; he extrapolates that the competitor is cooperative, then he will properly reduce his output to continue the cooperation as a ‘‘reward”1; Otherwise, if $\pi_c - \pi_{i,t} > 0$, the firm $i$ cannot realize the cooperative profit, and extrapolates that the competitor is not cooperative, then he will increase his output as ‘‘penalty”.2 For this case ,we get the dynamical systems of $q_1$ , and $q_2$ as follows:

$$q_{i,t+1} = q_{i,t} + u_i\left(\pi_c - \pi_{i,t}\right) = q_{i,t} + u_i\left[\pi_c - \left(a - b\sqrt{Q} - c_i\right)\right] \qquad (4)$$

where $u_i\left(i = 1,2\right)$ is a adjusting parameter, and $u_i > 0$ .In this model ,Since the firms do not need know the competitor’s related information, it is an adjusting strategy with incomplete information. Although its form is simple, it is based on the thoughts of ‘‘tit-for-tat” strategy in prisoners’ dilemma game. Now the question is that whether the firms can achieve a cooperative Pareto optimality. With above assumptions, the duopoly game with heterogeneous players is described by a two-dimensional nonlinear map $T\left(q_1(t), q_2(t)\right) \to \left(q_1(t+1), q_2(t+1)\right)$ defined as :

$$T:\begin{cases} q_1(t+1) = q_1(t) + u_1\left[\pi_c - (a - c)q_1(t) + bq_1(t)\sqrt{q_1(t) + q_2(t)}\right] \\ q_2(t+1) = q_2(t) + u_2\left[\pi_c - (a - c)q_2(t) + bq_2(t)\sqrt{q_1(t) + q_2(t)}\right] \end{cases} \qquad (5)$$

Where $q_i(t)$ denotes productions of period $t$ , $q_i(t+1)$ represent productions of period $t+1$

In the paper, we are considering an economic model where only nonnegative equilibrium points are meaningful. So we only study the nonnegative fixed points of the map (5), i.e. the solution of the nonlinear algebraic system as:

$$\begin{cases} \pi_c - (a - c)q_1 + bq_1\sqrt{q_1 + q_2} = 0 \\ \pi_c - (a - c)q_2 + bq_2\sqrt{q_1 + q_2} = 0 \end{cases} \qquad (6)$$

By setting $q_i(t+1) = q_i(t), i = 1,2$ in system (5), we obtained (6).

Then it is easy to work out an unique fixed point of system (6): $E = \left(q_1^*, q_2^*\right)$,where

$$q_1^* = q_2^* = q_c = \frac{2(a - c)^2}{9b^2}$$

The stability of these equilibriums is based on the eigenvalues of the Jacobian matrix of system (5)

$$J\left(q_1^*,q_2^*\right)=\begin{pmatrix} 1+u_1\left(c-a+\dfrac{b}{2\sqrt{q_1+q_2}}\right) & \dfrac{u_1 b}{2\sqrt{q_1+q_2}} \\ \dfrac{u_2 b}{2\sqrt{q_1+q_2}} & 1+u_2\left(c-a+\dfrac{b}{2\sqrt{q_1+q_2}}\right) \end{pmatrix}$$

We compute the Jacobian matrix $J$ at $E$ then get

$$J\left(q_1^*,q_2^*\right)=\begin{pmatrix} 1-\dfrac{u_1(a-c)}{6} & \dfrac{u_1(a-c)}{6} \\ \dfrac{u_2(a-c)}{6} & 1-\dfrac{u_2(a-c)}{6} \end{pmatrix}$$

By calculation, we get the characteristic polynomial $P(\lambda)$ of the matrix $J\left(q_1^*,q_2^*\right)$ as following:

$$p(\lambda)=\lambda^2-Tr\lambda+Det=0$$

Where $Tr$ is the trace and $Det$ is the determinant of the Jacobian matrix $J\left(q_1^*,q_2^*\right)$.

$$Tr=2-\frac{u_1(a-c)}{6}-\frac{u_2(a-c)}{6}$$

$$Det=1-\frac{u_1(a-c)}{6}-\frac{u_2(a-c)}{6}$$

Then we have two eigenvalues of matrix $J\left(q_1^*,q_2^*\right)$, $\lambda_1=1$ and $\lambda_2=1-\dfrac{(u_1+u_2)(a-c)}{6}$. If it holds that $u_i\,(i=1,2)$ is very small, we have $\left|\lambda_2\right|<1$. Since $\lambda_1=1$ is a critical condition we cannot know the stability of the system (5). But the following numerical experiments show that its stability is sensitive to the parameter

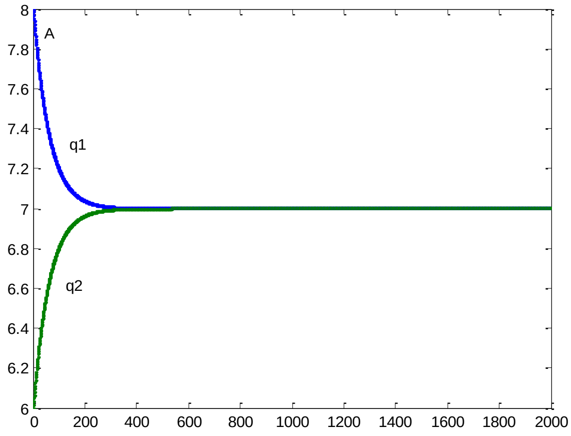


Fig. 1a. The output of the system (4) is stable

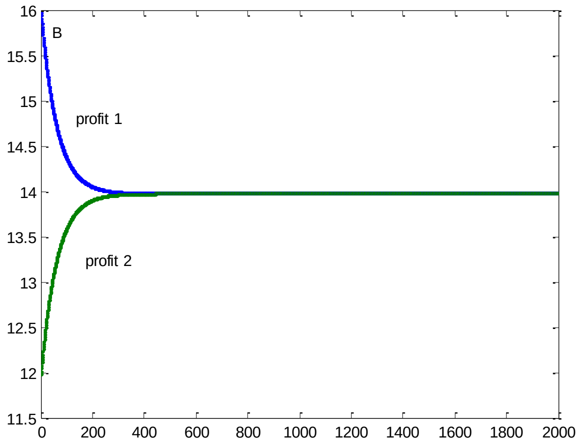


Fig. 1b. The profit of the system (4) is stable

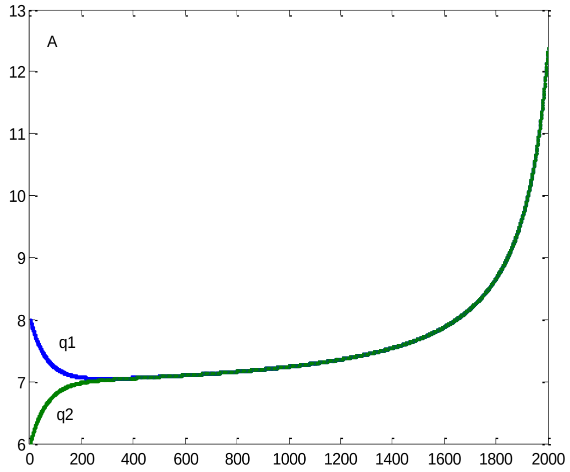


Fig. 2a The output t of the system (4) is unstable

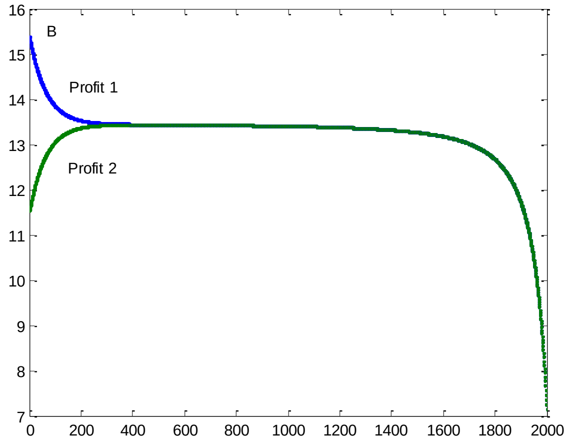


Fig. 2b. The profit of the system (4) is unstable

We take $a=8, c=2, u_1=u_2=0.0082$, and the initial value $q_{1,0}=8, q_{2,0}=6$. If we fix other parameters and vary one, for instance b, the stability of system changes . Fig. 1 shows that it is stable, but if a parameter changes slightly, it is the contrary (Fig. 2). And Fig. 2 shows that the output and the profit not only cannot achieve the Pareto optimality, but also appears the phenomenon of malignant competition – the outputs of both sides increase infinitely (Fig. 2A), while the profits approach to zero (Fig. 2B). That is to say, the firms in Cournot game cannot achieve the Pareto optimal equilibrium under the adjustment Eq. (4)

## 4. Delayed feedback control of the production system

4.1 By adding a time-delayed feedback control, we consider a new strategy:

$$\begin{cases} q_{1,t+1} = q_{1,t} + u_1\left(p_c - (a-c)q_{1,t} + b\sqrt{q_{1,t}+q_{2,t}}\right) + k\left(q_{1,t} - q_{1,t-1}\right) \\ q_{2,t+1} = q_{2,t} + u_2\left(p_c - (a-c)q_{2,t} + b\sqrt{q_{1,t}+q_{2,t}}\right) \end{cases} \quad (7)$$

Where $u_i\,(i=1,2)$ is an adjustment parameter , and $u_i>0$ , $k\left(q_{1,t}-q_{1,t-1}\right)$ is the delayed feedback control of the system. (7) equivalent to the following three-dimensional equations:

$$\begin{cases} q_{1,t+1} = q_{1,t} + u_1\left(p_c - (a-c)q_{1,t} + b\sqrt{q_{1,t}+q_{2,t}}\right) + k\left(q_{1,t} - q_{1,t-1}\right) \\ q_{2,t+1} = q_{2,t} + u_2\left(p_c - (a-c)q_{2,t} + b\sqrt{q_{1,t}+q_{2,t}}\right) \\ q_{3,t+1} = q_{1,t} \end{cases} \quad (8)$$

The Jacobian matrix at $E^* = \left(q_1^*, q_2^*\right)$ takes the form:

$$J(E^*) = \begin{pmatrix} k+1+u_1\left(c-a+\frac{b}{2\sqrt{q_1^*+q_2^*}}\right) & \frac{u_1 b}{2\sqrt{q_1^*+q_2^*}} & -k \\ \frac{u_2 b}{2\sqrt{q_1^*+q_2^*}} & 1+u_2\left(c-a+\frac{b}{2\sqrt{q_1^*+q_2^*}}\right) & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

By calculation, we get the characteristic polynomial $f(\lambda)$ of the matrix $J\left(q_1^*, q_2^*\right)$ as following:

$$f(\lambda) = \lambda^3 + B_1\lambda^2 + B_2\lambda + B_3 = 0$$

Where $B_1 = -k - 2 + \left(u_1+u_2\right)\left(a - c - \frac{3b^2}{4(a-c)}\right)$

$$B_2 = 2k + 1 + \left(u_1 + u_2 + u_2 k + u_1 u_2\left(-a + c + \frac{3b^2}{4(a-c)}\right)\right)$$
$$\ast\left(-a + c + \frac{3b^2}{4(a-c)}\right)\frac{9u_1u_2b^4}{16(a-c)^2}$$
$$B_3 = -k + u_2 k\left(a - c - \frac{3b^2}{4(a-c)}\right)$$

From Jury conditions, the necessary and sufficient conditions for $|\lambda_i| < 1, i = 1,2,3$ are:

$$\begin{cases} 1 + B_1 + B_2 + B_3 > 0 \\ 1 - B_1 + B_2 - B_3 > 0 \\ 1 - B_3^2 > \left|B_2 - B_1B_3\right| \\ \left|B_3\right| < 1 \end{cases} \quad (9)$$

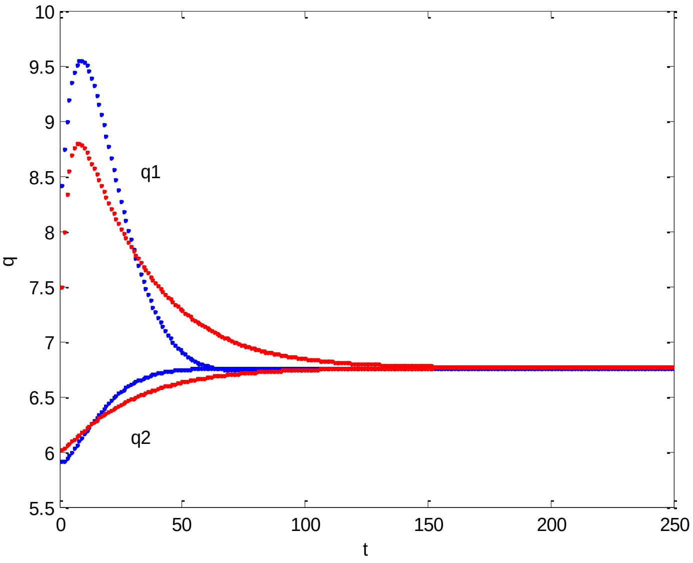


Fig3a

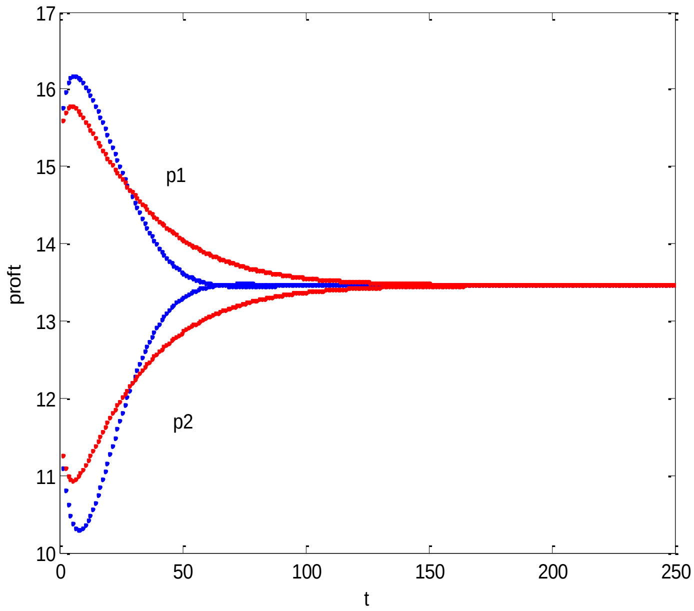


Fig3b

So the equilibrium point $E^*$ of the system (7) is stable, if the conditions in (9) are all satisfied.

We reconsider the unstable situation ( $a=8, b=1.09, c=2, u_1=u_2=0.0082, q_{1,0}=8, q_{2,0}=6$ ) in Section 3. Let $k=0.4$, now the output and profit system (7) become stable, as showed in Fig. 3(blue line) . In Fig.3a,the blue point shows that the changes of Productions1,2,when adds a time-delayed feedback control strategy. In Fig.3b,the blue point shows the proft.

## 4.2 By adding two time-delayed feedback control, we consider a new strategy:

$$\begin{cases} q_{1,t+1}=q_{1,t}+u_1(p_c-(a-c)q_{1,t}+bq_{1,t}\sqrt{q_{1,t}+q_{2,t}})+k_1(q_{1,t}-q_{1,t-1}) \\ q_{2,t+1}=q_{2,t}+u_2(p_c-(a-c)q_{2,t}+bq_{2,t}\sqrt{q_{1,t}+q_{2,t}})+k_2(q_{2,t}-q_{2,t-1}) \end{cases}$$

(10)

Where $u_i\,(i=1,2)$ is an adjustment parameter , and $u_i>0, k_i\left(q_{1,t}-q_{1,t-1}\right)$ is the delayed feedback control of the system. (10) equivalent to the following four-dimensional equations:

$$\begin{cases} q_{1,t+1}=q_{1,t}+u_1(p_c-(a-c)q_{1,t}+bq_{1,t}\sqrt{q_{1,t}+q_{2,t}})+k_1(q_{1,t}-q_{3,t}) \\ q_{2,t+1}=q_{2,t}+u_2(p_c-(a-c)q_{2,t}+bq_{2,t}\sqrt{q_{1,t}+q_{2,t}})+k_2(q_{2,t}-q_{4,t}) \\ q_{3,t+1}=q_{1,t} \\ q_{4,t+1}=q_{2,t} \end{cases}$$

(11)

The Jacobian matrix at $E^*=\left(q_1^*,q_2^*\right)$ takes the form :

$$J(E)=\begin{pmatrix} M_1 & \dfrac{bq_1^*}{2\sqrt{q_1^*+q_2^*}} & -k_1 & 0 \\ \dfrac{bq_2^*}{2\sqrt{q_1^*+q_2^*}} & M_2 & 0 & -k_2 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

By calculation, we get the characteristic polynomial $f(\lambda)$ of the matrix $J\left(q_1^*,q_2^*\right)$ as following:

$$f(\lambda)=\lambda^4+(M_1+M_2)\lambda^3+(M_1M_2+k_2)\lambda^2 \\ +\left[k_1-k_2M_1-\frac{b^2q_1^*q_2^*}{4(q_1^*+q_2^*)}\right]\lambda+(k_1k_2-k_1M_2)$$

Where

$$M_1=k_1+1-u_1(a-c)+u_1b\left(\frac{3q_1^*+2q_2^*}{2\sqrt{q_1^*+q_2^*}}+q_1^*\right)$$

$$M_2=k_2+1-u_2(a-c)+u_2b\left(\frac{2q_1^*+3q_2^*}{2\sqrt{q_1^*+q_2^*}}+q_2^*\right)$$

$$q_1^* = q_2^* = \frac{2(a-c)^2}{9b^2}$$

From Jury conditions, the necessary and sufficient conditions for $|l_i| < 1, i = 1,2,3,4$ are:

$$\begin{cases} 1 + B_1 + B_2 + B_3 + B_4 > 0 \\ 1 - B_1 - B_2 - B_3 + B_4 > 0 \\ |B_4| < 1 \\ |A_4| < |A_1| \\ |C_1| < |C_3| \end{cases} \qquad (12)$$

Where $B_1 = M_1 + M_2$ , $B_2 = M_1M_2 + k_2$ , $B_3 = k_1 - k_2M_1 - \frac{b^2 q_{1,t}q_{2,t}}{4(q_{1,t} + q_{2,t})}$ , $B_4 = k_1k_2 - k_1M_2$ ; $A_1 = 1 - B_1^2$ , $A_2 = B_1 - B_3B_4$ , $A_3 = B_2 - B_2B_4$ , $A_4 = B_3 - B_1B_4$ , $C_1 = A_4^2 - A_1^2$ , $C_2 = A_4A_3 - A_1A_2$ , $C_3 = A_4A_2 - A_1A_3$ .

So the equilibrium point $E^*$ of the system (10) is stable, if the conditions in (12) are all satisfied.

We reconsider the unstable situation ( $a = 8, b = 1.09, c = 2, u_1 = u_2 = 0.0082, q_{1,0} = 8, q_{2,0} = 6$ ) in Section 3. Because it(12) is so hard to solve, we give the control results $k_1 = 0.85$ , and $k_2 = 0.2$ . Let $k_1 = 0.85$, and $k_2 = 0.2$. Now the output and profit systems (10) become stable, as Fig. 3 shows. In Fig.3a,the red point shows that the changes of Productions1,2,when both manufacturers introduce time-delayed feedback control strategy. In Fig.3b,the blue point shows the proft.

Through numerical simulation，we obtained the results (Fig 3). This two methods can also control unstable system to achieve stable state. The stability of two strategies is at the same point ; It also shows that two manufacturers introducing time-delay feedback will make the early production more violently shocks . If the control is bad, it may lead the field production into chaos, thereby affecting the market stability and economic benefits. However, by introducing two time-delay feedbacks will make system achieve stability in shorter time than by adding one.

## 5. Conclusion

This article is about Cournot game for the competition of output. And we have studied two strategies of output's adjustment under incomplete information – the tit-for-tat strategy with time-delayed feedback. In conclusion, the cooperation may be achieved under the tit-for-tat strategy. But the stability of the adjustment system is sensitive to the parameters, and the Pareto Optimality cannot be assured. By introducing the feedback control to the cooperation intention of the players, the firms' cooperation can be achieved, and the Pareto Optimality is stable within the parameters' certain field. So the cooperation can be the result of such a strategy under the certain condition.

## References

[1] Zhang Jixiang, Da Qingli, Wang Yanhua. Analysis of nonlinear duopoly game with heterogeneous players[J]. Econ Model 2007;24:138–48.

[2] Agiza HN, Hegazi AS, Elsadany AA. The dynamics of Bowley's model with bounded rationality[J]. Chaos, Solitons & Fractals 2001;12:1705–17.

[3] Agiza HN, Elsadany AA. Nonlinear dynamics in the Cournot duopoly game with heterogeneous players[J]. Physica A 2003;320:512–24.

[4] Ahmed E, Agiza HN, Hassan SZ. On modifications of Puu's dynamical duopoly[J]. Chaos, Solitons & Fractals 2000;11:1025–8.

[5] Naimzada Ahmad K, Sbragia Lucia. Oligopoly games with nonlinear demand and cost functions: two boundedly rational adjustment processes[J]. Chaos,Solitons & Fractals 2006;29:707–22.

[6] Agiza HN, Hegazi AS, Elsadany AA. Complex dynamics and synchronization of a duopoly game with bounded rationality[J]. Math Comput Simul2002;58:133–46.

[7] Agliari Anna. Homoclinic connections and subcritical Neimark bifurcation in a duopoly model with adaptively adjusted productions[J]. Chaos, Solitons &Fractals 2006;29:739–55.

[8] Du Jianguo, Huang Tingwen. New results on stable region of Nash equilibrium of output game model[J]. Appl Math Comput 2007;192:12–9.

[9] Cafagna Vittorio, Coccorese Paolo. Dynamical systems and the arising of cooperation in a Cournot duopoly[J]. Chaos, Solitons & Fractals 2005;25:655–64.

[10] Elettreby MF, Hassan SZ. Dynamical multi-team

Cournot game[J]. Chaos, Solitons & Fractals 2006;27:666–72.

[11] Zhanwen Ding , Guiping Shi. Cooperation in a dynamical adjustment of duopoly game with incomplete information[J]. Chaos, Solitons and Fractals 42 (2009) 989–993

[12] Ding J, Yang W G, Yao H X, .Adaptive control the cournot duopoly production2model[J ] . Systems Engineering - Theory & Practice , 2008 ,2:111-118

# ICT-based Teaching and Learning in Higher Education – A Study

Dr. R. Krishnaveni[1], J. Meenakumari[2]

[1]Professor, PSG Institute of Management, PSG Institutions, Coimbatore, India
[2]Asst. Professor, Alliance University, Alliance Business School, Bangalore, India
Corresponding Addresses
{ krishnavenirm10@gmail.com , j_meenakumari@yahoo.com}

***Abstract***: Technology has become an indispensable tool in all aspects of life. It has transformed our life in many ways including the teaching-learning pattern. At present there is a transformation from traditional learning to a flexible learning scenario. Technology enhances ones learning by eliminating the geographical barriers, time and space constraints thereby enhancing life-long learning. In specific Information and Communication Technologies (ICT) have brought about significant changes in the higher education sector. ICT has been used in various aspects of teaching learning process in higher education. This paper presents the purpose of ICT-based teaching-learning in higher education institutions. It also identifies and presents the various items that contribute to effective ICT-based teaching-learning process. This paper brings out the extent to which the identified items contribute to ICT-based teaching-learning process in the present scenario. A path model for teaching-learning process was built and estimated. The path model was found fit to be implemented in higher education institutions to increase ICT-based teaching-learning process.

***Keywords***: Traditional learning, Transformation, Information and Communication Technology (ICT), Flexible learning, Path Model

## 1. Introduction

Knowledge has become the key factor in economic development. Knowledge revolution has also given rise to increased rate of innovations and a shorter product lifecycle. The advent of new economy based on advanced technologies and globalization combined with factors such as radical changes in the knowledge requirements and competition have enhanced the pressure on acquiring generic skills by all individuals in the current era.

Learning on a continued basis is required to bridge the skill gap between the requirements and competence of the individuals. Across the globe, information and communications technologies (ICTs) are changing the face of education. Our world is changing, and information and Communication technology (ICT) is central to this change Kader Asmal (2003). ICT and higher education changes are happening for improvement, innovation and for transformation. It has penetrated to all the aspects to learning and teaching process. The impact of ICT on traditional educational theories and practices are increasingly apparent. It has transformed and expanded the conventional boundaries of education. New innovations such as virtual colleges, laboratories, and universities are creating an abundance of additional areas of study surrounding this innovation.

A Technology–based learning systems unlike the traditional learning environment, is not bound by rigid timeframes, and location, and the learning experience is life-long. Rapid telecommunication and technological growth has paved the way for web-based education systems. We are in the transition period from a traditional to web-based learning system. Web-based education systems constitute one of the fastest growing areas in educational Technology, research and development. The mantra of web-based education system is “any time education anywhere” and “learning on the web rather than learning about the web”.

## 2. Objective of the Study

The Objective of this study is to determine various factors that contribute towards ICT-based teaching-learning processes. The extent of usage of technology for teaching-learning process in present higher education system and to identify the areas into which Technology is used to a larger extent in teaching -learning environment and to highlight the areas in which it could be used in an effective manner by proposing a validated model.

## 3. Theoretical Background

Education in India has evolved over the ages and continues to evolve, as witnessed from the progress of education right from ancient Indian times through the medieval period and pre-independence. Higher education in India gained momentum slowly and some of the ancient universities include Taxila, Vikramshila and Nalanda. Presently, India has hundreds of universities and thousands of colleges affiliated to them. A multitude of colleges have facility to focus on multiple disciplines. This has led to enhancement of the spread and quality of education in India. According to Whitworth and Berson (2003), ICT-enabled education has the potential to promote the development of students' decision-making and problem solving skills, data processing skills, and communication capabilities. ICT plays a major role for dealing with information and its transformation into knowledge, which is a basic requirement for citizens to become effective participants in this new scenario (Venezky, R.L. and C. Davis 2002).

ICT changes education from institution-centric to learner-centric, from classroom-based to being pervasively connected through e-learning and wireless technologies. New classes of learners are created, namely those who are not able to come to campus, or cannot afford a fixed timeframe, or those who would want to have a tailored program for their specific needs. ICT allows us to have the flexibility in space, time, and content.

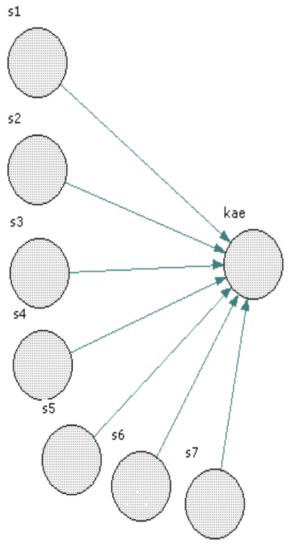
Learning using electronic means involves the acquisition of knowledge and skill using electronic technologies such as computer-and Internet-based courseware and local and wide area networks, and this is called e-learning (Encarta Dictionary, 2009). The use of technology in education, commonly defined as e-learning, has become a standard component in many courses. Technology applications are not limited to the classroom, and they are also replacing some classroom sessions with virtual sessions or fully replacing classroom courses with online courses (Paul Arabasz et.al. 2003).

The continuous innovation in ICT is causing an industrial and societal evolution based on information acquisition and knowledge dissemination (Branscomb, 1994), information networks represent the vehicles through which information and knowledge are being acquired and disseminated. Literature attests the power ICT can have in teaching and learning processes (Fonkoua, 2006; Newhouse, 2002). It has been suggested that using technology well in classrooms can even prepare students to be more effective citizens (John &Sutherland, 2004). In general, ICT can be considered as a vital tool to generate opportunities for attention to the increasing demand for higher education, as well as for improvement of academic processes and coordination of those with the society (Proyecto Académico 2007).

Further Demarest (1997), and Davenport et al., (1998), based on previous studies concluded that the process of knowledge management and the use of information technology can lower the cost of information usage and increase the speed of knowledge flow . From the above literature review, it is evident that ICT is playing a vital role in knowledge acquisition and there are lots of benefits of introducing ICT in the knowledge process. ICT is applied very effectively into a range of teaching strategies, including some very good interactive teaching involving questioning and discussion. ICT played an important role in helping teachers to demonstrate and reinforce key ideas during lessons using 'electronic blackboard', and e-mail was used efficiently to support homework (Douglas Osler 2000).Walter Omana and Theo Van der Wieda (2009) have clearly mentioned in their framework that knowledge capture/acquisition, knowledge store, knowledge share, knowledge enhancement, and knowledge dissemination can all be enhanced by technology and good policies.

## 4. Model building

Based on the above discussion, ICT is used as a tool for teaching and learning in higher education. The following path model for ICT-based teaching-learning process in higher education institutions with the various indicators was arrived at, as depicted below (Figure 1):



S1 - Usage of internet to supplement book information
S2 - Usage of ICT –based communication among faculties and students
S3 - Using Power Point Slides for teaching
S4- Evaluation of Test, Assignments and publication of results done electronically
S5- Expert interaction through ICT -based technology
S6 - Existence and extent of usage of Virtual library and Virtual leaning (e-learning)
S7- Usage of computers for multimedia-based delivery

**Figure 1.** Theoretical model for ICT-based teaching-learning process in higher education institutions

## 5. Methodology

This study is descriptive in nature with the population being approved higher education institutions. The sampling frame consists of 166 institutions and the sample size includes 50% of the total population. Random sampling technique was used for selecting the institutions. The instrument was validated for the items, and reliability and content validity test were also done. Further the theoretical model was validated using PLS.

## 6. Analysis

The following table depicts the important ICT factors that contribute to teaching-learning process in higher education institutions.

**Table 1.** Item categories generated for Teaching-learning Process

| Process | ICT factors |
|---|---|
| **ICT-based Teaching-Learning process** | Internet browsing to supplement book information |
| | Going through specialized papers / slides of various authors on the Internet |
| | Access to discussion boards / forums on the Web |
| | Usage of e-mails to interact with other professors / experts to enhance knowledge |
| | Usage of Technology for research work |
| | Existence of Virtual library |
| | Existence of Virtual learning (e-learning) |

The path model depicted in Figure 1 shows the significance of relationships between the constructs. An analysis was done to study the relationship between the factors and their impact on teaching learning process.

## 7. Findings and Discussion

It has been found from the literature review that ICT plays a major role in education and especially in teaching-learning process. The following were arrived at based on further analysis.

- The various items that contribute to ICT-based teaching-learning process are identified and depicted in Table 1
- Teaching-learning process has the average mean value of 3.72 which clearly indicates a good coverage of the various factors contributing to it.
- The demographic factors taken into consideration include the type of the institution, place and years of existence of the institution, the type of university to which they are affiliated, and the department and experience of the respondents. The demographic analysis did not reveal any statistically significant difference among the factors contributing to teaching-learning process

The contribution of individual items related to ICT-based Teaching-Learning process is represented in Table 2.

**Table 2.** Current contribution of items for ICT-based teaching-learning process

| **Knowledge Acquisition and Enhancement** | **Contribution %** |
|---|---|
| **Usage of Internet to supplement book information** | 96.5% |
| **Usage of ICT-based communication among faculties and students** | 69.2% |
| Usage of Power Point Slides for delivery of lectures | 89.5% |
| **Evaluation of Test, Assignments and publication of results done electronically** | 55.6% |
| Usage of Technology for research work and for expert discussions | 84.9% |
| **Existence and extent of usage of Virtual library and Virtual learning (e-learning)** | 62.2% |
| **Usage of computers for multimedia-based delivery** | 76.7% |

It is evident from the above data that both teachers and learners use ICT-based tools for teaching-learning process. There is a lifelong thirst for learning and knowledge enhancement and ICT acts as a facilitating tool for it. The data in the above table (Table 2) reveals that learners use ICT-based tools to the highest for supplementing book information and in their research work. Teachers use ICT –based tools for delivering lectures. ICT can be used to enhance the existing knowledge through expert interactions and through discussion board and forums. There is a transition from traditional learning system to e-learning system in the higher education scenario. The analysis reveals that there is much scope in the area of developing virtual libraries in higher education institutions. Traditional library systems are still in existence to a large extent though computers are used for managing the day-to-day transactions of the library system. This study further reveals that application of ICT-based tools in the area of examination and evaluation needs improvement.

The theoretical model represented in Figure 1 was estimated and evaluated using Partial Least Squares technique to arrive at the path model. The model fit was ensured with the statistically significant values of R-squared (61.3%) for ICT-based teaching-learning process. It was observed that there was a good correlation between the various indicators towards ICT-based Teaching Learning process. The outcome is depicted in Figure 2:

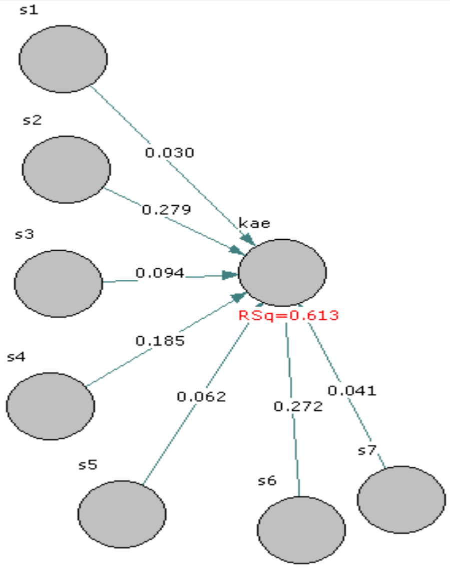


**Figure 2.** PLS Path model for ICT-based teaching-learning process in higher education institutions

## 8. Conclusion

At present Technology has become an indispensible tool for education. It is mainly intended to improve education access, so as to provide education for all. The outcome of this study could serve as an input for education planners to consider better utilization of ICT in various aspects relating to teaching-learning activities.

## References

[1] Bhattacharya, I. & Sharma, K., 'India in the knowledge economy – an electronic Paradigm', International Journal of Educational Management Vol. 21 No. 6, pp. 543-568, 2007

[2] Chickering et.al (1996),” Implementing the seven principles: Technology as lever”. *AAHE Bulletin, 49*(2), 3–6.Retrieved September 21, 2005, from http://2md.osu.edu/edtech/pdfs/seven_principles.pdf

[3] Chong Chee Keong, Sharaf Horani & Jacob Daniel A Study on the Use of ICT in Mathematics Teaching MOJIT, December 2005

[4] Demarest, M, Understanding knowledge management", *Long Range Planning*, Vol. 30 No.3, pp.374-84, 1997.

[5] Diem Ho , Research, Innovation and Knowledge Management: the ICT Factor Submitted to UNESCO, July 20, 2007

[6] Douglas Osler,, “The Use of ICT in Learning and Teaching” A Report by HM Inspectors of Schools: Scottish Executive Education, 2000

[7] Davenport, T.H., De Long, D.W., Beers, M.C,, "Successful knowledge management projects", Sloan Management Review, Vol. 39 No.2, pp.43-57, 1998.

[8] Hans N. Weiler, Higher education in India: Reflections on some critical issues, 2007

[9] Josephine Obioha The role of ICT in information seeking and use amongst research officers in research institutes in Nigeria: The Nigerian institute for oceanography & marine research institute experience The International Information & Library Review Volume 37, Issue 4, Pages 303-314, December 2005,

[10] Mooij, T, 'Design of educational and ICT conditions to integrate differences in learning: Contextual learning theory and a first transformation step in early education', Computers in Human Behavior 23(3), 1499—1530, 2007

[11] Sanyal, B. C, 'New functions of higher education and ICT to achieve education for all', Paper prepared for the Expert Roundtable on University and Technology-for-Literacy and Education Partnership in Developing Countries, International Institute for Educational Planning, UNESCO, September 10 to 12, Paris, 2001

## Author Biographies

**First Author** Dr. R.Krishnaveni (b. 1962) is presently Professor in PSGIM, PSG College of Technology, Coimbatore, Tamilnadu. She has twenty five years of teaching and research experience. Her publication includes 4 books, 60 research articles in international and national journals. She has been instrumental in organizing National level conference as well as workshops annually in the area of Business Research. Her popular book includes “Human Resource Development – a Researcher’s Perspective” (2008). She is the executive editor of the journal “Journal of contemporary research in management “. She is also the recipient of ‘Outstanding Woman Researcher’ award for the year 2009 from AIMS International held at IIM-Bangalore. She can be contacted at rmkrishnaveni@gmail.com

**Second Author** J.Meenakumari b (1973) is presently working with Alliance University in the department of Information Systems, Bangalore. She has submitted her doctoral thesis and prior to this she has obtained her Master’s Degree in Computer Science and Master’s of Philosophy degree with distinction. She has more than one decade of teaching experience. Her publication includes several articles in international and national journals and papers in International and National conferences. She has won two best paper awards in International conferences. She is a committee member of International Association of Computer Science and Information Technology (IACSIT). She can be contacted at j_meenakumari@yahoo.com

.

# Semantic Approach towards the Sensor Web Enablement

Deept Gupta, Satvika Khanna, Sakshi Bhatia

TIT&S, Bhiwani, India
(deeptiguptalight@gmail.com, satvika16oct@gmail.com, sakshi.bhatia@gmail.com)

***Abstract*** - Sensor webs for science have evolved considerably over the past few years. Sensor web provides an infrastructure that coordinates distributed, heterogeneous, a large number of sensor data resources. Many of today's sensor webs employ little autonomy. The deployment and usage of sensors is usually tightly coupled with the specific location, application, and the type of sensors being used. Various applications for sensor webs require data from heterogeneous sensors and even integrating multiple sensor webs. To be effective, this will require a capability for publishing and discovering sensor resources. Once this infrastructure is in place, it will be much easier to pull additional sensors into a particular sensor web application. In the demand of providing autonomy capabilities to sensors, a semantic approach to the sensor web technology is applied. It provides a proposed solution towards the challenges for sensor web technology. This article proposes the resolution of challenges like interoperability, autonomy, data integration by semantic sensor web approach. It allows sensor networks to interact with other sensors.

***Keywords:*** sensors, sensor webs, sensor web enablement, interoperability, semantics, semantic sensor web

## 1. INTRODUCTION

SENSOR networks are used in a broad variety of applications ranging from environmental monitoring and public health to disaster management and monitoring of public infrastructures. Sensors around the globe currently collect huge amount of data about the world. The rapid development and deployment of sensor networks and the lack of integration and communication between these networks is intensifying the existing problem of too much data and not enough knowledge. The combination of sensor networks with the Web, web services and database technologies, is termed as the Sensor Web. The term "Sensor Web" was first used by Kevin Delin of NASA in 1997, to describe a novel wireless sensor architecture where the individual pieces could act and coordinate as a whole.

As IP-enabled, affordable sensor devices of different types become available and are placed around, referred to as a "Sensing Cloud", in our environment, integrating the diverse sensory streams into the web can serve different user or machine queries.

General architecture of Sensor Web applications can be characterized by:

-variable and heterogeneous data, devices and networks.

-unreliable nodes and links, noise, uncertainty

- Vast data sources (sensors, images, GIS, etc.) in different settings (live, streaming, historical, and processed);

- Existence of multiple administrative domains

- need for managing multiple, concurrent, and uncoordinated queries to sensors.

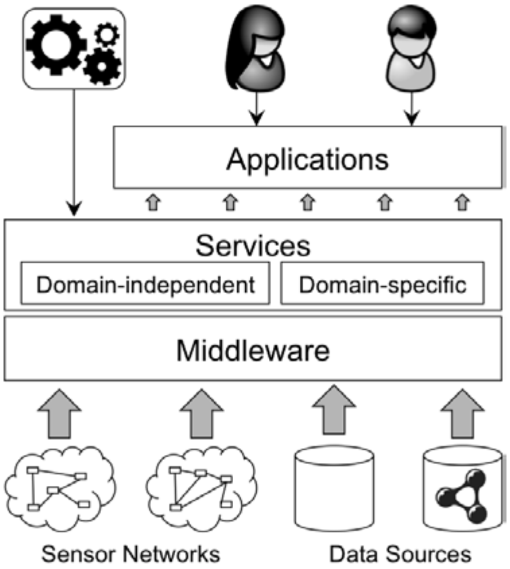


Figure 1. Working of sensor web

## 2. SENSOR WEB ENABLEMENT

Sensor technology, computer technology and network technology are advancing together while demand grows for ways to connect information systems with the real world. The SWE effort involves OGC members in developing the global framework of standards and best practices that make linking of diverse sensor related technologies fast and practical.

The Sensor Web Enablement (SWE) standards enable developers to make all types of sensors, transducers and sensor data repositories discoverable, accessible and useful via the Web. In much the same way that Hyper Text Markup Language (HTML) and Hypertext Transfer Protocol (HTTP) standards enabled the exchange of any type of information on the Web, the SWE initiative is focused on developing standards to enable the discovery, exchange, and processing of sensor observations, as well as the tasking of sensor systems. Sensor location is usually a critical parameter for sensors on the Web. The goal of SWE is to enable all types of Web and/or Internet-accessible sensors, instruments, and imaging devices to be accessible and, where applicable, controllable via the Web. It has a goal of allowing people to publish their sensor network data in such a way that other people's search and analysis systems can *automatically* find the information

The functionality that OCG has targeted within a sensor web includes:
-Discovery of sensor systems, observations, and observation processes that meet an application's or users immediate needs;
- Determination of a sensor's capabilities and quality of measurements;
- Access to sensor parameters that automatically allow software to process and geo-locate observations;
- Retrieval of real-time or time-series observations and coverage in standard encodings
- Tasking of sensors to acquire observations of interest;
- Subscription to and publishing of alerts to be issued by sensors or sensor services based upon certain criteria.

Sensor Web Enablement standards that have been built and prototyped by members of the OGC include the following pending OpenGIS Specifications:
1. **Observations & Measurements Schema (O&M)** – Schema for encoding observations and measurements from a sensor, both archived and real-time.
2. **Sensor Model Language (SensorML)** –Schema for describing sensors systems and processes; provides information needed for discovery of sensors, location of sensor observations, processing of low level sensor observations, and listing of taskable properties.
3. **Transducer Markup Language (TransducerML or TML)** – Schema for describing transducers and supporting real-time streaming of data to and from sensor systems.
4. **Sensor Observations Service (SOS)** - Standard web service interface for requesting, filtering, and retrieving observations and sensor system information. This is the intermediary between a client and an observation repository or near real-time sensor channel.
5. **Sensor Planning Service (SPS)** – Standard web service interface for requesting user-driven acquisitions and observations. This is the intermediary between a client and a sensor collection management environment.
6. **Sensor Alert Service (SAS)** – Standard web service interface for publishing and subscribing to alerts from sensors.
7. **Web Notification Services (WNS)** – Standard web service interface for asynchronous delivery of messages or alerts from SAS and SPS web services and other elements of service workflows.[2]

## 3. Management of sensor web data

The worldwide sensor web will generate too much data to visualize or analyze manually.
Most sensor network researchers would probably agree that we have placed too much attention on the networking of distributed sensing and too little on tools to manage, analyze, and understand the data. the sensor web must incorporate logical data abstractions and visualizations that can shield users from the complexities of the underlying sensing infrastructures but still propagate measures of uncertainty associated with calibration or sampling effects. A useful query on the worldwide sensor web might need to compare or combine data from many heterogeneous data sources maintained by independent entities. For example, while treating a patient, healthcare professionals might query hospitals for the patient's health profile and airports for his or her recent travels. They might then correlate this information with similar information from other patients suffering from similar diseases. [3]
SOS Implementation Specification is a critical element of the SWE architecture, defining the network centric data representations and operations for accessing and integrating observation data from sensor systems. The SOS is the intermediary between a client and an observation repository or near real-time sensor channel. Clients can also access SOS to obtain metadata information that describes the associated sensors, platforms, procedures and other metadata associated with observations. The client depends on registries that provide metadata for the different types of sensors and the kinds of data that they are capable of providing. Centralized registries for sensor-based data have appeared focused on the registration of sensor-based data sources, and on the provision of access to them in multiple ways.[6]

## 4. Challenges with Sensor Web Observations

Sensor webs encounters various challenges related to the characteristics of the data sources that are handled in typical Sensor Web applications and the creation of applications based on these data sources. Some of these challenges are discussed here.

First concern is related to the abstraction level in which sensor data can be obtained, processed and managed in general. Sensor data can be managed at a very low level, at the device- and network-centric levels, generally by means of using low-level programming languages and operating systems. But it can be also managed through higher-level formalisms (e.g., via declarative continuous queries over streams), thereby insulating clients and users from the infrastructural and syntactic heterogeneities of autonomously-deployed sensor networks
Second challenge is related to the adequate characterization and management of the quality of sensor

data. Issues like the unavailability of a piece of data over a period of time may have different meanings when seen from an application perspective: the sensor was not available, there was no event to trigger the data generation during that time, the communication with the sensor was broken, etc. Other issues like the accuracy of the sensed data may depend on a number of internal and external conditions to the sensor network. In summary, there are a number of quality characteristics that are relevant to the quality of service and that may affect the results obtained from a data observation process.

The sensor web is facing the problem of integration and fusion of data coming from autonomously-deployed sensor networks, with varying qualities of service and different throughput rates, geographical scales, etc. This is related not only with the integration of data coming from different sensor networks, but also with the combination of such data with data persisted in other sources, such as static data or archived sensor data. Even if the sensor web data sources used well-defined interfaces to publish their data, the complex and semantically disparate measures of data quality and uncertainty typically associated with sensor webs make data fusion a challenge.

It evolves in the problem of identify the location of relevant sensor-based data sources with which data integration and fusion tasks can be performed. The number of sensor networks being deployed in the real world is growing continuously,. As a result, more experiments and initiatives deploy sensor networks in different areas, and finding the right information to be used in integration and fusion tasks is highly relevant.

Finally, another important challenge has to do with the need to enable the rapid development of applications that are able to handle sensor data, taking into account the aforementioned characteristics and challenges. This includes dealing with data integrity and validation issues as well as the need for common interfaces and formats between applications, databases, sensor networks, etc. This challenge requires enabling the development of applications with different resource models and qualities of service (e.g., energy, bandwidth, processing, and storage) and facilitating the interaction with sensor data from the developer and user points of view.[4]

## 5. Semantic approach to sensor web-SEMANTIC SENSOR WEB

The sensible use of the term "semantics" refers to the meaning of *expressions* in a language. The semantics required to achieve interoperability is that of expressions built from symbols in service descriptions.

Much of the query-processing task in the worldwide sensor web will be automated; data must have a well-defined syntax and semantics. The Semantic Web can address many of the technical challenges of enabling interoperability among data from different sources. This technology enables information exchange by putting data with computer-processable meaning (semantics) on the World Wide Web.

The Semantic Web has three key aspects. First, data is encoded with self-describing XML identifiers, enabling a standard XML parser to parse the data. Second, the identifiers' meanings (properties) are expressed using the Resource Description Framework. RDF encodes the meaning in sets of triples, each triple being like an elementary sentence's subject, verb, and object, with each element defined by a URI (uniform resource identifier) on the Web. Ontologies express the relationships between identifiers. For example, two data sources can publish data in XML as "<Temperature><Celsius>20</Celsius></Temperature>" and "<Temperature> <Fahrenheit> 68 </Fahrenheit> </Temperature>." An associated RDF document can describe that Celsius and Fahrenheit are temperature units, and ontology can define the relationship between Celsius and Fahrenheit. So, a data-processing system can automatically infer that these two data points represent the same temperature value. Major industries are working to establish their own ontological standards for the Semantic Web.

Previously, the data processed by a GIS as well as its methods had resided locally and contained information that was sufficiently unambiguous in the respective information community. Now, both data and methods may be retrieved and combined in an ad hoc way from anywhere in the world, escaping their local contexts. They contain attributes, data types, and operations with meanings that differ from those implied by locally-held catalogues and manuals. Since the semantics specified by these local resources is not machine-readable, it cannot be shared with other systems. [4,9]

One of the main open issues in the development of applications for sensor network management is the definition of interoperability mechanisms among the several monitoring systems and heterogeneous data. In the last years, the Service-Oriented Architecture (SOA) approach has become predominant in many sensor network projects as it enables the cooperation and interoperability of different sensor platforms at a higher level of abstraction. The Semantic Sensor Web (SSW) proposes that sensor data be annotated with semantic metadata that will both increase interoperability and provide contextual information essential for situational knowledge.[4]

A number of sensor network ontologies have been created, which aim at describing different aspects of sensor-based data, from the device point of view (focusing on the hardware that is being used in order to generate the data) to the domain point of view (focusing on the types of data that can be generated from sensors and sensor networks in the context of specific domains). Several aspects are relevant in the development of most of these ontologies, such as the distinction between raw observed data and derived data, the representation of aspects like accuracy, or the consideration of observations and measurements.

In the context of identifying and locating relevant sensor-based data in the real world, sensor data registry interfaces are defined, and an appropriate infrastructure that can cope with the types of queries that are usually handled in sensor-based applications is being developed. These registries should provide support for spatio-temporal queries (e.g., "get sensor data sources that contain information about the temperature in this region for the last two days") and for metadata queries related to existing sensor network ontologies.[6] Semantic queries that are adapted to sensor-based data are formulated. They provide declarative querying infrastructure to define logical views over sensor network data and open the way for view and ontology-based techniques to be used.

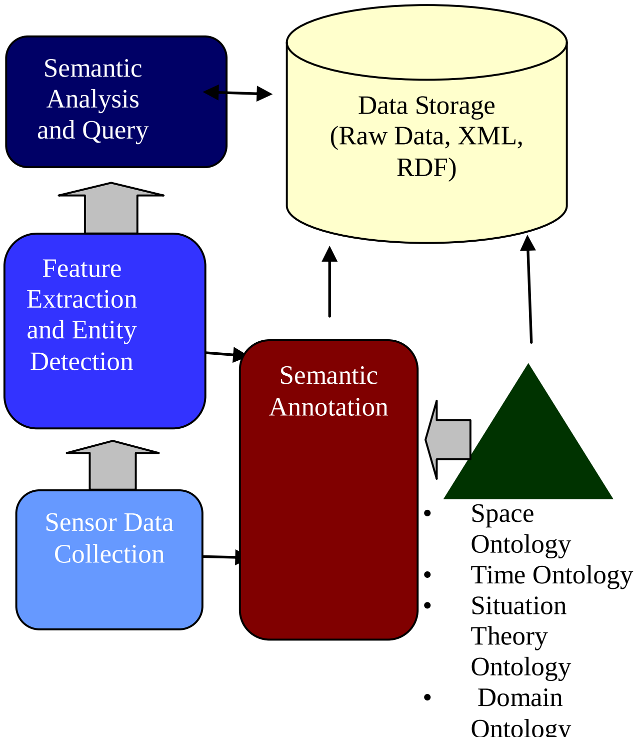

Semantic Analysis and Query
Data Storage (Raw Data, XML, RDF)
Feature Extraction and Entity Detection
Semantic Annotation
Sensor Data Collection
• Space Ontology
• Time Ontology
• Situation Theory Ontology
• Domain Ontology

Figure 2. Semantically query processing

## 6. Conclusion

Building more effective sensor webs involves many different challenges in the areas of information standardization and autonomy. The challenges in information standardization have evolved from the difficulty in the collection and analysis of information from many different types of sensors. We need to create data standards so that the different sensor data and the models that use them can be fused together to answer complex scientific questions. Different users have different views of the sensor data depending on their particular need. This problem is being addressed with the evolving concept of semantic view to current syntactic web technology. In this respect, ontologies are being created that will infuse metadata into the sensor data. This will allow data that can be filtered, summarized, and transformed, and will also allow features to be extracted into higher level features. In addition, the same data can be reused for different applications. The next generation semantic sensor web can be an effective proposed solutions of today's traditional sensor web technology.

## References

[1]. Vagan Terziyan and Oleksandr Kononenko, "Semantic Web Enabled Web Services: State-of-Art and Industrial Challenges",LNCS.

[2]. Mike Botts,George Percivall.Carl Reed,John Davidson," OGC® Sensor Web Enablement: Overview And High Level Architecture", *Proceedings of the 5th International ISCRAM Conference – Washington, DC, USA, May 2008* F. Fiedrich and B. Van de Walle, eds.

[3]. Magdalena Balazinska *University of Washington* Amol Deshpande *University of Maryland* Michael J. Franklin *University of California, Berkeley* Phillip B. Gibbons *Intel Research* Jim Gray and Suman Nath *Microsoft Research* Mark Hansen *University of California, Los Angeles* Michael Liebhold *Institute for the Future* Alexander Szalay *Johns Hopkins University* Vincent Tao *Microsot,* "Data Management in the Worldwide Sensor Web", PERVASIVEcomputing *Published by the IEEE Computer Society*

*[4]* Oscar Corcho and Raúl García-Castro," Five challenges for the Semantic Sensor Web"

[5] FREDDY DUITAMA,BRUNO DEFUDE ,AMEL BOUZEGHOUB,CLAIRE LECOCQ," A Framework for the Generation of Adaptive Courses Based on Semantic Metadata", 2005 Springer Science + Business Media, Inc. Manufactured in The Netherlands.

[6] Miao-Miao Wang;, Jian-Nong Cao2, Jing Li, and Sajal K. Dasi," Middleware for Wireless Sensor Networks: A Survey", A survey. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 23(3): 305,326 May 2008.

[7] Arne Br¨oring, Krzysztof Janowicz, Christoph Stasch, and Werner Kuhn," Semantic Challenges for Sensor Plug

and Play", W2GIS 2009, LNCS 5886, pp. 72–86, 2009. Springer-Verlag Berlin Heidelberg 2009.

[8] Beniamino Di Martino," An Ontology Matching Approach to Semantic Web Services Discovery", ISPA 2006 Ws, LNCS 4331, pp. 550 – 558, 2006.
Springer-Verlag Berlin Heidelberg 2006

[9]. Sam Bacharach, Open Geospatial Consortium, Inc. (OGC)," Implementations of OGC Sensor Web Enablement Standards", Article for December 2007 Sensors Magazine.

[10] Nicholas J. Kings, Caroline Gale, and John Davies," Knowledge Sharing on the Semantic Web", ESWC 2007, LNCS 4519, pp. 281–295, 2007. © Springer-Verlag Berlin Heidelberg 2007.

[11] Amit Sheth ," Semantic Sensor Web" ARC Research Network on Intelligent Sensors, Sensor Networks and Information Processing – ISSNIP talk Melbourne, August 1, 2008

# Design of Decentralized Controller Gain Scheduling For Power System Restoration Assessment in an Interconnected Power System

R.Jayanthi[1 *], I.A.Chidambaram[2]

*[1*] Assistant Professor, Department of Electrical Engineering, Annamalai University, Annamalainagar -608002,India*
*[2] Professor, Department of Electrical Engineering, Annamalai University, Annamalainagar -608002, India*
*[*] Corresponding Author: e-mail: rrj_pavi@yahoo.co.in, Tel +91-4144-237092*

## Abstract

The problem of restoration assessment in a Two-Area Two-Unit thermal Reheat Interconnected Power System (TATURIPS) has been investigated with spinning reserves such as Super conducting Magnetic Energy Storage device (SMES) and Gas Turbine (GT) units. The Proportional Integral (PI) controller gains are tuned using an Evolutionary Algorithm (EA) Particle Swarm Optimization (PSO) technique to find best parameter for the tuning of controller. Thus to exemplify the optimal parameter search PSO is used in an uncertainity area of the controller. Simulation results emphasis on the better settling time based stability performance of optimized PI controller in the TATURIPS with SMES / GT units when compared with that of the conventional controller in an interconnected power system.

***Keywords:*** Particle Swarm Optimization, Multi-Area power systems, Gas Turbine, Super Conducting Magnetic Energy Storage Device, settling time.

## 1. Introduction

In an extremely complex and highly meshed power system the disturbances may be propagated over a vast area within a very short period of time. Even a simple incident may degenerate the system very rapidly into a large-scale breakdown. Therefore, it is necessary to anticipate any critical situation within a very few hours or minutes before the real time operation, preventing cascading and limiting its consequences for restoration of power system, which has to be carried out by implementing remedial actions [1]. Generally, ordinary controllers are designed with Proportional-Integral (PI) controllers. However, since the “I” control parameters are usually tuned, it is incapable of obtaining good dynamic performance for various load and system change scenarios. In literatures, many control strategies have been suggested based on conventional linear control theory, variable structure control, a lot of artificial intelligence based robust controllers such as genetic algorithm, tabu search algorithm, fuzzy logic and neural networks based robust controller are used for PI controller parameters tuning [2], [3]. Since, Particle Swarm Optimization algorithm is an optimization method that finds the best parameters for controller in the uncertainty area of controller parameters and obtained controller is an optimal controller, it has been used in almost all sectors of industry and science. One of them is the load-frequency control. In this study, it is used to determine the parameters of a PI controller according to the system dynamics, control over frequency, deviations, control the input and inter-area tie-power oscillations. By optimizing the values of proportional ($K_P$) and integral ($K_I$) gains, the output of the system frequency seems to be improved. In this simulation study the proposed controller is simulated for a TATURIPS. To show effectiveness of proposed method and also compare the performance of these two controllers, several changes in demand of first area, demand of second area and demand of two areas simultaneously are applied. Simulation results indicate that the overshoots and settling times with the proposed PSO-PID controller are better than the output of the conventional controller [4-9]. PSO controllers guarantee the good performance under various load conditions

The expert system, which is used to bring up the faulted power system to the target system which allows the estimation and observation of the real restoration time, the degree of stability, the observation of the system voltage profile, power to be transmitted is done by following tools [10].

**1.1Generation Management:** This tool is responsible for connecting generators. Firstly, it started by finding the smallest black start generator in the solution and then it connects the generators in accordance with the generator sequence given by the PSO-solution.

**1.2 Restoration Path Management:** In every step of connecting a generator or load, an optimized path algorithm is used to find the shortest path. Moreover the Path Management is used to check the loading limits of every line proposed for connection.

**1.3 Time Management:** Since one of the main goals of using the expert system is to estimate the real restoration time, great attention has been given to the time required for every element in every stage of restoration.

**1.4 Load management:** During restoration, loads are restored based on the load priorities and system security considerations. The priorities of loads are calculated in accordance to the load importance. If two loads are in the same degree of priority, the nearest one is picked. Moreover, if two or more loads are in the same degree of priority and in the same distance, the load with the highest level of connectivity is picked.

**1.5 Role of expert system in this proposed work:** The main objectives of expert system in knowledge based restoration are

- By providing initial source of power immediately to the interconnected power system with SMES and Gas Turbine units.
- By optimizing the gain values of the PI controller using PSO technique for the two-area interconnected power system with SMES and Gas Turbine units for system restoration.
- The primary function of the expert system is to restore the interconnected thermal reheat power system even for small disturbances and to avoid excess under frequency deviations.

## 2. Modeling of a two-area interconnected thermal reheat power system

Due to the inherent characteristics of changing loads, the operating point of power system may change very much during a daily cycle. The generation changes must be made to match the load perturbation at the nominal conditions, if the normal state is to be maintained.

The mismatch in the real power balance affects primarily the system frequency but leaves the bus voltage magnitude essentially unaffected. In a power system, it is desirable to achieve better frequency constancy than obtained by the speed governing system alone. This requires that each area should take care of its own load changes, such that schedule tie power can be maintained. A two-area interconnected system dynamic model in state variable form can be conveniently obtained from the transfer function model.

The state variable equation of the minimum realization model of the 'N' area interconnected power system is expressed as [11].

$$\dot{X} = \mathrm{A}x + \mathrm{B}u + \Gamma d \qquad (2.1)$$

$$\mathrm{Y} = \mathrm{C}x \qquad (2.2)$$

Where, the system state vector x consists of the following variables as:

$$[x] = \left[\int \mathrm{ACE}_1 dt, \int ACE_2 dt, \Delta F_1, \Delta P_{g1}, \Delta X_{e1}, \Delta P_{tie}, \Delta F_2, \Delta P_{g2}, \Delta X_{e2}\right]^T$$

$$u = [U_1, \ldots . U_N]^T = [\Delta P_{C1}, \ldots . \Delta P_{CN}]^T$$

N – Control input vector

$$d = [d_1, \ldots . d_N]^T = [\Delta P_{D1}, \ldots . \Delta P_{DN}]^T$$

N – Disturbance input vector

$$y = [y_1, \ldots . y_N]^T$$

2N – Measurable output vector

A is system matrix, B is the input distribution matrix and Γ disturbance distribution matrix, x is the state vector, u is the control vector and *d* is the disturbance vector of load changes of appropriate dimensions.The typical values of system parameters for nominal operation condition are given in appendix. This study focuses on optimal tuning of controllers for LFC and tie-power control, settling time based optimization using PSO algorithm to ensure a better power system restoration assessment. The aim of the optimization is to search for the optimum controller parameter setting that maximizes the minimum damping ratio of the system. On the other hand in this study the goals are control of frequency and inter area tie-power with good oscillation damping and also obtaining a good performance under all operating conditions and various loads and finally designing a low-order controller for easy implementation.

## 3. Modeling of a Gas turbine

Amid growing concerns about Green house emissions, Gas turbines have been touted as a viable option, due to their higher efficiency and the lower green house gas emissions compared to other energy sources and fast starting capability which enables them to be often used as peak units that respond to peak demands [12]. Also, they can be profitably used in power system restoration for supplying power to the restoration areas as they have the advantages like, Quick start-up/shut-down, Low weight and size, Cost of installation is less, Low capital cost, Black-start capability, High efficiency, Requires low cranking power, Pollutant emission control etc.

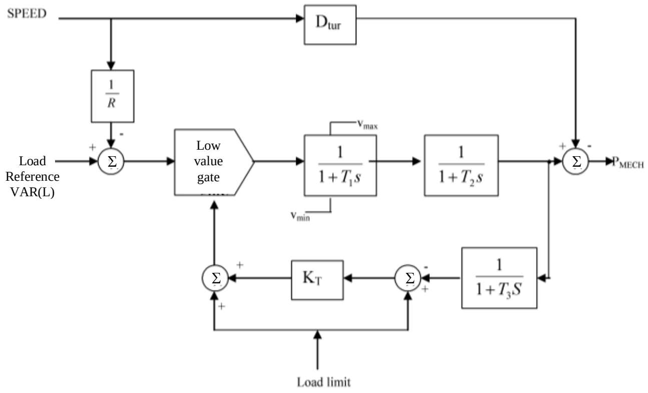


**Figure 1. Gas Turbine Model**

The continuous power plant output of a Gas turbine at the maximum depends upon frequency and temperature. It gives approximately two-thirds of the total power output of a typical combined cycle plant [13], [14]. When the load is suddenly increased the speed drops quickly, but the regulator reacts and increases the fuel flow to a maximum of 100% thereby improving the efficiency of the system. A model as shown in figure 1 has been implemented in the model used here.

## 4. Super Conducting Magnetic Energy Storage (SMES) device

The normal operation of a power system is continuously disturbed due to sudden small load perturbations. The problem lies in the fact that the inertia of the rotating parts is the only energy storage capacity in a power system. Thus, when the load-end of the transmission line experiences sudden load-end of the transmission line experiences sudden small load changes, the generators need continuous control to suppress undesirable oscillations in the control to suppress undesirable oscillations in the system.

The superconducting magnetic energy system is a fast acting device can swallow these oscillations and help in reducing the frequency and tie-line Power deviations for better performance of system disturbances. The Super conducting magnetic energy system is designed to store electric energy in the low loss superconducting coil. Power can be absorbed or released from the coil according to the system requirement. A super conducting magnetic energy storage(SMES) which is capable of controlling active and reactive power simultaneously has been expected as one if the most effective stabilizers of power oscillations [15].

Besides oscillation control, a SMES allows a load leveling, a power quality improvement and frequency stabilization. A typical SMES system includes three parts namely superconducting coil, power conditioning system and cooled refrigerator. From the practical point of view, a SMES unit with small storage capacity can be applied not only as a fast compensation device for power consumptions of large loads, but also as a robust stabilizer for frequency oscillations.

### 4.1. SMES Unit:

The schematic diagram in Figure 2 shows the configuration of a thyristor controlled SMES unit [16]. The SMES unit contains DC superconducting Coil and converter which is connected by Y–D/Y–Y transformer. The inductor is initially charged to its rated current $I_{d0}$ by applying a small positive voltage. Once the current reaches the rated value, it is maintained constant by reducing the voltage across the inductor to zero since the coil is superconducting. Neglecting the transformer and the converter losses, the DC voltage is given by

$$E_d = 2Vd0 \cos \alpha - 2 I_d R_c \qquad (4.1)$$

Where $E_d$ is DC voltage applied to the inductor (kV), firing angle (α), $I_d$ is current flowing through the inductor (kA). $R_c$ is equivalent commutating resistance (V) and $V_{d0}$ is maximum circuit bridge voltage (kV). Charge and discharge of SMES unit are controlled through change of commutation angle α.

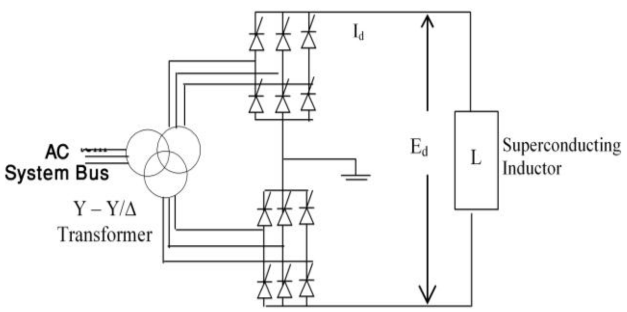


**Figure 2. The schematic diagram of SMES unit**

In AGC operation, the dc voltage $E_d$ across the superconducting inductor is continuously controlled depending on the sensed area control error (ACE) signal. Moreover, the inductor current deviation is used as a negative feedback signal in the SMES control loop. So, the current variable of SMES unit is intended to be settling to its steady state value. If the load is used as a negative feedback signal in the SMES control demand changes suddenly, the feedback provides the prompt restoration of current. The inductor current must be restored to its nominal value quickly after a system disturbance, so that it can respond to the next load disturbance immediately . As a result, the energy stored at any instant is given by

$$W_L = L I_d^2/2 \quad \text{MJ} \qquad (4.2)$$

Where L = inductance of SMES, in Henry

Equations of inductor voltage deviation and current deviation for each area in Laplace domain are as follows:

$$\Delta E_{di}(s) = \left(\frac{K_{SMES}}{1+sT_{dci}}\right)\left[\beta_1 \Delta F_1(s) + \Delta P_{tie1}(s)\right] - \frac{K_{id}}{1+sT_{dci}} \Delta I_{di}(s) \qquad (4.3)$$

$$\Delta I_{di}(s) = (1/sL_i) * \Delta E_{di}(s) \qquad (4.4)$$

Where

$\Delta E_{di}(s)$ = converter voltage deviation applied to inductor in SMES unit
$K_{SMES}$ = Gain of the control loop SMES
$T_{dci}$ = Converter time constant in SMES unit
$K_{id}$ = gain for feedback ΔId in SMES unit.
$\Delta I_{di}(s)$ = inductor current deviation in SMES unit

The deviation in the inductor real power of SMES unit is expressed in time domain as follows:

$$\Delta P_{SMESi} = \Delta E_{di} I_{doi} + \Delta I_{di} \Delta E_{di} \qquad (4.5)$$

Figure 3 shows the block diagram of the SMES unit. To achieve quick restoration of the current, the inductor current deviation can be sensed and used as a negative feed back signal in the SMES control loop.

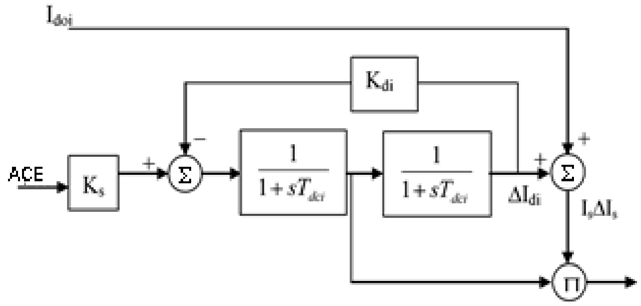


**Figure 3. Block diagram of SMES unit**

In a two-area interconnected thermal power system under study with the sudden small disturbances which continuously disturb the normal operation of power system. As a result the requirement of frequency controls of areas beyond the governor capabilities SMES is located in area1 absorbs and supply required power to compensate the load fluctuations.

Tie-line power flow monitoring is also required in order to avoid the blackout of the power system. The Input of the integral controller of each area is

$$ACE_i = \beta_i \Delta f_i + \Delta P_{tie\ i} \qquad (4.6)$$

Where,

$\beta_i$ = frequency bias in area i
$\Delta f_i$ = frequency deviation in area i
$\Delta P_{tie\ i}$ = Net tie power flow deviation in area i

The application of energy storages to electrical power system can be grouped into two categories.

1. Like conventional pumped hydro plant storage meant for load leveling application.
2. To improve the dynamic performance of power system.

SMES have the following advantages like: The time delay during charge and discharging is quite short, Capable of controlling the both active and reactive power simultaneously, Loss of power is less, High reliability, High efficiency.

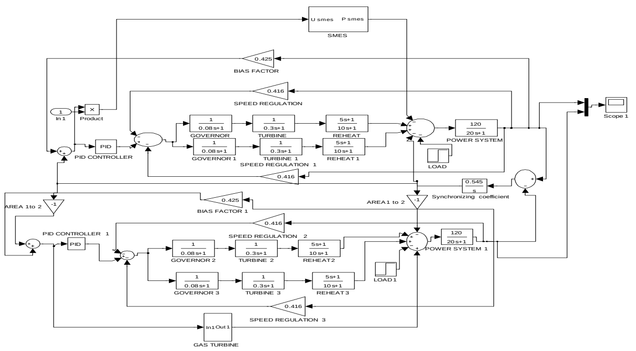


**Figure 4. Simulink model of TATURIPS with SMES and GT units**

## 5. Controller design using particle swarm optimization technique for the power system restoration problem

This is a population based search technique. Each individual potential solution in PSO is called particle. Each particle in a swarm fly around in a multidimensional search space based on its own experience and experience by neighbouring particles. Let in search space '*S*' in n-dimension with the swarm consists of 'N' particles. Let, at instant 't', the particle 'i' has its position defined by

$$X_t^i = \{x_1^i,\ x_2^i,\ \ .\ \ .\ \ .x_n^i\}$$

Velocity $V_t^i = \{v_1^i,\ v_2^i,\ \ .\ \ .\ \ .v_n^i\}$ in variable space '*S*'. Velocity and position of each particle in the next generation (time step) can be calculated as

$$V_{t+1}^i = \omega V_t^i + C_1.rand().(P_t^i - X_t^i) + C_2\ Rand()(P_t^g - X_t^i)$$

$$X_{t+1}^i = X_t^i + V_{t+1}^i$$

Where,

$N$ - number of particle in swarm
$\omega$ - inertia weight
$C_1,\ C_2$ - acceleration constant
$rand()\ Rand()$ - Uniform random value in the range [0, 1]
$P_t^i$ - best-position that particle '*i*' could find so far
$P_t^g$ - global best at generation 't'

Performance of PSO depends on selection of inertia weight ($\omega$), Max velocity ($V_{\max}$) and acceleration constant ( $C_1,\ C_2$ ).

**Effect of Inertia weight ($\omega$)**

Suitable weight factor helps in quick convergence. Large weight factor facilitates global exploration (i.e. searching of new area). While small weight factor facilitates local exploration (so wise to choose large weight factor for initial iterations and gradually smaller weight factor for successive iterations). Generally, $\omega$ 0.9 at beginning and 0.4 at end [17].

**Effect of Max velocity ($V_{\max}$)**

With no restriction on the max velocity of the particle, velocity may become infinitely large. If $V_{\max}$ is very low particle may not explore sufficiently. If $V_{\max}$ is very high it may oscillate about optimal solution. Therefore, velocity clamping effect has to be introduced to avoid '*swarm explosion*' [18]. Generally, max velocity is set as 10-20% of dynamic range of each variable. Velocity can be controlled within a band

$$V_{\max} = V_{ini} - \frac{V_{ini} - V_{fin}}{iter_{\max}}\ iter$$

**Effect of Acceleration constant ( $C_1, C_2$ )**

$C_1$ is called ***Cognitive Parameter*** which pulls each particle towards local best position. $C_1, C_2$ is called ***Social Parameter*** which pulls the particle towards global best position. Generally, $C_1, C_2$ are chosen between 0 to 4. The design steps of PSO based PI controller is as follows.

1. The algorithm parameters like number of generation, population, inertia weight and constants are initialized.
2. The values of the parameters $K_P$ and $K_i$ initialized randomly.
3. The fitness function of each particle in each generation is calculated.
4. The local best of each article and the global best of the particles are calculated.
5. The position, velocity, local best and global best in each generation is updated
6. Repeat the steps 3 to 5 until the maximum iteration reached or the best solution is found.

### 5.1 Simulink model of PSO Based PI Controller

The PSO algorithm is used to search an optimal parameter set containing $K_P$ and $K_i$ . The parameters used for tuning the PSO algorithm and simulink models are tabulated in table below:

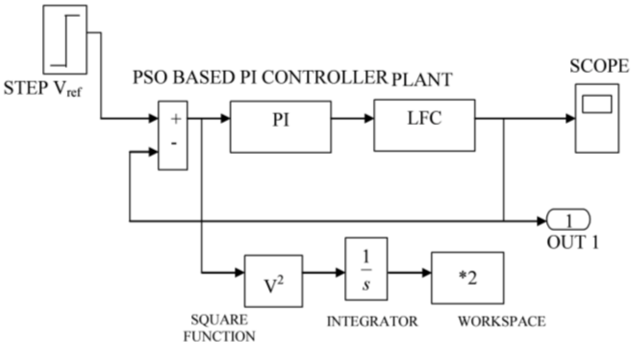


**Figure 5. Simulink model of plant with PSO Algorithm based PI Controller**

**Table 1:** ***Parameters values tuned for PSO***

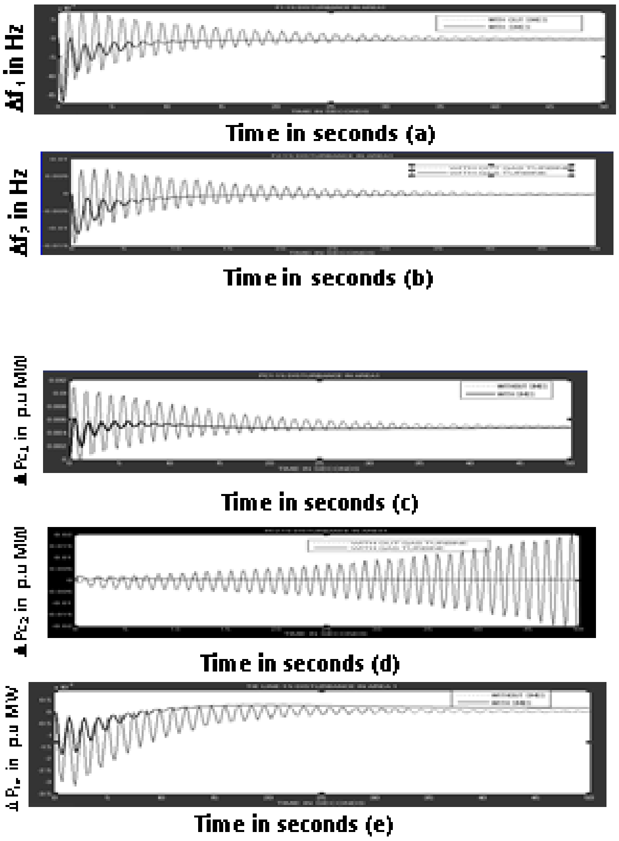
| **Parameters** | **LFC** |
|---|---|
| Population size | 5 |
| Number of generations | 10 |
| Inertia weight (w) | 0.8 |
| Cognitive coefficient (C1) | 2.05 |
| Social coefficient (C2) | 2.05 |

## 6. SIMULATION RESULTS

**Table 2:** Proportional plus Integral controller gains for 0.01p.u. MW step-load change in Area-1 and Area-2

| **Power System** | **Gain values** | |
|---|---|---|
| | $K_P$ | $K_I$ |
| Conventional in Area1&Area 2 | 0.95 | 0.30 |
| With SMES in Area-1 | 0.52 | 0.26 |
| With Gas Turbine in Area-2 | 0.65 | 0.24 |

**CASE 1:** Comparison of frequency deviations, control input requirements and tie-line power deviations in a two-area interconnected thermal reheat power system for 0.01p.u. MW load change in area-1.



Time in seconds (a)

Time in seconds (b)

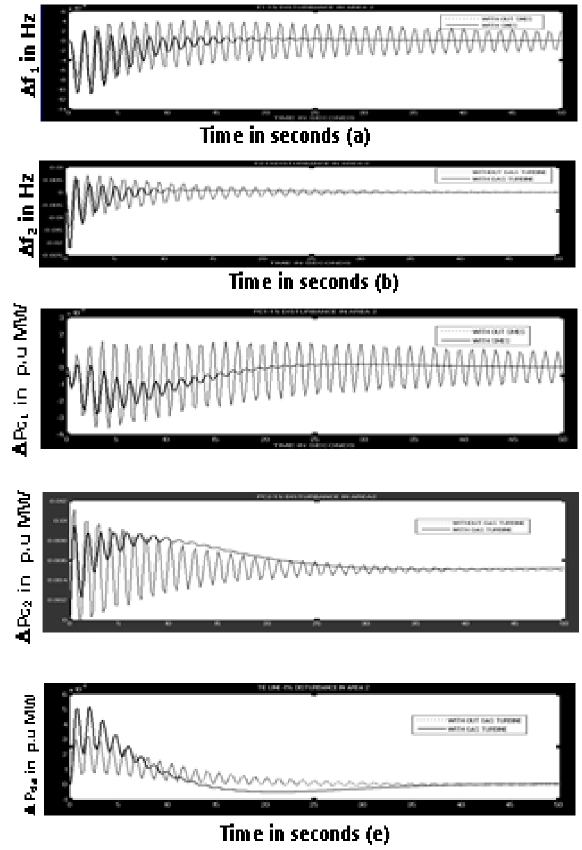
Time in seconds (c)

Time in seconds (d)

Time in seconds (e)

(a) Frequency deviation without and with SMES
(b) Frequency deviation without and with GT
(c) Control input requirement without and with SMES
(d) Control input requirement without and with GT
(e) Tie-line power deviations without and with SMES

**CASE 2:** Comparison of frequency deviations, control input requirements and tie-line power deviations in a two-area interconnected thermal reheat power system for 0.01p.u. MW load change in area-2.



Time in seconds (a)

Time in seconds (b)

Time in seconds (e)

(a) Frequency deviation without and with SMES
(b) Frequency deviation without and with GT
(c) Control input requirement without and with SMES
(d) Control input requirement without and with GT
(e) Tie-line power deviations without and with GT

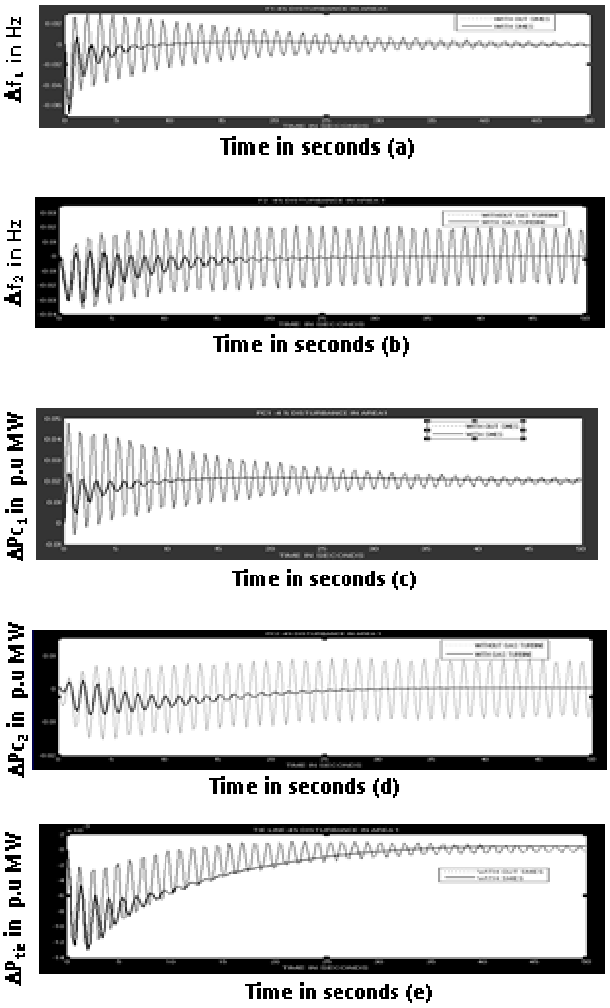
| Area1 and Area2 | Change in frequency H.Z | $\Delta$Ptie$_1$ |
|---|---|---|
| Without SMES | More than 50 sec($\Delta F_1$) | More than 50 sec |
| With SMES | 8 sec($\Delta F_1$) | 12 sec |
| Without Gas turbine | More than 50 sec($\Delta F_2$) | More than 50 sec |
| With Gas turbine | 15 sec($\Delta F_2$) | 25 sec |

**Table 3:** Settling Time (in seconds) of the output response without and with SMES and GT For 0.01p.u. MW step load change in area-1 and area-2 respectively.

**Table 4:** Proportional plus Integral controller gains for 0.04p.u.MW step load change in Area-1 and Area-2

| Power system | Gain values | |
|---|---|---|
| | $K_P$ | $K_I$ |
| Conventional in Area1&Area 2 | 0.98 | 0.22 |
| With SMES in Area-1 | 0.64 | 0.15 |
| With Gas turbine in Area-2 | 0.7 | 0.26 |

**CASE 3:** Comparison of frequency deviations, control input requirements and tie-line power deviations in a two-area interconnected thermal reheat power system for 0.04p.u. MW load change in area-1.



Time in seconds (a)

Time in seconds (b)

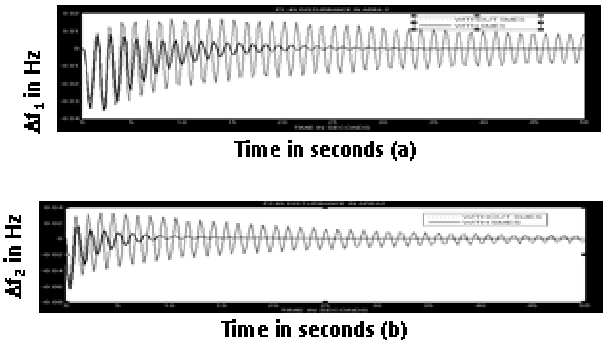
Time in seconds (c)

Time in seconds (d)

Time in seconds (e)

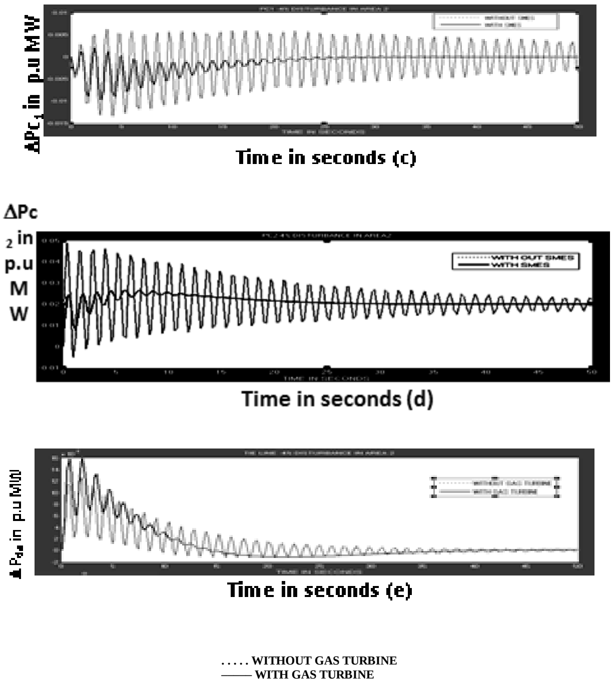
(a) Frequency deviation without and with SMES
(b) Frequency deviation without and with GT
(c) Control input requirement without and with SMES
(d) Control input requirement without and with GT
(e) Tie-line power deviations without and with SMES

**CASE 4 :** Comparison of frequency deviations, control input requirements and tie-line power deviations in a two-area Interconnected thermal reheat power system for 0.04p.u. MW load change in area-2.



Time in seconds (a)

Time in seconds (b)



(a) Frequency deviation without and with SMES
(b) Frequency deviation without and with GT
(c) Control input requirement without and with SMES
(d) Control input requirement without and with GT
(e) Tie-line power deviations without and with GT

| Area1 and Area 2 | Change in frequency H.Z | $\Delta Ptie_1$ |
|---|---|---|
| Without SMES | More than 50 sec($\Delta F_1$) | More than 50 sec |
| With SMES | 9 sec($\Delta F_1$) | 18sec |
| Without SMES | More than 50 sec($\Delta F_2$) | More than 50 sec |
| With SMES | 10 sec($\Delta F_2$) | 18 sec |

**5Table 5:** Settling Time (in seconds) of the output response without and with SMES and Gas turbine. For 0.01p.u. MW step load change in area-1 and area-2 respectively.

## CONCLUSION

In this paper a PSO controller by considering SMES and GT units as spinning reserves for power system restoration problem has been designed. It is clear from the simulation results that the transient responses of frequency and tie-line, better system control has been achieved using the spinning reserves along with the global optimization controller for restoring the system. There is improvement of the dynamic performance of TATURIPS attaining the restoration of the system under consideration in a short duration of time. Hence, the proposed controller yields good transient response with a minimum settling time and further, this work is being extended with the consideration of the system non-linearity.

## REFERENCES

1. Unbehauen, H., Kocaarslan, I., (June 1990), *Experimental Modelling and Simulation of a Power Plant*, Proceedings of European Simulation Multi Conference, Nürnberg, Germany, pp. 474 - 478, 10-13.
2. Chang, C.S., Weihui Fu, (1997), *Area load frequency control using fuzzy gain scheduling of PI controllers*, Electric Power systems Research, 42, pp. 145-152.
3. Kumar, A. O.P. Malik, G.S. Hope, (Jan. 1985), *Variablestructure-system control applied to AGC of an interconnected power system*, IEE Proceedings, Vol. 132, Pt. C, No. 1, pp. 23-29.
4. Shayegi, H., H.A. Shayanfar and O.P. Malik, (2007), Robust decentralized neural networks based LFC in a deregulated power system. Elec. Power Syst. Res., 77: 241-251.
5. Shayeghi, H. *et al*., (2007), Robust modified GA based multi-stage fuzzy LFC. Energy Conversion and Manag., 48: 1656-1670.
6. Lim, K.Y. *et al*., (1996), Robust decentralized load frequency control of multi-area power system. IEE Proceedings-C, 143 (5): 377-386.
7. Wang, Y., D.J. Hill and G. Guo, (1998), Robust decentralized control for multi-machine power system, IEEE Trans. on circuits and systems: Fund. Theory and Applications., Vol. 45, No. 3.
8. ChaoOu ,Weixing Lin,June (2006), “Comparison between PSO and GA for Parameters Optimization of PID Controller”, Proceedings of the IEEE International Conference on Mechatronics and Automation, pp.2471-2475.
9. Aldeen, M. and J.F. Marah, (1991). Decentralized PI design method for inter-connected power systems. IEE Proc.-C, Vol. 138, No. 4.
10. Adibi M.M., R.J Kafka, D.P Milanic, (Aug 1994), “Expert system requirements for power system Restoration”, IEEE Transactions on Power Systems, Vol.9, pp.1592-1598.
11. Chidambaram, I.A. and S.Velusami, (2005), “Design of decentralized biased controllers for load-frequency control of interconnected power systems”, Electric Power Components and Systems, Vol. 33, No.12, pp.1313-1331.
12. Soon Klat Yee, Jovica, F.Michael Hughcs, (Feb. 2008), “Over view and comparative Analysis of gas turbine Model for system stability studies”, IEEE Transactions on Power Systems, Vol.23, pp.108-118.
13. Barsali, S., D.Polio, A.Pratico, R.Salvati, M.Sforna, (Aug. 2008), “Restoration islands supplied by gas turbine”, Electrical Power System Research, Vol. 78, pp.2004-2010.
14. Nagpal M., A.Moshref, G.K Morision, P.Kundur, (Jan 2001) “Experience with Testing and modeling of gas turbine”, Proceedings of the IEEE/PES 2001 Winter Meeting, Columbus USA, pp. 652-656.
15. Demiroren, A. (2002) “Application of a self-tuning to power system with SMES”, European Transactions

on Electrical Power (ETEP), vol. 12, N0.2, pp. 101-109.

16. Tripathy, S.C., R. Balasubramanian, P.S. Chandramohanan Nair, (1997) Adaptive automatic generation control with superconducting magnetic energy storage in power systems, IEEE Trans. Energy Convers.7 (3) 434-441.
17. Kennedy, J. and R.C. Eberhart, (1995), Particle swarm optimization. In: Proc. IEEE Int. Conf. on Neural Network, Perth, Australia, pp: 1942-1948.
18. Ghoshal, S.P., (June 2004),”Optimizations of PID gains by particle swarm optimizations in fuzzy based automatic generation control”, Electrical power system Research, Vol.2, pp.203-212.

## Appendix

### Data for TATURIPS [11]

$Pr_1$=$Pr_2$=2000MW
$Kp_1$=$Kp_2$=120Hz/p.u
$Tp_1$=$Tp_2$=20sec.
$Tp_1$=$Tt_2$=0.3 sec.
$Tg_1$=$Tg_2$=0.08sec.
$Kr_1$=$Kr_2$=0.5
$Tr_1$=$Tr_2$=10 sec.
$R_1$=$R_2$=2.4Hz/p.u MW.
$a_{12}$=-1
$T_{12}$=0.545 p.u MW/Hz
$\beta_1 = \beta_2$ =0.425 p.u. MW/Hz

### Data for the SMES unit [15]

L=2.65H
$T_{dc}$=0.03 sec
$I_{do}$=4.5KA
$K_{id}$=0.2 KV/KA
$K_{SMES}$=100KV/unit MW

### Data for the Gas turbine model [12]

$T_1$=10 sec
$T_2$=0.1sec
$T_3$= 3sec
$K_t$=1
$K_r$=0.04
Dturb=0.03
Maximum and minimum valve position = 1and -0.1

## Biographical Notes

**R. Jayanthi** received her B.E and M.E degrees from Faculty of Engineering and Technology, Annamalai University, Annamalai Nagar, Chidambaram, India in 1994 and 2007 respectively. Currently working as an Assistant Professor in the Department of Electrical Engineering, Annamalai University, Annamalai Nagar, since 2007. She is currently working towards the Ph.D. degree. Her research interest includes power system operation and control.

**Dr. I.A.Chidambaram** (1966) received Bachelor of Engineering in Electrical and Electronics Engineering (1987) Master of Engineering in Power System Engineering (1992) and Ph.D. in Electrical Engineering (2007) from Annamalai University, Annamalainagar.

During 1988 - 1993 he was working as Lecturer in the Department of Electrical Engineering, Annamalai University and from 2007 he is working as Professor in the Department of Electrical Engineering, Annamalai University, and Annamalainagar. He is a member of ISTE and Indian Science Congress (ISC). His research interests are in power systems, electrical measurements and control systems. (Electrical Measurements Laboratory, Department of Electrical Engineering, Annamalai University, Annamalainagar 608002, Tamilnadu, India, Tel:-91-04144-238501,Fax:-91-04144-238275) driacdm@yahoo.com/ driacdm@gmail.com

# The Comparative study and Performance of HCM and MFPCM Algorithms on Iris data set

VUDA SREENIVASARAO

Professor & Head CSE, IT Dept. St .Mary's College of Engg. & Technology, Hyderabad ,India.

**Abstract:** Data mining technology has emerged as a means for identifying patterns and trends from large quantities of data. Data mining is a computational intelligence discipline that contributes tools for data analysis, discovery of new knowledge, and autonomous decision making. Clustering is a primary data description method in data mining which group's most similar data. The data clustering is an important problem in a wide variety of fields. Including data mining, pattern recognition, and bioinformatics. It aims to organize a collection of data items into clusters, such that items within a cluster are more similar to each other than they are items in the other clusters. There are various algorithms used to solve this problem In this paper, we use HCM (Hard C -mean) clustering algorithm and MFPCM (Modified Fuzzy Possibilistic C - mean) clustering algorithm. In this paper we compare the performance analysis of Hard C mean (HCM) clustering algorithm and compare it with Modified Fuzzy possibilistic C mean algorithm. In this we compared HCM and MFPCM algorithm on different data sets. We measure complexity of HCM and MFPCM at different data sets. HCM clustering is a clustering technique which is separated from Modified Fuzzy Possibililstic C mean that employs Possibililstic partitioning.

**Keywords**: Data clustering Algorithm, Portioning, Data Mining, Hard C Mean, Modified Fuzzy Possibililstic C mean.

## 1. Introduction:

Data analysis is considered as a very important science in the real world. Data mining technology has emerged as a means for identifying patterns and trends from large quantities of data. Data mining is a computational intelligence discipline that contributes tools for data analysis, discovery of new knowledge, and autonomous decision making. The task of processing large volume of data has accelerated the interest in this field. As mentioned in Mosley (2005) data mining is the analysis of observational datasets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. Data mining discovers description through clustering visualization, association, sequential analysis. Clustering is a primary data description method in data mining which group's most similar data. Data clustering is a common technique for data analysis, which is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. Cluster analysis is a technique for classifying data; it is a method for finding clusters of a data set with most similarity in the same cluster and most dissimilarity between different clusters. The conventional clustering methods put each point of the data set to exactly one cluster. Since 1965, Zadeh proposed fuzzy sets in order to come closer of the physical world. Zadeh introduced the idea of partial memberships described by membership functions. Clustering algorithm partitions an unlabelled set of data into groups according to the similarity. Compared with the data classification, the data clustering is an unsupervised learning process, it does not need a labeled data set as training data, but the performance of the data clustering algorithm is often much poorer. Although the data classification has better performance, it needs a labeled data set as training data and labeled data for the classification is often very difficult and expensive to obtain. So there are many algorithms are proposed to improve the clustering performance. *Clustering* is the classification of similar objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset share some common trait.

Clustering technique is used for combining observed objects into clusters (groups), which satisfy two main criteria:

- Each group or cluster should be homogeneous objects that belong to the same group are similar to each other.
- Each group of cluster should be different from other clusters, that is, objects that belong to one cluster should be different from the objects of other clusters.

Clustering can be considered the most important unsupervised learning problem. So, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be the process of organizing objects into groups whose members are similar in some way. A cluster is therefore a collection of objects, which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. There are many clustering methods available, and each of them may give a different grouping of a dataset. The choice of a particular method will depend on the type of output desired, the known performance of method with particular types of data, the hardware and software facilities available and the size of the dataset.

## 2. Hard C- Mean clustering algorithm :

In non fuzzy or hard clustering, data is divided into crisp clusters, where each data point belongs to exactly one cluster.

- Used to classify data in crisp set
- Each data point will be assigned to only one cluster
- Clusters are also known as partitions
- U is a matrix with c rows and n columns
- The cardinality gives number of unique c partitions for n data points

In this clustering technique partial membership is not allowed. HCM is used to classify data in a crisp sense. By this we mean that each data point will be assigned to one and only one data cluster. In this sense, these clusters are also called as partitions that are partitions of the data. In case of hard c mean each data element can be a member of one and only one cluster at a time. In other words we can say that the sum of membership grades of each data point in all clusters is equal to one and in HCM membership grade of a specific data point in a specific cluster is one and in all the remaining clusters its membership grade is zero. Also number of clusters that is can't be less than or equal to one and they can't be equal to or greater than number of data elements because if number of clusters is equal to one than all data elements will lie-in same cluster and if number of clusters is equal to number of data elements than each data elements will lie in its own separate cluster. That is each cluster is having only one data point in this special case. The steps of HCM algorithm given below.

1. fix c(2<=c<n) and initialize the U matrix

$$U^{(0)} \in M_C$$

Then for r=0, 1, 2, 3……………

2. Calculate the center vectors{ $V^{®}$ with $U^{®}$ }

3. Update $U^{®}$ calculate the updated characteristic function(for a all i,k).

$$X_{ik}^{(r+1)} = \begin{cases} 1, d_i^{(r)} = \min d_{jk}^{(r)} \; forallj\in c \\ 0, otherwise \end{cases}$$

4. if $||U^{(0r-1)}-U^{®}||<=\delta$(tolerance level)

STOP: otherwise set r=r+1 and return to step 2.In step 4 the notation || || is any matrix norm such as the Euclidean norm.

## 3. Modified Fuzzy Possibililstic C - Mean Algorithm :

The FPCM algorithm attempts to partition a finite collection of elements X={x1, x2, x3………xn} into a collection of c fuzzy clusters with respect to some given criterion. Given a finite set of data, the algorithm returns a list of c cluster centers V, such that V=vi, i=1,2,3……………,c And a partition matrix U such that U=uij,i=1,2,3,……………c, j=1,2,……………n Where uij is a numerical value in [0, 1] that tells the degree to which the elements xj belongs to the i-th cluster. Defines a family of fuzzy sets {Ai, i=1,2,3……..c} as a fuzzy c partition on a universe of data points X

1. Fuzzy set allows for degree of membership

2. A single point can have partial membership in more than one class.

3.There can be no empty classes and no class that contains no data points

The steps of Modified Fuzzy Possibililstic C - Mean Algorithm given below:

1. the objective function of the Modified Fuzzy Possibililstic C - Mean Algorithm can be formulated as follows:

$$J_{MFPCM} = \sum_{i-1}^{c} \sum_{j-1}^{n} (\mu_{ij}^{m} w_{ji}{}^{m} d^{2m}(x_j, v) + t_{ij}^{\eta} w_{ji}{}^{\eta} d^{2\eta}(x_j, v_i))$$

2. Calculate U = {μ $_{ij}$} represents a fuzzy partition matrix, is defined as:

$$u_{ij} = \left[ \sum_{k-1}^{c} \left( \frac{d?\mathbf{x}_{\mathbf{j}}, v_{\mathbf{i}})}{d?\mathbf{x}_{\mathbf{j}}, v_{\mathbf{k}})} \right)^{2m/(m-1)} \right]^{-1}$$

3. Calculate T = {t $_{ij}$} represents a typical partition matrix, is defined as :

$$t_{ij} = \left[ \sum_{k-1}^{n} \left( \frac{d?\mathbf{x}_{\mathbf{j}}, v_{\mathbf{i}})}{d?\mathbf{x}_{\mathbf{j}}, v_{\mathbf{k}})} \right)^{2\eta/(\eta-1)} \right]^{-1}$$

4. Calculate V = {v $_{ij}$} represents c centers of the clusters, is defined as:

$$v_i = \frac{\sum_{j-1}^{n} \left( \mu_{ij}^{m} w_{ji}^{m} + t_{ij}^{\eta} w_{ji}^{\eta} \right)^{*} x_j}{\sum_{j-1}^{n} \left( \mu_{ij}^{m} w_{ji}^{m} + t_{ij}^{\eta} w_{ji}^{\eta} \right)}$$

## 4. Results:

### 4.1 Time complexity of HCM and MFPCM by varying no. of Clusters on Iris Data:

The implementation of HCM & MFPCM is done on iris Data in MATLAB. The data t contains 3 classes of 150 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other two, the latter are NOT linearly separable from each other. The data set contain four attribute which are given below

The time complexity of HCM [11] is O(ndc$^2$i) and time complexity of MFPCCM is O(ncdi). Now keeping no. of data points constant, lets assume n=100, d=3, i=20 and varying no. of clusters, we obtain the following table and graph. Where n= number of data point, c= number of cluster, d= dimension, i= number of iteration

**Table 4.1 Time Complexity when Number of cluster varying**

| S.No. | Number of Cluster | HCM Time Complexity | MFPCM Time Complexity |
|---|---|---|---|
| 1 | 1 | 1000 | 2000 |
| 2 | 2 | 9000 | 5000 |
| 3 | 3 | 24000 | 8500 |
| 4 | 4 | 47000 | 10500 |

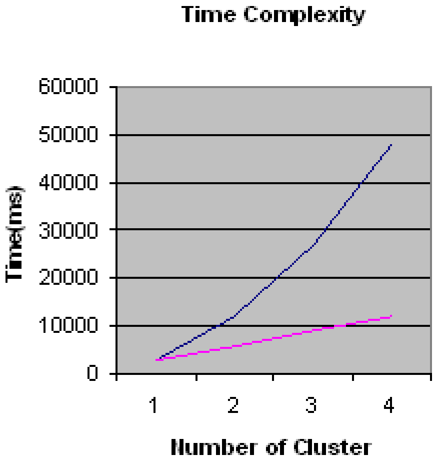


**Figure 4.1 Time complexity of HCM and MFPCM by varying no. of Clusters**

Now keeping no. of cluster constant, lets assume n=140, d=3, c=3 and varying no. of Iteration, we obtain the following table and graph.

**Table4.2 Time Complexity when Number of Iterations varying**

| S.No. | Number of Iteration | HCM Time Complexity | MFPCM Time Complexity |
|---|---|---|---|
| 1 | 5 | 6000 | 3000 |
| 2 | 10 | 10000 | 6000 |
| 3 | 15 | 16000 | 8000 |
| 4 | 22 | 25000 | 12000 |

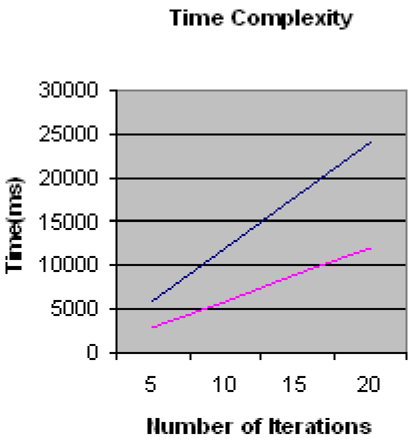


**Figure 4.2 Time complexity of HCM and MFPCM by varying no. of Iterations**

### 4.2. Comparison of space complexity of HCM and MFPCM :

The space complexity of HCM is O(nd+nc) and MFPCM is O(cd). Now keeping no. of data points constant, lets assume n=140, d=3 and varying no. of clusters we obtain the following graph.

**Table4.3 Space Complexity when Number of Clusters varying**

| S.No. | Number of Cluster | HCM Space Complexity | MFPCM Space Complexity |
|---|---|---|---|
| 1 | 10 | 500 | 4 |
| 2 | 15 | 700 | 8 |
| 3 | 20 | 900 | 12 |
| 4 | 25 | 1100 | 16 |

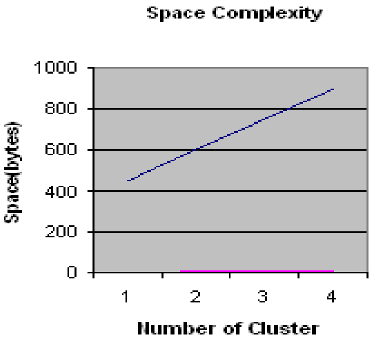


**Figure4.3 space complexities of HCM and MFPCM by varying number of clusters**

### 4.3 Complexity Analysis of HCM Algorithm :

The asymptotic efficiency of the algorithm has following notations:

i number HCM over entire dataset.

n number of data points.

c number of clusters

d number of dimensions

The time complexity of the Hard c mean algorithm is O(ndc$^2$i), where empirically I grows very slowly with n,c and d.

The memory complexity of HCM is O(nd + nc), where nf is the size of data set and nc the size of U matrix.

For data sets, which cannot be loaded into memory, HCM will have disk accesses every iteration. Thus the disk input output complexity will be O(ndi) It is likely that for those data sets the U matrix cannot be kept in memory too. Thus, it will increase the disk input/output complexity further

### 4.4 Complexity Analysis of MFPCM Algorithm

The asymptotic efficiency of the algorithm has following notations:

i number of k means passes over entire dataset.

n number of data points.

c number of clusters

d number of dimensions

The time complexity of the hard c mean algorithm is O (ncdi), where empirically I grows very slowly with n, c and d.

The memory complexity of MFPCM is cd

I/O complexity of MFPCM is ndi

**Table 4.4. Comparative Analysis of Complexities of HCM and MFPCM**

| Algorithm | Time complexity | Space complexity | I/O complexity |
|---|---|---|---|
| HCM | O(ndc$^2$i), | O(nd + nc) | O(ndi) |
| MFPCM | O(ncdi), | cd | ndi |

### 5. Conclusion:

In partitioning based clustering algorithms, the number of final cluster (k) needs to be defined beforehand. Also, algorithms have problems like susceptible to local optima, sensitive to outliers, memory space and unknown number of iteration steps required to cluster. The time complexity of the MFPCM is O(ncdi) The memory complexity of MFPCM is cd and the input output complexity will be O(ndi). Fuzzy clustering, which constitute the oldest component of soft computing, are suitable for handling the issues related to understandability of patterns, incomplete/noisy data, mixed media information and human interaction, and can provide approximate solutions faster. They have been mainly used in discovering association rules and functional dependencies and image retrieval. The time complexity of the Hard C Mean algorithm is O(ndc$^2$i). The memory complexity of MFPCM is O(nd + nc),and the disk input output complexity will be O(ndi)

## 6. References:

[1] ude hemanth.D, D.Selvathi and J.Anitha,"Effective Fuzzy Clustering Algorithm for Abnormal MR Brain Image Segmentation",Page umber 609-614, International/Advance Computing Conference (IACC 2009),IEEE,2009.

[2 ] Sorin Istrail, "An Overview of Clustering Methods", With Applications to Bioinformatics.

[3 ] Wei Wang, Chunheng Wang, Xia Cui, Ai Wang, "*A Clustering Algorithm Combine the FCM algorithm with Supervised Learning Normal Mixture Model*"**,** IEEE 2008.

[4] Deepak Agrawal "*Web Data Clustering using FCM and Proximity Hints from Text as well as Hyperlink-structure*"**,** IEEE 2008.

[5] M. Brej and M. Sonka, "*Object localization and border detection criteria design in edge-based image segmentation automated learning from examples*"**,** IEEE Transactions on Medical imaging, vol. 19, pp. 973-985, 2000.

[6] S. Chen and D. Zhang, *"Robust image segmentation using FCM with spatial constraints based on new kernel-induced distance measure"*, IEEE Transactions on Systems, Man and Cybernetics, vol. 34, pp. 1907-1916, 1998.

[7] O. Sojodishijani, V. Rostami and A. R. Ramli, "*Real time color image segmentation with non-symmetric Gaussian membership functions",* Fifth International Conference on Computer Graphics, Imaging and Visualization, pp. 165-170, 2008.

[8] M. S. Yanp, K.L. Wu and J. Yub, *"A novel fuzzy clustering algorithm",* IEEE International Symposium on Computational Intelligence in Robotics and Automation, vol. 2, pp. 647- 652, 2003.

[9] L. Hui, *"Method of image segmentation on high-resolution image and classification for land covers",* Fourth International Conference on Natural Computation, vol. 5, pp. 563-566, 2008.

[10] D. L. Pham, *"Spatial models for fuzzy clustering",* Laboratory of Personality and Cognition, Gerontology Research Center, 2001.

[11] R. J. Almeida and J. M. C. Sousa**,** *"Comparison of fuzzy clustering algorithms for Classification",* International Symposium on Evolving Fuzzy Systems, pp. 112-117, 2006.

[12] Mohamed Fadhel Saad and Adel M.Alimi " Modified Fuzzy Poossibilistic C-means" ,International multi conference of Engineers and computer scientists -2009 Vol1

[13] Vuda sreenivasarao and Dr.S.Vidyavathi**,** "Comparative investigations and performance analysis of FCM and MFPCM algorithms on IRIS data", India Journal of computer science and engineering, Volume 1, Issue 2, July 2010, pp 145-151.

[14] Vuda sreenivasarao and Dr.S.Vidyavathi, "Comparative analysis of Fuzzy C-Mean and modified fuzzy possibilistic C-Mean algorithms in Data mining", International Journal of Computer Science and Technology , Volume 1, Issue no 1, Sep 2010, pp 100-102.

## 7. Author profile:

**VUDA SREENIVASARAO** received his M.Tech degree in Computer Science & Engg from the Satyabama University, in 2007.Currently working as Professor & Head in the Department of CSE & IT at St.Mary's college of Engineering & Technology, Hyderabad, India.. He is currently pursuing the PhD degree in CSE Depart in Singhania University, Rajasthan. His main research interests are Data Mining, Network Security, and Artificial Intelligence. He has got 10years of teaching experience .He has published 21 research papers in various international journals. He is a life member of various professional societies like MIACSIT, MISTE. MIAENG.

# LMI Approach for Stability of Neural Networks

K. Ratchagit

Department of Mathematics
Faculty of Science, Maejo University
Chiang Mai 50290, Thailand
e-mail: kreangkri@mju.ac.th

***Abstract***: In this paper, we derive a sufficient condition for asymptotic stability of the zero solution of delay-difference control system of Hopfield neural networks in terms of certain matrix inequalities by using a discrete version of the Lyapunov second method.

***Keywords***: Asymptotic stability, Hopfield neural networks, lyapunov function, delay-difference control system, matrix inequalities.

## 1. Introduction

In recent decades, Hopfield neural networks have been extensively studied in many aspects and successfully applied to many fields such as pattern identifying, voice recognizing, system controlling, signal processing systems, static image treatment, and solving nonlinear algebraic system, etc. Such applications are based on the existence of equilibrium points, and qualitative properties of systems. In electronic implementation, time delays occur due to some reasons such as circuit integration, switching delays of the amplifiers and communication delays, etc. Therefore, the study of the asymptotic stability of Hopfield neural networks with delays is of particular importance to manufacturing high quality microelectronic Hopfield neural networks.

While stability analysis of continuous-time neural networks can employ the stability theory of differential system (Liu *et al.* 2003), it is much harder to study the stability of discrete-time neural networks (Elaydi and Peterson 1990) with time delays (Arik 2005) or impulses (Gubta and Jin 1996). The techniques currently available in the literature for discrete-time systems are mostly based on the construction Lyapunov second method (Hale 1977). For Lyapunov second method, it is well known that no general rule exists to guide the construction of a proper Lyapunov function for a given system. In fact, the construction of the Lyapunov function becomes a very difficult task.

In this paper, we consider delay-difference control system of Hopfield neural networks of the form

$$v(k+1) = -Av(k) + BS(v(k-h)) + Cu(k) + f\,, \qquad (1)$$

where $v(k) \in \Omega \subseteq \mathbf{R}^n$ is the neuron state vector, $h \geq 0,$ $A = diag\{a_1, \ldots, a_n\}$, $a_i \geq 0$, $i = 1, 2, ..., n$ is the $n \times n$ constant relaxation matrix, $B$ is the $n \times n$ constant weight matrix, $C$ is $n \times m$ constant matrix, $u(k) \in \mathbf{R}^m$ is the control vector, $f = (f_1, \ldots, f_n) \in \mathbf{R}^n$ is the constant external input vector and $S(z) = [s_1(z_1), \ldots, s_n(z_n)]^T$ with $s_i \in C^1\left[\mathbf{R}, (-1,1)\right]$ where $s_i$ is the neuron activations and monotonically increasing for each $i = 1, 2, ..., n$.

The asymptotic stability of the zero solution of the delay-differential system of Hopfield neural networks has been developed during the past several years. We refer to monographs by Burton (Burton 1993) and Ye (Ye 1944) and the references cited therein. Much less is known regarding the asymptotic stability of the zero solution of the delay-difference control system of Hopfield neural networks. Therefore, the purpose of this paper is to establish sufficient condition for the asymptotic stability of the zero solution of equation (1) in terms of certain matrix inequalities.

## 2. Preliminaries

The following notations will be used throughout the paper. $\mathbf{R}^+$ denotes the set of all non-negative real numbers; $\mathbf{Z}^+$ denotes the set of all non-negative integers; $\mathbf{R}^n$ denotes the n-finite-dimensional Euclidean space with the Euclidean norm $\|.\|$ and the scalar product between $x$ and $y$ is defined by $x^T y$; $\mathbf{R}^{n \times m}$ denotes the set of all $(n \times m)$-matrices; and $A^T$ denotes the transpose of the matrix $A$; Matrix $Q \in \mathbf{R}^{n \times n}$ is positive semidefinite $(Q \geq 0)$ if $x^T Qx \geq 0,$ for all $x \in \mathbf{R}^n$. If $x^T Qx > 0 (x^T Qx < 0,$ resp.) for any $x \neq 0$, then $Q$ is positive (negative, resp.) definite and denoted by $Q > 0, (Q < 0,$ resp.). It is easy to verify that $Q > 0,$ $(Q < 0,$ resp.) iff $\exists \beta > 0:$
$x^T Qx \geq \beta \|x\|^2, \forall x \in \mathbf{R}^n,$
$(\exists \beta > 0: x^T Qx \leq -\beta \|x\|^2, \forall x \in \mathbf{R}^n,$ resp.) .

**Lemma 2.1** (Hale 1977) The zero solution of difference system is asymptotic stability if there exists a positive definite function $V(x): \mathbf{R}^n \to \mathbf{R}^+$ such that

$$\exists \beta > 0: \Delta V(x(k)) = V(x(k+1)) - V(x(k)) \leq -\beta \|x(k)\|^2,$$

along the solution of the system. In case the above condition holds for all $x(k) \in V_\delta$, we say that the zero solution is locally asymptotically stable.

**Lemma 2.2** For any constant symmetric matrix $M \in \mathbf{R}^{n\times n}$, $M = M^T > 0$, scalar $s \in \mathbf{Z}^+ /\{0\}$, vector function $W : [0, s] \to \mathbf{R}^n$, we have

$$s\sum_{i=0}^{s-1}(w^T(i)Mw(i)) \geq \left(\sum_{i=0}^{s-1} w(i)\right)^T M \left(\sum_{i=0}^{s-1} w(i)\right).$$

We present the following technical lemmas, which will be used in the proof of our main result.

## 3. Main results

In this section, we consider the sufficient condition for asymptotic stability of the zero solution $v^*$ of (1) in terms of certain matrix inequalities. Without loss of generality, we can assume that $v^* = 0, S(0) = 0$ and $f$ =0 (for otherwise, we let $x = v - v^*$ and define

$$S(x) = S(x + v^*) - S(v^*)) .$$

The new form of equation (1) is now given by

$$x(k+1) = -Ax(k) + BS(x(k-h)) + Cu(k) . \qquad (2)$$

This is a basic requirement for controller design. Now, we are interested designing a feedback controller for the system equation (2) as

$$u(k) = Kx(k),$$

where $K$ is $n\times m$ constant control gain matrix.

The new form of (2) is now given by

$$x(k+1) = -Ax(k) + BS(x(k-h)) + CKx(k) . \qquad (3)$$

Throughout this paper we assume the neuron activations $s_i(x_i)$, $i = 1, 2, \ldots, n$ is bounded and monotonically nondecreasing on $\mathbf{R}$, and $s_i(x_i)$ is Lipschitz continuous, that is, there exist constant $l_i > 0, i = 1, 2, \ldots, n$ such that

$$|s_i(r_1) - s_i(r_2)| \leq l_i\,|r_1 - r_2|,\ \forall r_1, r_2 \in \mathbf{R}. \qquad (4)$$

By condition equation (4), $s_i(x_i)$ satisfy

$$|s_i(x_i)| \leq l_i\,|x_i|,\ i = 1, 2, ..., n . \qquad (5)$$

**Theorem 3.1** The zero solution of the delay-difference control system (3) is asymptotically stable if there exist symmetric positive definite matrices $P, G, W$ and $L = diag[l_1, \ldots, l_n] > 0$ satisfying the following matrix inequalities of the form

$$\psi = \begin{pmatrix} (1,1) & 0 & 0 \\ 0 & (2,2) & 0 \\ 0 & 0 & (3,3) \end{pmatrix} < 0, \qquad (6)$$

where

$$(1,1) = hG + W,$$

$$(2,2) = -W,$$

$$(3,3) = -hG.$$

**Proof** Consider the Lyapunov function

$V(y(k)) = V_1(y(k)) + V_2(y(k))$, where

$$V_1(y(k)) = \sum_{i=k-h}^{k-1} (h - k + i)x^T(i)Gx(i),$$

$$V_2(y(k)) = \sum_{i=k-h}^{k-1} x^T(i)Wx(i),$$

$G$ and $W$ being symmetric positive definite solutions of (6) and $y(k) = [x(k), x(k-h)]$.

Then difference of $V(y(k))$ along trajectory of solution of (3) is given by

$$\Delta V(y(k)) = \Delta V_1(y(k)) + \Delta V_2(y(k)),$$

where

$$\begin{aligned}\Delta V_1(y(k)) &= \Delta\left(\sum_{i=k-h}^{k-1} (h-k+i)x^T(i)Gx(i)\right)\\ &= hx^T(k)Gx(k) - \sum_{i=k-h+1}^{k} x^T(i)Gx(i),\\ \Delta V_2(y(k)) &= \Delta\left(\sum_{i=k-h}^{k-1} x^T(i)Wx(i)\right)\\ &= x^T(k)Wx(k) - x^T(k-h)Wx(k-h),\end{aligned} \qquad (7)$$

Then we have

$$\Delta V \leq x^T(k)[hG + W]x(k) - x^T(k-h)Wx(k-h) - \sum_{i=k-h}^{k-1} x^T(i)Gx(i).$$

Using Lemma 2.2, we obtain

$$\sum_{i=k-h}^{k-1} x^T(i)Gx(i) \geq \left(\frac{1}{h}\sum_{i=k-h}^{k-1} x(i)\right)^T (hG)\left(\frac{1}{h}\sum_{i=k-h}^{k-1} x(i)\right).$$

From the above inequality it follows that:

$$\Delta V \le x^T(k)[hG+W]x(k) - x^T(k-h)Wx(k-h)$$
$$-\left(\frac{1}{h}\sum_{i=k-h}^{k-1} x(i)\right)^T (hG)\left(\frac{1}{h}\sum_{i=k-h}^{k-1} x(i)\right)$$

$$=\left( x^T(k), x^T(k-h), (\frac{1}{h}\sum_{i=k-h}^{k-1} x(i))^T \right)\begin{pmatrix} (1,1) & 0 & 0 \\ 0 & (2,2) & 0 \\ 0 & 0 & (3,3) \end{pmatrix}$$

$$\times\begin{pmatrix} x(k) \\ x(k-h) \\ (\frac{1}{h}\sum_{i=k-h+1}^{k} x(i)) \end{pmatrix}$$

$$= y^T(k)\psi y(k)$$

where

$$(1,1)=hG+W,$$

$$(2,2)=-W,$$

$$(3,3)=-hG,$$

$$y(k)=\begin{pmatrix} x(k) \\ x(k-h) \\ (\frac{1}{h}\sum_{i=k-h}^{k-1} x(i)) \end{pmatrix}.$$

By the condition (6), $\Delta V(y(k))$ is negative definite, namely there is a number $\beta > 0$ such that $\Delta V(y(k)) \le -\beta\|y(k)\|^2$, and hence, the asymptotic stability of the system immediately follows from Lemma 2.1. This completes the proof.

## 4. Conclusions

In this paper, based on a discrete analog of the Lyapunov second method, we have established a sufficient condition for the asymptotic stability of delay-difference control system of Hopfield neural networks in terms of certain matrix inequalities.

## Acknowledgements

This work was supported by the Thai Research Fund Grant, the Higher Education Commission and Maejo University, Thailand.

## References

[1] K. Abdelwahab and R.B. Guenther, An introduction to numerical methods a MATLAB approach, Chapman and Hall/CRC, New York, 2002.

[2] S. Barnett and R.G. Cameron, Introduction to Mathematical Control Theory, Oxford, Clarendon Press, 1985.

[3] J. Lu, G. Chen, A new chaotic attractor coined, Int. J. Bifurc. Chaos 12 (2002) 659-661.

[4] V.N. Phat, Constrained Control Problems of Discrete Processes, World Scientific Publisher, Singapore-NewJersey-London,1996.

[5] V.N. Phat, Introduction to Mathematical Control Theory, Hanoi National University Publisher, Hanoi, 2001.

[6] VN Phat, J. Jiang, A.V. Savkin and I. Petersen, Robust stabilization of linear uncertain discrete-time systems via a limited communication channel. Systems and Control Letters. 53(2004), 347-360 (SCI)

[7] VN Phat and J. Jiang, Feedback stabilization of nonlinear discrete-time systems via a digital communication channel. Int. J. of Math. and Math. Sci., 1(2005), 43-56.

[8] VN Phat, Robust stability and stabilizability of uncertain linear hybrid systems with state delays. IEEE Trans. on CAS II, 52(2005), 94-98 (SCI)

[9] VN Phat, N.M. Linh and T.D. Phuong, Sufficient conditions for strong stability of non-linear time-varying control systems with state delays. Acta Math. Vietnamica, 30(2005), 69-86.

[10] VN Phat and A.V. Savkin, Robust set-valued state estimation for linear uncertain systems in Hilbert spaces. Nonl. Func. Anal. Appl., 10(2005), 285-298.

[11] VN Phat and S. Pairote, Global stabilization of linear periodically time-varying switched systems via matrix inequalities. J. Control Theory Appl. 1(2006), 26-31.

[12] P. Niamsup and VN Phat, Stability of linear time-varying delay systems and applications to control problems, J. Comput. Appll. Math. 194(2006), 343-356 .

[13] VN Phat, Global stabilization for linear continuous time-varying systems Appl. Math. Comput. 175(2006), 1730-1743.

[14] VN Phat and P. Niamsup, Stabilization of linear non-autonomous systems with norm bounded controls. J. Optim. Theory Appl., 131(2006), 135-149.

[15] S. Pairote and VN Phat, Exponential stability of switched linear systems with time-varying delay, Elect. J. Diff. Equations, 59(2007), 1-10

[16] Q.P. Ha, H. Trinh and VN Phat, Design of Reduced-Order Observers for Global State Feedback Control of Multi-Agent Systems, Int. J. of Aut. Control, N2/3, 1(2007), 165-181.

[17] VN Phat and PT Nam, Exponential stability and stabilization of uncertain linear time-varying systems using parameter dependent Lyapunov function. Int. J. of Control, 80(2007), 1333-1341.

[18] P.T. Nam and VN Phat, Robust exponential stability and stabilization of linear uncertain polytopic time-delay systems. J. Control Theory Appl., 6(2008), 163-170

[19] P. Niamsup, K. Mukdasai and VN Phat, Linear uncertain non-autonomous time-delay systems: Stability and stabilizability via Riccati equations. Elect. J. Diff. Equations., 26(2008), 1-10.

[20] VN Phat, D.Q. Vinh and N. S. Bay, L2−stabilization and H∞ control for linear non-autonomous time-delay systems in Hilbert spaces via Riccati equations, Adv. in Nonl. Var. Ineq., 11(2008), 75-86.

[21] P. Niamsup and K. Mukdasai and VN Phat, Improved exponential stability for time- varying systems with nonlinear delayed perturbations, Appl. Math. Comput., 204(2008), 490-495.

[22] VN Phat and Q.P. Ha, New characterization of stabilizability via Riccati equations for LTV systems. IMA J. Math. Contr. Inform., 25(2008), 419-429.

[23] PT Nam and VN Phat, An improved stability criterion for a class of neutral deferential equations. Appl. Math. Letters, 22(2009), 31-35.

[24] VN Phat, T. Bormat and P. Niamsup, Switching design for exponential stability of a class of nonlinear hybrid time-delay systems, Nonlinear Analysis: Hybrid Systems, 3(2009), 1-10

[25] VN Phat and PT Nam, Robust stabilization of linear systems with delayed state and control, J. Optim. Theory Appl., 140(2009), 287-299.

[26] L.V. Hien, Q.P. Haand VN Phat, Stability and stabilization of switched linear dynamic systems with time delay and uncertainties Appl. Math. Comput, 210(2009), 223-231.

[27] VN Phat and LV Hien, An application of Razumikhin theorem to exponential stability for linear non-autonomous systems with arbitrary time-varying delays, Appl. Math. Letters, 22(2009), 1412-1417.

[28] LV Hien and VN Phat, Exponential stability and stabilization of a class of uncertain linear time-delay systems, J. of the Franklin Institute, 346(2009), 611-625.

[29] LV Hien and VN Phat, Delay feedback control in exponential stabilization of linear time-varying systems with input delay, IMA J. Math. Contr. Inform., 26(2009), 163-177.

[30] LV Hien and VN Phat, Exponential stabilization for a class of hybrid systems with mixed delays in state and control, Nonlinear Analysis: Hybrid Systems, 3(2009), 259-265.

[31] VN Phat, Memoryless H∞ controller design for switched nonlinear systems with mixed time-varying delays, Int. J. of Control, 82(2009), 1889-1898,

[32] VN Phat and Q.P. Ha, H∞ control and exponential stability for a class of nonlinear non-autonomous systems with time-varying delay, J. Optim. Theory Appl., 142(2009), 603-618.

[33] P. Niamsup and VN Phat, H∞ optimal control of LTV systems with time-varying delay via controllability approach, ScienceAsia, 35(2009), 284-289.

# Intelligent Messaging Service in an InfoStation-based University Network

Liam Merwick, Ivan Ganchev, Máirtín O'Droma
Telecommunications Research Centre.
University of Limerick, Ireland

***Abstract:*** This paper shows how a communication infrastructure consisting of mobile devices, InfoStations and an intelligent gateway can be combined to create a messaging service for a campus sized area. It allows for fast and efficient delivery of messages to a group of users through the provision of two-tier address space architecture. A particularly novel part is the creation of an intelligent central message processing agent which decides which device, and in what format, the message should be forwarded based on a user's preferences and the presence (or not) of their registered devices on the network. The benefit of this 'Intelligent Assistant' is the delivery of messages to a user on the device they are most likely to be able to access at any moment in time and thus deliver messages in a timely manner. A system was successfully prototyped which could deliver messages in SMS and email format and was designed so that further message formats could easily be integrated.

***Keywords:*** mobile messaging, InfoStation, Bluetooth, SMS, two-tier address space

## 1. Introduction

In recent years communications technology has advanced considerably and people own multiple communication devices (mobile SMS/MMS, PDA, desktop computer, laptop) and have many means of being contacted (email, phone, Instant Messaging). Often, when trying to get a message to a person, there is no way of knowing the best way of notifying him/her in a timely manner - he/she may be away from their desk or have their phone switched off. When sending a message, some protocols allow for the sender to request to be notified when the message is received or read [Faj98], but the sender is still required to decide which of the recipient's devices/addresses to send the message to (possibly without any knowledge of the recipient's schedule or location). It is of limited benefit to either party if the message is sent to a device that the user has no access to at that time.

The aim of our research was to build a system which allows a person to send a message to another user in the 'best' possible way, i.e., the messaging service can dynamically decide to route that message to the other person based on that person's current location, contact preferences and other criteria (e.g. urgency or price the user is willing to pay). The end result should be that the recipient(s) should get the message in a more timely manner and in a way most suited to them as they will have more control over how and where the messages are received.

The major system components (Fig. 1) are an application that runs on the mobile device (MobileApp), one or more InfoStations through which the mobile devices connect to the messaging service, a central processing application (MessageRedirector) and a web interface which allows a user to update his/her details and addresses (ContactApp).

The system prototype was built upon the service-orientated network architecture described in [Iva06], [Iva07], which is used to deliver lectures, tutorials and tests on a university campus network. However our prototype has implemented software which has a MessageRedirector (cf. Section 4.5) as the central processing component instead of the InfoStation Centre in that architecture.

The remainder of the paper is organised as follows. In Section 2, we introduce some typical use cases of an Intelligent Messaging Service Section 3 describes related work in this field. The architecture of the messaging service is defined in Section 4 and how it is implemented in Section 5. Finally, we describe our conclusions and future work in Section 6.

## 2. Typical Use Cases

The operation of the messaging service can be best described by some typical examples of its use.

### 2.1. Sending a Message

While parking his car at the university car-park, a lecturer receives an urgent message about a meeting rescheduled for the same day. He notices that the meeting will be held in the same time as a lecture. The lecturer urgently needs to broadcast a message notification to the entire student class about canceling/postponing the lecture. The lecturer types and sends the message on his mobile device which is connected to the nearest base station of the messaging service (e.g. deployed at the car park). The messaging service then decides what is the most appropriate, quickest and cheapest way of delivering this message to each student in the class according to his current individual location (and device in possession) specified in his profile. All registered users (lecturers and students) have profiles containing, among other things, information about the best way of forwarding urgent messages to them at any particular moment, e.g. by SMS/MMS, email, fax, voice mail or otherwise.

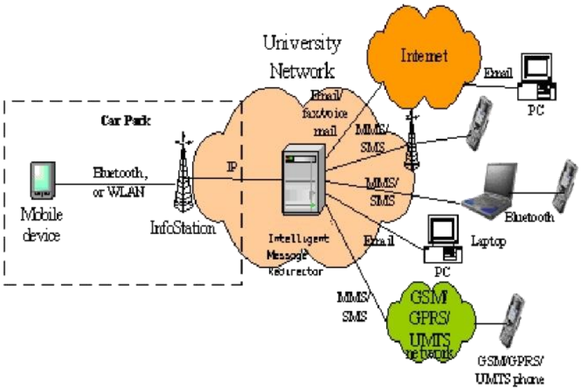


Figure 1: Messaging Service Components

### 2.2. Changing a User's Preferred Contact Address

A student studying in the library or a professor giving a lecture, to avoid distractions, wishes to receive messages via email rather than via SMS. Similarly, a lecturer working at his computer in his office may prefer to handle incoming messages via email. The user logs into the web interface. Following a successful login, the user's details, including a list of addresses, are displayed. The user raises the priority value in the field associated with his preferred email address so that it is the highest priority address. The web interface validates the details before inserting them into the database. The Intelligent Assistant that is part of the Message Redirector will route any messages received to the user's email account until the priorities are re-adjusted.

### 2.3. Adding an address to a profile

The user logs into the web interface which interacts with the database back-end containing the users' details and addresses. Following a successful login, the user's details are displayed including a list of addresses. The user selects the option to add a new address and fills in the requested details. The web interface validates the details before inserting them into the database.

## 3. Related Work

There are numerous software applications available which allow for communications via mobile devices. These applications are provided by both the mobile carriers (e.g. standard applications such as SMS and MMS) as well as open source (e.g. numerous downloadable Java applets) and proprietary (e.g. Nokia Presence) offerings which build upon the platforms provided by the carriers (e.g. Brew, Vodafone Live!) and device manufacturers (e.g. Nokia's Series 40/60).

The research undertaken, and described in this paper, focuses on creating new infrastructure to route messages and where possible leverages existing applications to actually send and receive messages. Many individuals and organisations have put considerable effort into providing tools to send/receive message for the various standard formats (and in many cases multiple high quality applications exist for each message format – each focusing on different usage scenarios). Enabling users to select from pre-existing applications allows users to choose the application that best suits their needs and allows us to focus on extending the capabilities of the messaging service instead of re-implementing the clients.

### 3.1. State Of The Art

#### 3.1.1. A Unified Messaging System

One instance of a unified messaging system is GlobalCom [Bar02], which is a suite of web-based tools that can be used to send messages in various formats. They used a single message-independent format to store the messages and implemented their own clients (such as mobile text, email, chatrooms, etc.) which access the message store database and convert the messages to the standard format (SMS, IMAP, etc.) for display and transmission [HBN03]. By implementing the system in this way, GlobalCom allows the user to choose what device and application to use to access their messages (as opposed to the system trying to decide the most likely client being used by them at that moment in time). However, a drawback to the system is that it increases the number of applications/clients that a user needs to operate as they will still have to interact with people outside of this closed messaging system and thus continue to have to use their regular email/chat clients, etc.

#### 3.1.2. A Proxy-Based Platform

The iMobile EE [CHJ+03] project aims to hide the complexity of multiple devices and content sources by acting as a message gateway that allows mobile devices using various protocols on different access networks to relay messages to each other. This is a continuation of the original iMobile [RCCC01] project and as well as implementing a message proxy, it seeks to provide proxies for information such as stock quotes, weather and flight information. The iMobile research approaches the messaging problem in a similar way to our project, where the processing of messages is carried out on a single server which in turn routes the messages to the other clients. iMobile allows devices to connect to the iMobile server[1] over the GSM network. This differs from our research, which implemented Bluetooth InfoStations to allow inter-device networking within the campus without depending on commercial providers (part of the project's remit was to be able to minimise the cost of sending the messages by utilising the University network or creating new communications infrastructure over which the operation costs could be controlled).

### 3.2. Innovations

Where our research differs from many of the other projects in this field is in the logic which decides how to deliver the message to the recipient. Other applications have focused on improving the connection between the mobile devices such as improved bandwidth, latency, etc. This has been termed "Always Best Connected (ABC)" [Mat06] and the message is transmitted in a single format (e.g. SMS).

Rather than introduce new message formats or new applications for communicating, this research sought to build upon existing applications in an attempt to create a universal messaging service. This allows the user to continue using their messaging applications of choice and to hide the translation between the message formats in the messaging service infrastructure. The next section describes the architecture of the messaging service and gives a high-level view of how all the components interact (the actual implementation details are discussed in Section 5).

## 4. System Architecture

This section describes the overall architecture of the messaging service and how each part of the system interacts with all the other components.

A graphical overview of how the system components interact is given in Fig. 2 and by following the arrows in the diagram, some of the possible message paths though the system can be traced. A number of use cases, given in Section 2, describe how a user would interact with the system.

### 4.1. Two-Tier Address Space

One of the central innovations in this project is the use of two-tier address space architecture. Each user in the messaging service is identified by a unique userID (more details in Section 4.2). This userID can be considered a 'virtual address' which the messaging service maps to a specific contact address ('real address') in the list of contact

[1] *the equivalent of the MessageRedirector in our project*

addresses that the recipient has provided.

An analogy for this dynamic one-to-many address translation would be a demultiplexer (a digital switching device that has one input and several outputs). The sender uses the userID to identify the recipient; the messaging service demultiplexes the userID input and outputs the required contact address (the Intelligent Assistant provides the 'control bits' to select the contact address). A graphical representation of the operation is depicted in Fig. 3.

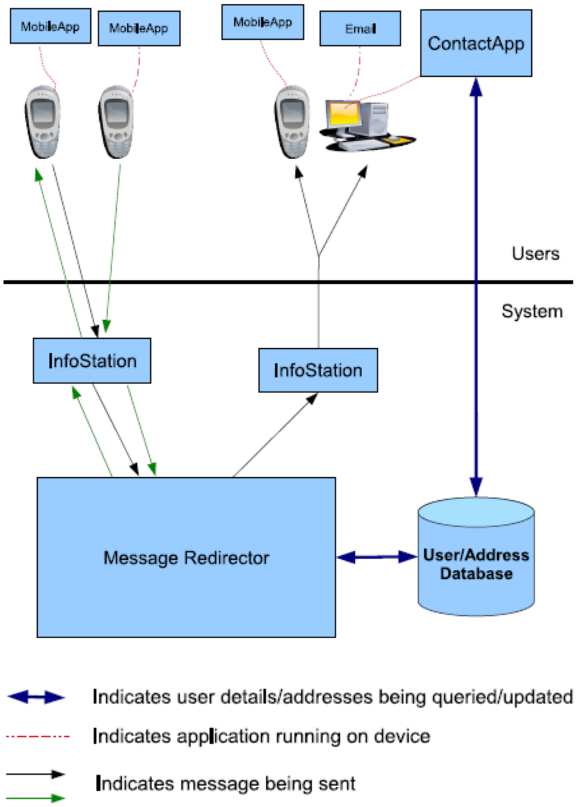

MobileApp
MobileApp
MobileApp
Email
ContactApp
Users
System
InfoStation
InfoStation
Message Redirector
User/Address Database
Indicates user details/addresses being queried/updated
Indicates application running on device
Indicates message being sent

*Figure 2: Messaging Service Components*

Benefits of this address space architecture include:

- The recipient gets the message delivered to the most suitable address/device.
- The sender need only maintain a single contact address for the recipient.
- If the recipient does not want to be disturbed, they can forward messages to an address that won't result in an interruption.
- The recipient can add new contact addresses without having to inform all their contacts. Once added to the ContactApp, the new address will be used if it meets the criteria decided by the Intelligent Assistant.

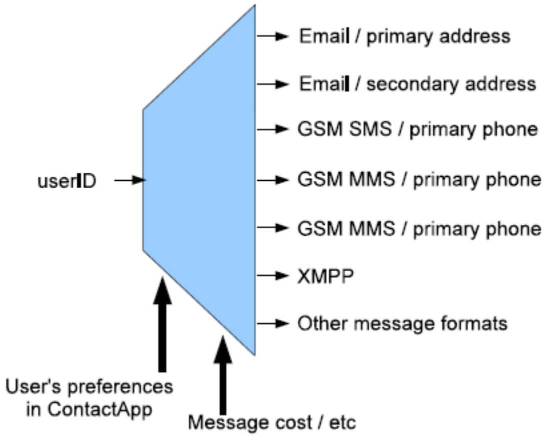


*Figure 3: Two-tier Address Space - an address demultiplexer*

## 4.2. Database

Each user of the messaging service has a unique identifier in the format of an RFC2822 [Res01] address (i.e. like the format of email addresses) with a userID part and a domain part (userid@domain).

Each user profile has one or more addresses associated with it. Each address entry consists of three pieces of information: (1) the contact address (e.g. email address in the RFC2822 format, mobile phone number), (2) the type of address (email, SMS, etc.), and (3) a user assigned priority to allow the user indicate to the system the ordering of the formats that they would prefer to receive messages (e.g. via email rather than via Instant Messaging (IM)). In the absence of an ordering specified by the user, a default ordering of the formats is provided by the system, which favours formats that have a lower cost per transaction (e.g. send as email rather than SMS, which would incur a network provider charge).

The user list and address list are stored in a database accessible to all the infrastructure components. Each address has a link to the associated user details via a "foreign key".

Multiple user profiles can also be grouped in hierarchies (in the same way email addresses can be added to mailing lists) so that messages can easily be sent to multiple users with each one receiving the message in their preferred format as described in

Section 2.

It is also possible for a user to access the database via the web interface (c.f. Section 4.6) and update their profile. The operations supported include:

- Add/delete contact methods (“addresses”) by which messages can be sent to the user.
- Disable/enable specific contact methods in the database.
- Select a default method of contact.
- Set the ‘priority’ of an address in order to give user a hint to the Intelligent Assistant (c.f. Section 4.5.2) as to which address to select.

## 4.3. Mobile Application

The mobile application (MobileApp) runs on the users’ mobile devices (such as a mobile phone or PDA). This is what is known as the Mobile Station in GSM parlance [Sco97], [ETS98a]. The application allows the user to send messages in various standard formats such as SMS [ETS98b] and interfaces with external messaging systems such as email.

### 4.3.1 Functionality

The MobileApp provides a message editor which allows a user to compose text messages.

The MobileApp can also save received messages and has an application-specific AddressBook to store the userID/domain tuple of contacts (This address data (userID/domain tuple) is also sent as part of the message format and the MessageRedirector can query the database (c.f. Section 4.2) to get the recipient’s contact details).

The application on the mobile device controls when the messages are pushed from the mobile device and pulled from the InfoStation. It polls the InfoStation on a regular basis while it is within range to indicate that it is still within the cell and to check to see if any messages are available for the user.

### 4.3.2. Network Access

The messaging service has to be able to cope with having intermittent network access as the user will not always be within range of an InfoStation (c.f. Section 4.4).

Messages which fail to be sent will not be automatically retried asynchronously (this is a common design, e.g. [Hos02] and implementations of the SMS specification [ETS98b] do this). If an attempt to send a message fails for whatever reason, the sender is given the option of retrying immediately or saving the message to persistent storage. Implementing asynchronous retrying would be too complex, may not be what the user wants, as the information contained in the message may have a limited useful lifespan and repeated attempts at sending the message may use too much power and reduce the battery life of the mobile device.

The MobileApp only connects to one InfoStation at a time and there is no ’hand-off’ (when a mobile device switches seamlessly to a new InfoStation, discussed in [TWB96], [AMH+99], [BB95]). This will not be an issue as sending a message is an instantaneous event (unlike a voice call, which can last for a considerable amount of time, during which the user may be in motion and move out of range of the InfoStation to which they were connected).

## 4.4. InfoStation

The messaging service contains one or more InfoStations which provide the coverage necessary for the users to connect with the messaging service. The requirements for the InfoStation hardware are ruggedness, reliability and low cost as there could be quite a number of them in remote and outdoor locations in a large university campus.

A graphical representation of the major components in the InfoStation is depicted in Fig. 4

The InfoStation provides a service for the mobile device to connect and upload any messages that the user has created and wishes to send to other users of the messaging service. It then places these messages on its Send (Message) Queue, which the MessageRedirector will be watching and will read from when a new message arrives.

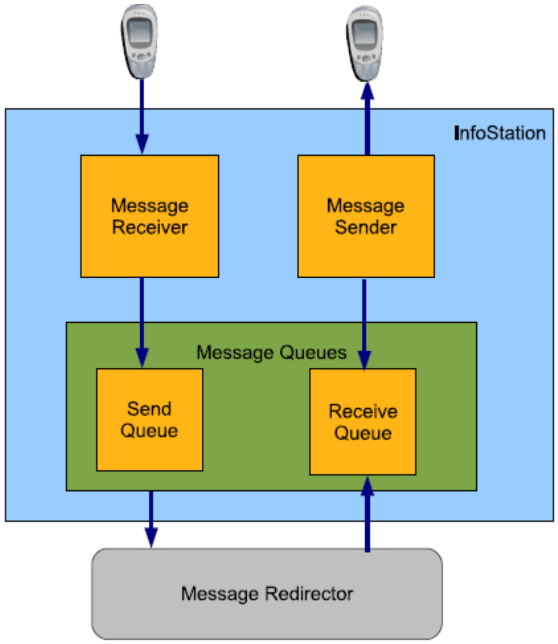


*Figure 4: Diagram of the InfoStation internals*

The InfoStation also provides a service for the mobile device to connect and download messages intended for its user. The InfoStation will have been polling the Receive (Message) Queue and will download any messages that the MessageRedirector has placed on it to be routed to a mobile device, which is known to be connected to that InfoStation. The mobile device checks on a regular basis and downloads any messages which are outstanding for that user.

## 4.5. MessageRedirector

The MessageRedirector (MR) is the central controlling component of the messaging service and contains some of the novel innovations referred to in Section 3.2. Each InfoStation will pass any message it receives to the MessageRedirector for processing as can be seen in Fig. 5.

The next subsections describe the major modules that make up the MessageRedirector and how the modules work in unison.

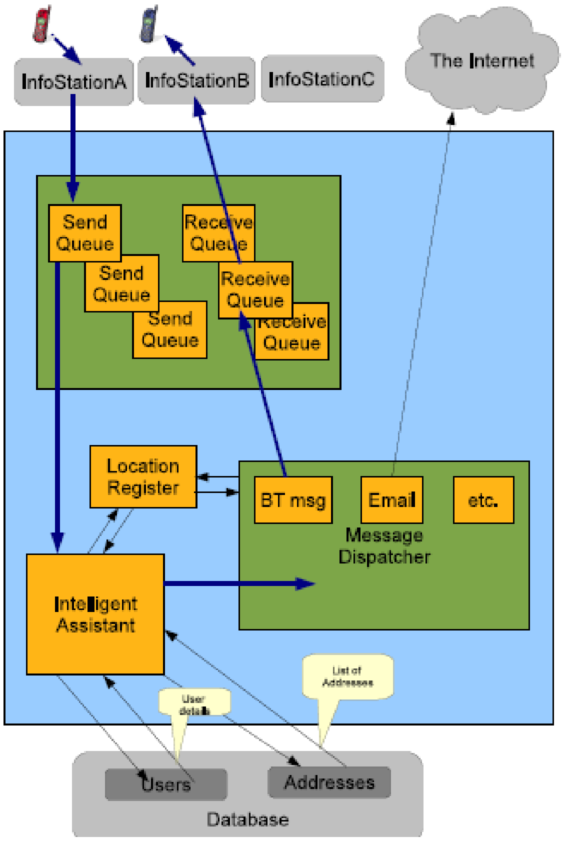


*Figure 5: Diagram of the MessageRedirector internals*

### 4.5.1. Overview of operations

When the MessageRedirector receives a message, it parses the message format and gets the sender's and recipient's userIDs as well as decoding the actual message payload. It provides the recipient's userID to the Intelligent Assistant, which provides the MessageRedirector with their preferred contact address. The MessageRedirector then inputs the sender's userID, the recipient's contact address and the message payload to the Message Dispatcher, which sends the message to the recipient.

### 4.5.2. Intelligent Assistant

It is the Intelligent Assistant that decides what address the message is sent to. The Assistant queries the database for all the addresses associated with the recipient's userID and decides the user's preferred contact address. The Assistant takes into account factors such as the urgency of the message, the cost associated with sending the message to the various addresses and passes back to the MessageRedirector the address to which it thinks the message should be forwarded.

### 4.5.3. Message Dispatcher

Once the preferred address has been discovered, the MessageRedirector passes the address and the message text to the Message Dispatcher module, which sends the message in the required format (which can be decoded from the preferred address). Knowledge of the various message formats is limited to the Message Dispatcher module and this encapsulation of message processing allows the MessageRedirector itself to have no knowledge of the message formats. Adding support for a new format is a matter of plugging in an extra format implementation to the Message Dispatcher without having to modify the MessageRedirector itself.

### 4.5.4. Location Register

The messaging service keeps track of each user's current location using a Location Register, which contains the address of the InfoStation to which the user's mobile device is connected. The MessageRedirector uses this information to route the messages to the correct InfoStation, which in turn sends the message to the mobile device. This is based on the same principle as the Home Location Register (HLR) in GSM [Sco97], [CLA02].

### 4.6. Web Interface

A web interface to the messaging service is provided so that users can modify their profile, which contains their address details.

The messaging service administrator can create an account/password for a user using the web interface (called ContactApp).

When the user logs in and is successfully authenticated, the application displays the user's details and the list of addresses that they have registered. The user can then add or remove addresses, modify their details or change the preferred priority of the addresses (i.e. control the order in which the MessageRedirector selects an address in order to forward a message to the user).

### 4.7. Privacy Issues

Depending on how this project was implemented, there could have been privacy concerns with the information that may have been available. For example, if delivery notification was enabled, it could be possible to track a user's presence depending on how a message was delivered to them (e.g. a person would have to be on campus to receive a Bluetooth message).

This is an issue that other projects have also encountered: [JPB05] (mentioned in Section 3) designed a system where the sender provides the context in which a message should be delivered - the sender does not have to know the recipient's current presence status or indeed should not find out that status without their permission. As in this project the DeDe team also chose not to provide delivery notifications.

## 5. Implementation

This section describes how the messaging service is implemented. Section 5.1 describes the messaging service components that the user interacts with and Section 5.2 details the necessary background components required to distribute the messages.

### 5.1. User Interface

The user interacts with the messaging service through the mobile application (MobileApp, running on the mobile device) and the web interface (ContactApp, accessible via any web browser). Typically a user would use the MobileApp more frequently than the ContactApp; once addresses are entered in the ContactApp, a user would probably only log in to change their preferred contact address once or twice a day whereas they would send and receive messages via the MobileApp throughout the day.

### 5.1.1. Mobile Application

The MobileApp application was written in the Java programming language which allows it to run on the wide range of phones and mobile devices that run J2ME (a specialised virtual machine specifically for low-powered mobile devices). Details of the MobileApp were provided in Section 4.3.

In order to guarantee a usable response and a satisfactory user experience, the MobileApp User Interface (UI) and the message sending/receiving components have separate software threads of execution. This multithreaded approach means that the user can navigate the application's menus while messages are being sent and received in the background and there is no risk of the mobile device display locking up if a user goes out of range of an InfoStation, etc.

A screenshot of the application's main menu is shown in Fig. 6. This menu shows a list of the features available in the application, which will be explained in further sections.

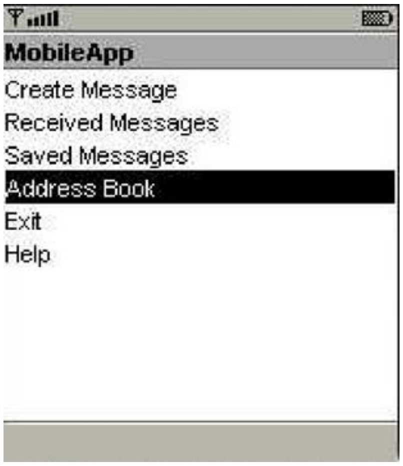


*Figure 6: Menu of features of MobileApp*

## Message Editor

A central component of the MobileApp is the message editor which enables the user to compose the messages they wish to send. It utilises the comprehensive editing features of the J2ME platform such as predictive text entry.

## Address Book

Users of MobileApp can save the addresses (userid/domain tuples) of people they frequently send messages to. The address book saves and retrieves the contact details to/from persistent storage on the mobile device using the J2ME *RecordStore* class. This maintains the address list on the mobile device across sessions and after the device has been powered OFF.

## Message Stores

Other features of the MobileApp include the ability to save received and unsent messages (e.g. messages composed while out of range of an InfoStation). The messages are also saved to persistent storage on the mobile device using the J2ME *RecordStore* class.

## 5.1.2. Web Interface

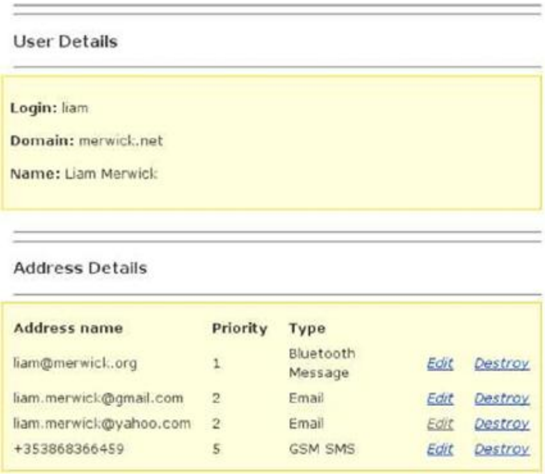
The web interface allows a user to control how messages are routed to them and is accessible using any web browser (c.f. Section 4.6).

The web application's main requirement is that it can access the database containing the user details and that a web server is running on the machine on which the web application is hosted (the open-source Apache web server (httpd) is used in this prototype).

The first page presented to the user when they connect to the website is an authentication page with an option to jump to a sign-up page, which provides the opportunity for a user to subscribe to the service, where they could provide their username and password, if this is their first time using it.

A screenshot of the key part of the main menu that is presented to the user upon login is shown in Fig. 7. The upper part of the page shows the userID of the person currently logged in and lists the operations that the user can perform on the addresses. Other parts of the ContactApp, not shown here, permit operations such as logging out of the session, adding a new address or modifying their details.

The upper part of the page shows the user's details and the lower part shows the addresses that are associated with this user. The addresses are displayed in the order of the priority that the user assigned to them. Addresses can be deleted, disabled or modified from this page.



User Details

**Login:** liam

**Domain:** merwick.net

**Name:** Liam Merwick

Address Details

| Address name | Priority | Type | | |
|---|---|---|---|---|
| liam@merwick.org | 1 | Bluetooth Message | Edit | Destroy |
| liam.merwick@gmail.com | 2 | Email | Edit | Destroy |
| liam.merwick@yahoo.com | 2 | Email | Edit | Destroy |
| +353868366459 | 5 | GSM SMS | Edit | Destroy |

*Figure 7: Menu of features of ContactApp*

The format of these addresses was explained in Section 4.2. In addition to the system not delivering messages to addresses it knows are not available (e.g. the device is out of range), the user can also manually disable an address so that the messaging service does not consider it when it decides which address to select.

When a user adds a new address by which they can be contacted, he/she enters the address, a priority relative to their other addresses and then selects the address type from a drop down list of all the address types supported.

## 5.2. Prototype System Infrastructure

This section details the implementation of the infrastructure used by the messaging service. The user does not directly interface with these components but they are essential to the operation of the system. The infrastructure is made up of the MessageRedirector and InfoStation applications provided by this project along with externally provided supporting applications such as the Message Queue.

### 5.2.1. InfoStation

In the prototype implementation for this project the InfoStation Java application runs on basic PC hardware but when deployed in a production messaging service in a University, a low cost ruggedised system could be built and used. A USB dongle was added to the PC to provide Bluetooth connectivity.

The InfoStation application is a multithreaded application, with two threads waiting for Bluetooth connections. The first thread receives messages users wish to send and another thread pushes messages to the users' mobile device when it periodically connects to check for messages. In order to communicate with the MessageRedirector, two message queues are created for each InfoStation in the system (c.f. Figure 4).

### 5.2.2. MessageRedirector

For this project, the MessageRedirector also runs on basic PC hardware and is implemented in Java. In addition to the architecture and requirements listed in Section 4.5, the MessageRedirector initialises some of the general infrastructural components such as the Java Message Queues which are used by the InfoStations to communicate with the MessageRedirector.

#### Intelligent Assistant

The 'Intelligent Assistant' is the "brains" of the MessageRedirector. It is the module within the MessageRedirector which parses the message received from an InfoStation, decodes the message receiver's address and decides what format/device belonging to that user should receive the message.

Each message received by the Message Redirector is parsed to decode the sender's and recipient's addresses as well as the message payload.

#### Message Dispatcher

Given that we have a range of message formats (and possibly more in the future), we wish to have an extensible way of sending messages once we know the format they are to be sent in. For this reason, the component in the MessageRedirector that processes and routes the messages is implemented using the Factory software pattern [EGV95]. A generic MsgDispatcher class is sub-classed by the specific dispatcher implementation for each message type (e.g. BluetoothMsgDispatcher, EmailMsgDispatcher, etc.). When the Intelligent Assistant returns the recipient's contact address, it also informs the MessageRedirector of the address type and the message payload is input to the Message Dispatcher class of that type.

### Location Register

As described in Section 4.5.4, the Location Register is a module within the MessageRedirector which keeps track of the InfoStation to which a user's mobile device is currently connected. Each time a message is received by the MessageRedirector from an InfoStation, the MessageRedirector calls the setUserLocation method in the Location Register with the sender's details and information on the location of the InfoStation to which the sender is connected. The Location Register stores this in an in-memory HashMap so that the MessageRedirector can later find out what InfoStation to send a message to in order for the InfoStation to push it to the user's mobile device.

If a mobile device has not been in contact with an InfoStation within the previous two minutes, it is deemed to be out of range and the entry is considered dormant in the Location Register. However, it is not evicted from the cache so that a record is kept of a user's last known location. Instead, any requests to the Location Register for that user via the isUserConnected API routine return an 'out of range' error until contact is made by the device again.

Naturally, a support application to provide self-learning, smart user location management could be developed which would be campus- and user-specific. For instance knowing that certain users can be within range of specific InfoStations (e.g. restaurants, meeting rooms, etc.) for extended periods of time (greater than two minutes) would allow the Intelligent Assistant to correlate a location returned by the Location Register with a list of delivery preferences (controlled by the user) for specific locations and to choose a delivery method based on the user's preferences (with the user having to explicitly change their address preferences when they entered the area).

## 6. Conclusions

This paper has described the realisation of an Intelligent Messaging Service in an InfoStation-based university network along with a detailed explanation of the underlying components which make up the system. A functional messaging service was developed consisting of all the necessary components to allow end-to-end message delivery including the mobile device application, InfoStations, central processing unit as well as the web infrastructure to manage the users' profile and contact information. The system was implemented within a single building, rather than campus wide, but we were successful in sending messages between mobile devices as well as to external messaging services (e.g. to email servers hosted on the Internet). Delivery times were measured as being comparable with existing formats (e.g. an email between two users) which demonstrated that inserting an Intelligent Assistant into the communication's critical path and performing delivery decisions based on a user's location was feasible.

Future work includes extending the number of message formats supported, implementing Wi-Fi connectivity support on both application running on the mobile devices (MobileApp) as well as the InfoStation, improving the interaction with standard applications provided by the phone platforms (e.g. augmenting the standard contact list to include the address details for the unified messaging service) and adding a billing solution for the message formats which incur a charge when being sent. Also an important augmentation of the Location Register support for the Message Redirector would be the development of self-learning smart location management functionality.

## References

1. [AMH+99] I.F. Akyildiz, J. McNair, J.S.M. Ho, H. Uzunalioglu, and W. Wang. Mobility Management in Next-Generation Wireless Systems. *Proc. of the IEEE,* vol. 87, no. 8, pp. 1347-1384, Aug. 1999.

2. [Bar02] D. Barber. Globalcom: a unified messaging system using synchronous and asynchronous forms. *Proc. of the inaugural conference on the Principles and Practice of Programming in Java, PPPJ '02/IRE '02,* .pp 141–144; Held in National University of Ireland Maynooth, County Kildare, Ireland, June 13-14, 2002. J. Waldron, J.F. Power (Eds.). ACM International Conference Proceeding Series 25 ACM.

3. [BB95] Ajay V. Bakre and B. R. Badrinath. Handoff and systems support for indirect TCP/IP. In*: Proc. of the 2nd Symposium on Mobile and Location Independent Computing, MLICS '95,* pp. 11–24; Berkeley, CA, USA, 1995. USENIX Association; ACM Digital Library.

4. [CHJ+03] Yih-Farn Chen, Huale Huang, R. Jana, T. Jim, M. Hiltunen, S. John, S. Jora, R. Muthumanickam, and Bin Wei. Immobile ee: an enterprise mobile service platform. ACM *Wireless Networks*, vol. 9, no.4, pp. 283–297, 2003.

5. [EGV95] Ralph Johnson Erich Gamma, Richard Helm and John Vlissides. *Design patterns: elements of reusable object-oriented software*. Addison-Wesley Professional Computing Book Series. 1995.

6. [ETS98a] European Telecommunications Standards Institute. *ETSI GSM Technical Specification 04.22*, GSM Radio Link Protocol for data and telematic services, Version 6.1. November 1998.

7. [ETS98b] European Telecommunications Standards Institute. *ETSI SMS specification: ETSI TS 100 901 (GSM 03.40 version 7.3.0 Release 1998)*. 1998.

8. [Faj98] R. Fajman. *IETF RFC 2298 - An Extensible Message Format for Message Disposition Notifications*. 28 pages.1998.

9. [HBN03] P. Healy, D. Barber, and B. Nolan. Developing unified messaging system apps in java. Proc. *Proceedings of the 2nd international conference on Principles and Practice of Programming in Java, PPPJ '03,* pp. 137–138. Held in Kilkenny, Ireland, June 16-18, 2003. J. Waldron, J.F. Power (Eds). ACM International Conference Proceeding Series 42 ACM.

10. [Hos02] Ashima Hosalkar. Building Mobile Applications with J2EE, J2EE-J2ME and J2EE Extended Application Servers. *Proc. of the Mid-Atlantic Student Workshop on Programming Languages and Systems, MASPLAS'02*, Pace University, White Plains, NY, USA, April, 2002.

11. [Iva06] I. Ganchev, S. Stojanov, M. O'Droma and D. Meere. An InfoStation-Based University Campus System for the Provision of mLearning Services. *Proc. of the IEEE Sixth IEEE International Conference on Advanced Learning Technologies, ICALT '06* pp. 195–199. 5–7 July, Kerkrade, The Netherlands. IEEE Computer Society. 2006.

12. [Iva07] I. Ganchev, S. Stojanov, M. O'Droma and D. Meere. An InfoStation-Based University Campus System Supporting Intelligent Mobile Services. In *Journal of Computers*, vol. 3, pp. 21–33. Academy Publisher, 2007.

13. [Jon04] J. Häkkilä and J. Mäntyjärvi. User experiences on combining location sensitive mobile phone applications and multimedia messaging. *Proc. of the 3rd International Conference on Mobile and Ubiquitous Multimedia, MUM'04,* pp. 179–185, New York, NY, USA. ACM Press. 2004.

14. [JPB05] J. Younghee, P. Persson, and J. Blom. DeDe: design and evaluation of a context-enhanced mobile messaging System. *Proc. of the SIGCHI conference on Human factors in computing systems, CHI '05* pp. 351–360, New York, NY, USA. ACM Press. 2005.

15. [Mat06] Matthias Siebert, I. Ganchev, M. O'Droma, F. Bader, H. Chaouchi, I. Armuelles,

I. Demeure and F. McEvoy. A 4G generic ANWIRE system and service integration architecture. *SIGMOBILE Mob. Comput. Commun. Rev.* 10(1), pp. 13–30. 2006.

16. [RCCC01] Chung-Hwa Herman Rao, Yih Fam Robin Chen, Ming-Feng Chen, and Di-Fa Chang. iMobile: a proxy based platform for mobile services. *Proc. of the first workshop on Wireless mobile internet, WMI '01*, pp. 3–10, New York, NY, USA, 2001. ACM Press. 2001

17. [Res01] P. Resnick, Ed.. *IETF RFC 2822 - Internet Message Format*, Internet Engineering Task Force. 2001.

18. [Sco97] J. Scourias. Overview: The global system for mobile communications, 1996. URL(2010): http://ccnga.uwaterloo.ca/~jscouria /GSM/gsmreport.html

19. [TWB96] M. S. Taylor, W. Waung, and M. Banan. *Internetwork Mobility the CDPD Approach*, Book. ISBN: 0-13-209693-5. 1997.

20. [CLA02] E. Clayirci. I.F.Akyildiz, User mobility pattern scheme for location update and paging in wireless systems. *Mobile Computing, IEEE Transactions on*. Vol 1 Issue 3. pp 236 – 247. 2002

# Optimized Blood Pressure Control during Surgery

[1]Mrinmoy Chakraborty [2]Achintya Das [3]P.P.Sarker
[1]Dr. B.C. Roy Engineering College, Durgapur, West Bengal, India
chakraborty_wb@yahoo.co.in
[2]Kalyani Govt. Engineering College, Kalyani, Nadia, India
achintya_das123@yahoo.co.in
[3]DTS, Kalyani University,Kalyani, Nadia

***Abstract***: The present work describes the design and analysis of a blood pressure control system during surgery using anesthesia in an efficient way. The total system under consideration is supposed to consist of the patient's body dynamics, suitable compensator, and surgical disturbance operated in the closed loop manner with unity negative feedback yielding controlled blood pressure as output. With appropriate control of valve setting of the anesthetic reagent chamber, the vapor of the reagent is emanated. This vapor emanation as the process is concerned is further mingled with surgical disturbance where the output is fed to the patient for controlled dose of anesthesia. With the application of controlled dose of anesthesia, the regulated blood pressure of the patient becomes achievable. As the output which is fed back in the unity negative feedback path to the compensator input of an error detector whose other input receives the desired blood pressure for the study of the patient is now pressure track under desired environment. The optimality of the performance for the system is considered to be attained with proportional gain of one proportional plus integral (PI) compensator, so chosen that the integral square error becomes a minimum. The present work considers the minimization using particle swarm optimization (PSO) which searches the maxima of the reciprocal of the ISE with in the range of the value of the system gain as obtained by Routh Hurwitz criterion. The overall system is found to be stable, controllable and observable. The system is also analyzed in sampled data control domain (z domain). The stability in z domain is analyzed using Jury's Stability test. The overall system offers an in-depth sight for the biomedical engineering application of a blood pressure control system operated in optimal condition. MATLAB software is appropriately used in the entire analysis.

***Keywords***: Stability, PI compensator, Integral Square Error, Sample data control

## 1. Introduction

Control of blood pressure during surgery using anesthesia is extremely important matter in medical science, because the blood pressure of a patient during surgery using anesthesia normally varies over a range, although the safe surgery demands the maintenance of constant mean arterial pressure to some requisite value. The level of arterial blood pressure is postulated to be a proxy for depth of anesthesia during surgery. The level of anesthesia is required to be controlled for healthy environment of the surgery for controlled level of mean arterial pressure

A block diagram of the system is shown in Figure: 1[1, 2], where the impact of surgery is represented by the disturbance D(s).

In the system one compensator which controls the overall system for appropriate performance criteria of the system is used.

The compensator considers the adaptive noise cancellation method towards the system design for having a desired control system.

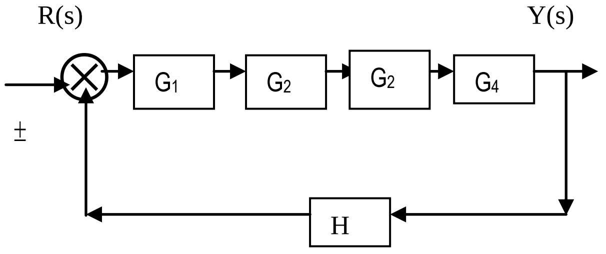


**Figure 1: Block Diagram of blood pressure control system.**

$G_1(s) = (ks+k_1) / s$; $G_2(s)=1/s$ $G_3(s) =1/s$; $G_4 (s) = 1/(s+2)^2$;
$H (s) =1$;
R(s) =Desired blood pressure, Y(s) =Actual blood pressure.

## 2. Blood pressure control

There is wide range of blood pressure in healthy subjects. Increase in blood pressure occurs with age. In the in the same individual, transient variations in blood pressure is common [3], nervousness, excitement, exertion, fatigue, cold and fatigue may raise the normal level to some extent but in these conditions, systolic pressure is affected more than the diastolic one.

The systolic blood pressure is mainly determined by the force of contraction of the left ventricle. The diastolic blood pressure is regulated by the arteriolar resistance, which converts the intermittent output of the heart into a continuous capillary blood flow. During systole the large musculo-elastic arteries are distended and during diastole their elastic recoil helps to maintain the arterial pressure.

The arterial pressure consists of systolic and diastolic pressures. The mean arterial pressure is the average arterial pressure throughout each cardiac cycle of the heart beat. The arterial pressure usually remains nearer to diastolic level than to systolic level during a greater portion of the pulse cycle.

The mean arterial pressure is lowest immediately after the birth, measuring about 70 mm of Hg at birth and reaching an average of amount 110 mm Hg in the normal old person and as high as 130 mm Hg in person with arteriosclerosis [4].During surgery due to loss of blood, the mean arterial pressure falls (Figure 2:).The blood pressure is, however, very importantly required to be maintained to specific safe value for normal surgery operation., The surgery operation using anesthesia again having other contribution to changing blood pressure. So the anesthesia dose is ultimately required to be controlled for safe surgical environment.

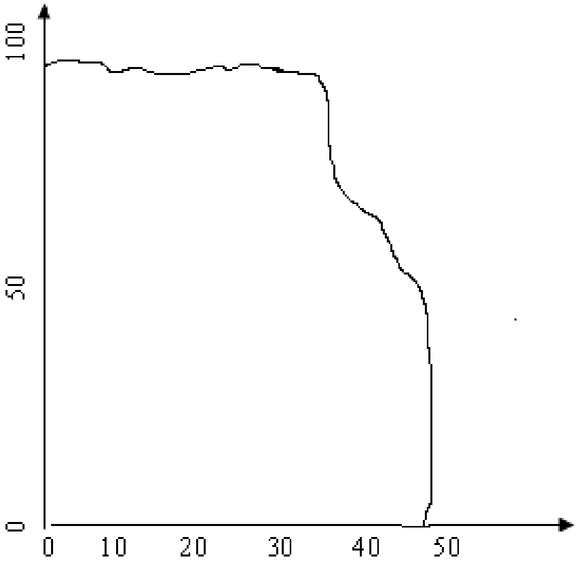


**Figure 2: Hemorrhage effect on arterial pressure.**

## 3.Conversion from s-domain to z-domain [5]:

Z-transform helps in the analysis and design of sample data control system, as Laplace transform does in the analysis and design of continuous data control system.

The z-transform $F(z)$ of a sample data control signal $f(KT)$ is defined by the relation:

$$F(z) = \sum_{K=0}^{\infty} f(KT) z^{-K} \quad \ldots (1)$$

The above relation is derived from the Laplace transformation as applied to sample data control signal.

Assuming $e^{sT} = z$ be the concerned transformation variable in Laplace Transformation,

we have

$$sT = \ell nz$$

i.e. $$s^{-1} = \frac{T}{\ell nz} \quad \ldots (2)$$

Using power series expansion of $\ell nz$, the above equation becomes:

$$s^{-1} = \frac{T}{2}\left[\frac{1}{u} - \frac{4}{3}u - \frac{4}{45}u^3 - \frac{44}{945}u^5 \ldots\ldots\ldots\right] \quad \ldots (3)$$

Where

$$u = \frac{1 - z^{-1}}{1 + z^{-1}} \quad \ldots (4)$$

In general, for any positive integral value of $n$

$$s^n = \left(\frac{T}{2}\right)^n \left[\frac{1}{u} - \frac{1}{3}u - \frac{4}{45}u^3 - \frac{44}{945}u^s\right]^n \quad \ldots (5)$$

By using binomial expansion in the above equation for various values of $n$, we may have the transformation from $s$ to $z$ domain.

### 4. Integral square error (J) [6]:

Instead of the time domain calculation of $J$ (integral square error), the complex frequency domain can be used. According to a theorem in mathematics by Parseval

$$J = I_{SE} = \int_0^{\infty} e^2(t)dt = \frac{1}{2\pi j} \int_{-j\infty}^{j\infty} E(s)E(-s)ds \ldots (6)$$

Where $E(s)$ can be expressed as follows:

$$E(S) = \frac{N_{n-1}s^{n-1} + \ldots + N_1 s + N_0}{D_n s^n + D_{n-1}s^{n-1} + \ldots + D_1 s + D_0} \quad \ldots (7)$$

assuming type 1 behavior.

$J$ follows from complex variable theory. To clarify the effect of system order, the subscript for $J$ will be the system order. For an nth –order system.

$$J_n = (-1)^{n-1} \frac{B_n}{2D_n H_n} \quad \ldots (8)$$

Where $H_n$ and $B_n$ are determinants. $H_n$ is the determinant of the $n \times n$ Hurwitz matrix. The first two rows of the Hurwitz matrix are formed from the coefficients of D(s), while the remaining rows consist of right-shifted versions of the first two rows until the $n \times n$ matrix is formed. Thus we write

$$H_n = \begin{bmatrix} D_{n-1} & D_{n-3}\ldots\ldots\ldots & \\ D_n & D_{n-2}\ldots\ldots\ldots & \\ 0 & D_{n-1} & D_{n-3}\ldots \\ 0 & D_{n-2}\ldots\ldots\ldots\ldots & \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & & \end{bmatrix} \quad \ldots (9)$$

The determinant $B_n$ is found by first calculating

$$N(s)N(-s) = b_{2n-2}s^{2n-2} + \ldots + b_2 s^2 + b_0 \quad \ldots (10)$$

Then first row of the Hurwitz matrix is replaced by the coefficients of $N(s)N(-s)$ while the remaining rows are unchanged.

$$B_n = \begin{bmatrix} b_{2n-2} & \ldots\ldots & b_2 \ldots b_0 \\ D_n & D_{n-2}\ldots\ldots\ldots & \\ 0 & D_{n-1} & D_{n-3}\cdots \\ 0 & D_n & D_{n-2}\cdots \\ \ldots\ldots\ldots\ldots\ldots\ldots & \ldots\ldots\ldots\ldots & \end{bmatrix} \qquad \ldots (11)$$

## 5. PARTICLE SWARM OPTIMIZATION (PSO) [6,7,8] FOR THE COST FUNCTION J$_n$

The cost function **J$_n$**, a function of k will be minimized i.e. the reciprocal of the same will be maximized using PSO.

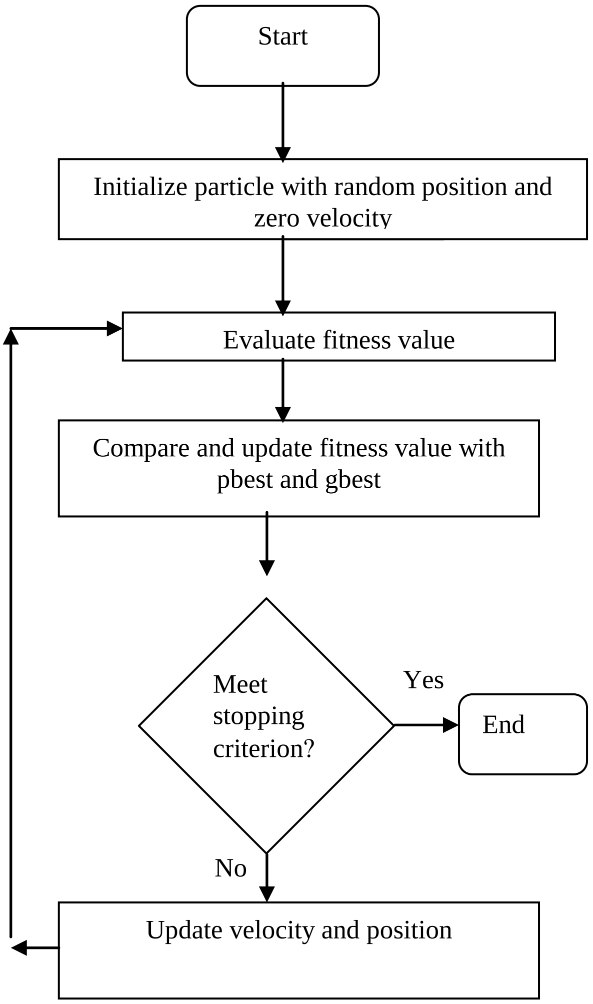

Start
Initialize particle with random position and zero velocity
Evaluate fitness value
Compare and update fitness value with pbest and gbest
Meet stopping criterion?
Yes
End
No
Update velocity and position

Fig .3: Flow chart for particle swarm optimization

pbest: the best solution (fitness) a particle has achieved so far, gbest: the global best solution for all particles. gbestk : the value of k for J$_n$ is minimum.

## . 6. SYSTEM DESIGN:

Under present situation, overall transfer function of the system is given by

$$\frac{Y(s)}{R(s)} = T(s)$$

$$= \frac{(ks+k_1)}{s^4+4s^3+4s^2+ks+k_1}.(12)$$

Thus for the entire system the characteristics equation is $s^4+4s^3+4s^2+ks+k_1=0$. For this characteristics equation to make the design problem with stability, the Routh array is constructed as below:

$$\begin{array}{llll} s^4 & 1 & 4 & k_1 \\ s^3 & 4 & k & \\ s^2 & 16-k & k_1 & \\ s^1 & \dfrac{16k-k^2-4k_1}{16-k} & 0 & \\ s^0 & k_1 & & \end{array}$$

For stability

$k_1 > 0$ ,let $k_1$=1, 16-k<0, i.e. k<16

Now

$$\frac{Y(s)}{R(s)} = T(s)$$

$$= \frac{(ks+1)}{s^4+4s^3+4s^2+ks+1}.(12)$$

$$T_E(s) = \frac{1-T(s)}{s}$$

$$= \frac{s^3+4s^2+4s}{s^4+4s^3+4s^2+ks+1} \qquad \ldots(14)$$

$$N_3 = 1, N_2 = 4,\ N_1 = 4,\ N_0 = 0 \qquad \ldots (15)$$

$$D_4 = 1, D_3 = 4\ , D_2 = 4\ , \qquad \ldots(16)$$

$$D_1 = k\ \ ,\ \ D_0 = 1 \qquad \ldots(17)$$

$$J_4 = -\frac{B_4}{2D_4H_4} \qquad \ldots (18)$$

$$H_3 = \begin{bmatrix} D_3 & D_1 & 0 & 0 \\ D_4 & D_2 & D_0 & 0 \\ 0 & D_3 & D_1 & 0 \\ D_4 & D_2 & D_0 & 0 \end{bmatrix} \qquad \ldots (19)$$

$$N(s) = s^3 + 4s^2 + 4s \qquad \ldots (20)$$

$$N(-s) = -s^3 + 4s^2 - 4s \qquad . \ldots (21)$$

$$J_4 = -\frac{B_4}{2D_4H_4}$$

$$J_4 = \frac{(12k+60)}{(32k-32-2k^2)}$$

$J_4$ will be minimum for k=6 and the minimum value of the same is obtained by Particle swarm optimization method.

## 7. METHODS AND MATERIAL

So long any control system is considered in continuous data control system (continuous time domain ↔ Laplace domain), the system analysis and study get restricted for any change in the system parameter, or input variation for easy and ready study. To circumvent this problem sample data (s.d.) control system makes study and analysis easy and ready available with variation in system parameter and also the input. For this reason the system is also studied in sample data control model. The stability of the present system is tested by Jury's stability test which guarantees the stability of the overall system. Needless to mention, any stable system when operated in s.d. mode, the system is not necessarily to be guaranteed to remain stable in the s.d. mode also, there being the enhancement of the order of the system.

As any control system deserves to reach its steady state by which the system finally runs, and follows the input at that state, the designed parameter K is accordingly decided, the other desirable characteristic performances being also available in the system.

## 8. Matlab output files [8]

```
H4 =
[ 4, k, 0, 0]
[ 1, 4, 1, 0]
[ 0, 4, k, 0]
[ 0, 1, 4, 1]

 H =16*k-16-k^2

 B4 =
[  -1,   8, -16,   0]
[   1,   4,   1,   0]
[   0,   4,   k,   0]
[   0,   1,   4,   1]

B =-12*k-60
J4 =-(-12*k-60)/(32*k-32-2*k^2)
Pbestx=6
n =    12    60
d =    -2    32   -32
nd =  24         240       -2304
dd =   4       -128        1152       -2048        1024
k =   -16
        6
s =  -3.3117
   -0.2514 + 1.2506i
   -0.2514 - 1.2506i
   -0.1856
a =    -4    -4    -6    -1
        1     0     0     0
        0     1     0     0
        0     0     1     0
b =     1
        0
        0
        0
c =     0     0     6     1
d =     0
t = 0     0     6   -23
    0     6     1   -24
    6     1     0   -36
    1     0     0    -6
dobs =   4
t1 = 1     -4    12  -38
     0      1    -4   12
     0      0     1   -4
     0      0     0    1
rcont =     4

Transfer function:
         6 s + 1
-----------------------------
s^4 + 4 s^3 + 4 s^2 + 6 s + 1
 Margins =   1.7430    1.8257    5.8176    1.7256
Transfer function:
         6 s + 1
-----------------------------
s^4 + 4 s^3 + 4 s^2 + 6 s + 1

Transfer function:
0.00663 z^3 + 0.01558 z^2 - 0.01675 z - 0.004375
-----------------------------------------------
 z^4 - 3.322 z^3 + 4.127 z^2 - 2.253 z + 0.4493

Sampling time: 0.2
Enter length of output vector = 5
Type in the numerator coefficients =
   [0.00663  0.01558 -0.01675  -0.004375]
Type in the denominator coefficients =
    [1 -3.322  4.127  -2.253   0.4493]

Coefficients of the power series expansion
   0.0066    0.0376    0.0808    0.1238    0.1596
d =    1.0000   -3.3220    4.1270   -2.2530    0.4493
t =0.9250 + 0.2335i
   0.9250 - 0.2335i
   0.9553
   0.5168
v = 0.0013
z = 11.1513
s =0.4493   -2.2530    4.1270   -3.3220    1.0000
   1.0000   -3.3220    4.1270   -2.2530    0.4493
  -0.7981    2.3097   -2.2727    0.7604         0
   0.7604   -2.2727    2.3097    2.3097         0
   0.0588   -0.1152    0.0576         0         0
a0 =    0.4493 a4 =  1 b0 =    0.7981 b3 =    0.7604
```

c0 = 0.0588c2 = 0.0576

## 9. Output plots

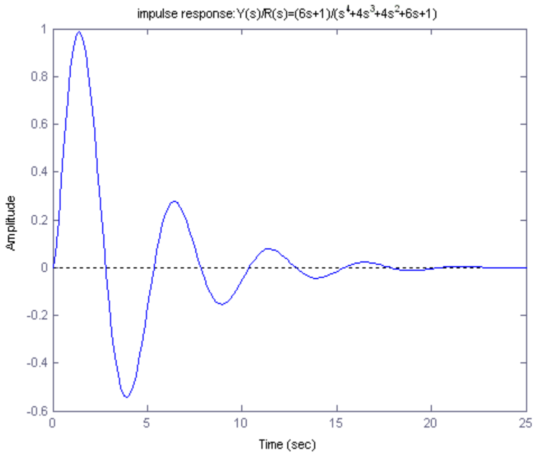


**Figure 4**

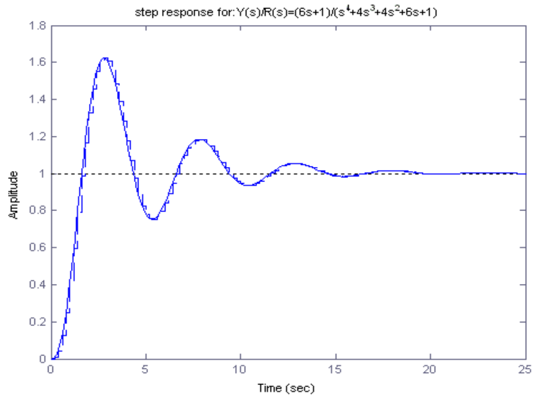


**Figure 5**

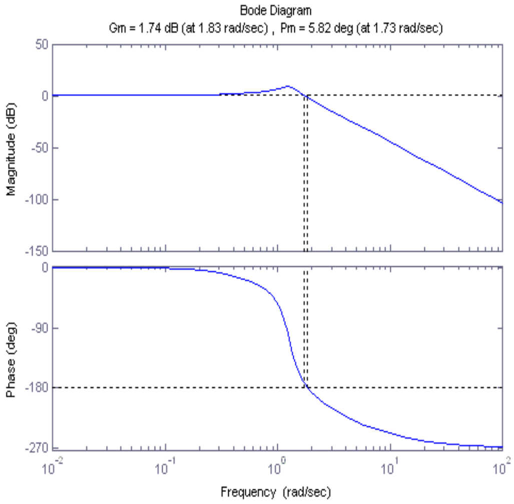


**Figure 6**

## 10. Conclusion

Since the system is stable in both continuous and sample data control system, controllable, observable and having appropriate gain and phase margin so the design of a blood pressure control system becomes feasible.

## 11. References*:*

1. R. C. Dorf, and R. H. Bishop, "Modern Control Systems", Addison Wesley, pp. 748-749, First Indian Reprint 1999.
2. R.Meier,'Control of Blood Pressure During Anesthesia', IEEE Control system, pp. 12-16, December 1992.
3. R.B.Scott, 'Prices text book of the practice of medicine', Tenth edition, The English language book of society and oxford university press, 1960.
4. A. C. Guyton, "Text Book of Medical Physiology", W.B.Saunders Company, Sixth Edition, pp-136, 1981.
5. Achintya Das, "Control Systems", Vol2, Matrix Educare, pp. 97-98, March 2009.
6. R.T.Stefani, B. Shahian, C.J.Savant, G.H, Hostetter,"Design of Feedback Control Sysrtems", Oxford University Press, pp. 210-211, 2002.
7. Carlisle A. Dozier G "An of the self PSO", Proceedings of 2001 workshop on Particle Swarm Optimization", Indianapolis, IN pp.1-6, 2001.
8. M. Abido, "Optimal Design of power system stabilizers using particle swarm optimization", IEEE Transactions on energy Conservation, 17:406-413, 2002.
9. Z.L.Gaing, "A particle swarm optimization approach for optimum design for pid controller in avr system" IEEE Transactions on energy Conservation, 19:384-391, 2004.

# Classifying Squint Risk Level of Squint Eye Patients from Pattern Visual Evoked Potential Signals

R.Kalaivaazhi[1] and Dr.D.Kumar[2]

[1]Anjalai Ammal Mahalingam Engineering College, Anna University,
[2]Periyar Maniyammai University,
[1]Assistant Professor, Department Of Information Technology,AAMEC,Kovilvenni, Thiruvarur (dt) , Tamilnadu, India
[2]Dean Research , Periyar Maniyammai University, Vallam, Thanjavur(dt), Tamilnadu, India.
{vazhi@hotmail.com , kumar_durai@yahoo.com }

***Abstract:*** Classification of squint risk level of squint eye patient is a classical problem. In this study, genetic algorithm(GA) and adaptive neuro fuzzy inference system (ANFIS) are used in the classification of squint risk level from pattern visual evoked potential (P-VEP) signal parameters. The squint risk level is classified based on the extracted parameters like energy, variance, peaks, sharp and spike waves, duration, events, covariance and P100 latency from the P-VEP of the patient. This paper focuses on comparison of genetic algorithm ( GA) and adaptive neuro fuzzy inference system(ANFIS) in classification. Genetic algorithm (GA) and ANFIS are applied on the code converter's classified risk levels to optimize risk levels that characterize the patient. The Performance Index(PI) and Quality Value (QV) are calculated for these methods. A group of ten patients with known squint findings are used in this study. High PI such as 93.33% and 97.83% for GA and ANFIS are obtained at QV of 20.14 and 24.59.

***Keyword:*** P-VEP Signals, P100 latency, Squint eye, Genetic Algorithm, Adaptive Neuro Fuzzy Inference System, Risk Level.

## 1. Introduction:

The recognition of specific waveforms and features in the Pattern Visual Evoked Potential (P-VEP) for classification of squint risk levels has been the subject of much research. Techniques used are ranged from statistical methods to syntactic and knowledge based approaches. All of these methods require the definition of a set of features (or symbols and tokens) to be detected, and a pattern matcher to compare the observed values with the ideal, prototypical ones. An alternative approach, inspired by the configuration of the human brain, involves the use of artificial neural networks and fuzzy inference system (ANFIS). One specific ANFIS architecture is Sugeno FIS and RBFN(Radial Base Function Neural Network) with five layers between the input and output nodes. Training ANFIS is achieved by generalized least mean square algorithm. This research focused on classification of squint risk levels from PVEP signals through ANFIS and Genetic Algorithm (GA). The GA is a type of natural evolutionary algorithm that models biological process to optimize highly complex cost functions by allowing a population composed of many individuals to evolve under specific rules to a state that maximizes the fitness. John Holland developed this method in 1975 [1]. Many researchers share the intuitions that if the space to be searched is large, is known not to be perfectly smooth and unimodal (i.e., consists of a single smooth 'hill'), or is not well understood, or if the fitness function is noisy, and if the task does not require a global optimum to be found, i.e., if quickly finding a sufficiently good solution is enough – a GA will have a good chance off being competitive with or surpassing other optimization methods [2]. A comparison of GA and ANFIS as a classification and optimization tools for bio medical engineers with a useful application of squint risk level classification is analyzed.

### 1.1 Back Ground:

Visual evoked potential (VEP) is an evoked electrophysiological potential that can be extracted, using signal averaging, from the electroencephalographic activity recorded at the scalp. The VEP can provide important diagnostic information regarding the functional integrity of the visual system. Pattern reversal is the preferred technique for most clinical purposes. The results of pattern reversal stimuli are less variable in wave form and timing than the results elicited by other stimuli. Diagnosis of presence of squint is rarely difficult, but objectively determining what the symptomatic patient sees can be challenging. The pattern reversal stimulus consists of black and white checks that change phase (i.e., black to white and white to black) abruptly and repeatedly at a specified number of reversals per second. There must be no overall change in the luminance of the screen. This requires equal numbers of light and dark elements in the display. Background luminance of screen and room should approximate the mean for onset/offset of each check.

When a patient is diagnosed with squint, the latent potential of vision improvement is very important when deciding on therapy. Recently, various attempts have been made to assess which factors present at the time of diagnosis reflect the final visual outcome after squint treatment.[28-33] It has been reported that pattern visual-evoked response acuity correlates with the best-corrected Snellen acuity in normal subjects.[33] [34] Increases in the amplitude on pattern visual evoked potential (P-VEP) appear to reflect vision improvement during squint treatment. [30] Among patients with squint, strabismus amblyopia, those with an eccentric fixation had a relatively delayed P100 latency and less vision improvement after 6 months of squint treatment when compared with patients who had a central fixation [28].

To investigate whether P100 latency could predict visual outcomes in patients with functional squint including, patients were grouped by P100 latency on P-VEP at the time of initial diagnosis, and visual improvement was compared after occlusion therapy between the two groups. Also, differences in P100 latency by type of amblyopia were sought.

## 2. Methodology

10 patients whose visual abnormalities could not be explained by the findings of ophthalmological, neurological and psychiatric examinations were included. Control group comprised of 24 age and gender matched normal volunteers. Examinations of patients and subjects included visual acuities with Landolt rings, tests of pupillary reaction to light, visual fields tests looking for signs of tubular constriction, ophthalmoscopic examination and presence of squint eye tests.

P-VEP was always performed with appropriate refractive correction. This investigation was performed according to the standard guidelines after informed consent was obtained from all subjects. For P-VEP recording, each subject viewed a white and black checkerboard pattern on a television monitor. One experimenter monitored the patients' ocular fixation, which was directed toward the TV screen in a shielded room as a monocular P-VEP was recorded. The check sizes were $1^o$, 30`and 15`. Visual acuity of 0.2 corresponded to $1^o$ pattern, 0.4 corresponded to the 30` pattern and 0.7 to the 15` pattern.

The checks were reversed at 0.7 Hz. The computer analysis time of the P-VEP was 512milliseconds. One hundred P-VEP responses were averaged per session. The latency and amplitude of P100 for 3 consecutive check size were measured in both groups. The P100 component was used to estimate objective visual acuity.

A paper record of 16 channel P-VEP data is acquired from a clinical P-VEP monitoring system through 10-20 international electrode placing method. The PVEP signal was band pass filtered between 0.5 Hz and 50Hz using five pole analog Butter worth filters to remove the artifacts. With an P-VEP signal free of artifacts, a reasonably accurate detection of squint is possible; however, difficulties arise with artifacts. This problem increases the number of false detection that commonly plagues all classification systems. With the help of neurologist, we had selected artifact free PVEP records with distinct features. These records were scanned by Umax 6696 scanner with a resolution of 600dpi. Since the EEG records are over a continuous duration of about thirty seconds, they are divided into epochs of two second duration each by scanning into a bitmap image of size 400x100 pixels.

A two second epoch is long enough to detect any significant changes in activity and presence of artifacts and also short enough to avoid any repetition or redundancy in the signal [1] [2] [3]. The P-VEP signal has a maximum frequency of 50Hz and so, each epoch is sampled at a frequency of 200Hz using graphics programming in C. Each sample corresponds to the instantaneous amplitude values of the signal, totaling 400 values for an epoch. The different parameters used for quantification of the P-VEP are computed using these amplitude values by suitable programming codes.

The parameters are obtained for three different continuous epochs at discrete times in order to locate variations and differences in the presence of squint activity. We used ten P-VEP records for both training and testing. These P-VEP records had an average length of six seconds and total length of 60 seconds. The patients had an average age of 31 years. A total of 480 epochs of 2 seconds duration are used. General features of the test records are as follows.

Record 1and 4: High risk level with peaks and spikes.
Record 3 and6: Patient under clinical observation after two weeks of intensive drug therapy. Record 2and 8: Very High risk level with energy, Peaks and spikes.
Record 5and 7: Medium risk level with variance, energy, peaks and spikes.
Record 9and 10: Low risk level with variance, energy, peaks and spikes with occasional medium risk levels

### 2.1 Feature Extraction and Code Converter System

The various parameters obtained by sampling are given as inputs to the code converter system as shown in fig. 1. These parameters are defined as follows [9], [10], [11].

1. The energy in each two-second epoch is given by

$$E = \sum_{i=1}^{n} x_i^2 \quad \text{……(1)}$$

Where $x_i$ is signal sample value and n is number of samples. The normalized energy is taken by dividing the energy term by 1000.
2. The total number of positive and negative peaks exceeding a threshold is found.
3. Spikes are detected when the zero crossing duration of predominantly high amplitude peaks in the P-VEP waveform lies between 20 ms and 70 ms and sharp waves are detected when the duration lies between 70ms and 120ms.
4. The total numbers of spike and sharp waves in an epoch are recorded as events.
5. The variance is computed as σ given by

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n} \quad \text{……(2)}$$

Where $\mu = \frac{\sum_{i=1}^{n} x_i}{n}$ is the average amplitude of the epoch.

6. The average duration is given by

$$D = \frac{\sum_{i=1}^{p} t_i}{p} \quad \text{……(3)}$$

Where $t_i$ is one peak to peak duration and $p$ is the number of such durations.

7. Covariance of Duration which is defined as the variation of the average duration is

$$CD = \frac{\sum_{i=1}^{p} (D - t_i)^2}{pD^2} \quad \text{…….(4)}$$

8. P100 latency are calculated using standard deviation.

$$s = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \overline{x})^2}{N-1}} \quad \text{……..(5)}$$

In this formula, x is the value of the mean, N is the sample size, and $x_i$ represents each data value from i=1 to i=N.. The ∑ symbol indicates that you must add up the sum $(x_1 - x)^2 + (x_2 - x)^2 + (x_3 - x)^2 + (x_4 - x)^2 + (x_5 - x)^2. . . + (x_N - x)^2$.

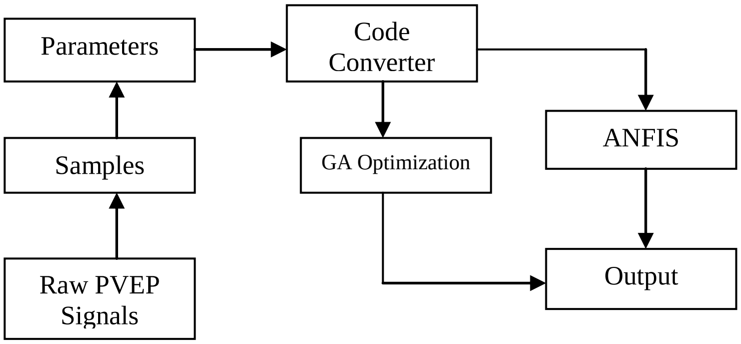


**Figure.1.**Block diagram of Genetic Algorithm and ANFIS based Classification system

The average values of extracted parameters in each 2 seconds epoch over sixteen channels of the patient record 5 is listed in the Table I

**Table I** Average values of Extracted Parameters from Patient Record 5

| Parameters | Epoch1 | Epoch2 | Epoch3 |
|---|---|---|---|
| Energy | 5.2869 | 8.581 | 10.10 |
| Variance | 1.1397 | 2.121 | 2.322 |
| Peaks | 1 | 2 | 2 |
| Total | 9 | 38 | 35 |
| Sharp & Spike | 8 | 6 | 6 |
| Total | 122 | 91 | 87 |
| Event | 12 | 10 | 10 |
| Total | 185 | 154 | 145 |
| Average duration | 3.798 | 4.042 | 3.883 |
| Covariance | 0.5793 | 0.5123 | 0.5941 |
| P100 latency(SD) | 0.04 | 0.05 | 0.03 |

With the help of expert's knowledge and our experiences with the references [12],[13],[14], we have identified the following parametric ranges for five linguistic risk levels (very low, low, medium, high and very high) in the clinical description for the patients which is shown in table II

**Table II** Parameter Ranges for Various Risk Levels

| Risk levels<br>Normalized Parameters | Normal | Low | Medium | High | Very High |
|---|---|---|---|---|---|
| Energy | 0-1 | 0.7-3.6 | 2.9-8.2 | 7.6-11 | 9.2-30 |
| Variance | 0-0.3 | 0.15-0.45 | 0.4-2,2 | 1,6-4.3 | 3.8-10 |
| Peaks | 0-2 | 1-4 | 3-8 | 6-16 | 12-20 |
| Events | 0-2 | 1-5 | 4-10 | 7-16 | 15-28 |
| Sharp Waves | 0-2 | 1-5 | 4-8 | 7-11 | 10-12 |
| Average Duration | 0-0.3 | 0.15-0.45 | 0.4-2.4 | 1.8-4.6 | 3.6-10 |
| Covariance | 0-0.05 | 0.025-0.1 | 0.09-0.04 | 0.28-0.64 | 0.54-1 |
| P100 latency (msec) | 120 | < 120 | 120 - 130 | 130 - 140 | > 140 |

The output of code converter is encoded into the strings of seven codes corresponding to each P-VEP signal parameter based on the squint risk levels threshold values as set in the table II. The expert defined threshold values are containing noise in the form of overlapping ranges. Therefore we have encoded the patient risk level into the next level of risk instead of a lower level. Likewise, if the P100 latency is 130 – 140 msec then the code converter output is High risk level instead of Normal level [12].

### 2.2 Code Converter as a Pre Classifier

The encoding method processes the sampled output values as individual code. Since working on definite alphabets is easier than processing numbers with large decimal accuracy, we encode the outputs as a string of alphabets. The alphabetical representation of the five classifications of the outputs is shown in table III.

**Table III** Representation of Risk Level Classifications

| Risk Level | Representation |
|---|---|
| Normal | N |
| Low | L |
| Medium | M |
| High | H |
| Very High | V |

By encoding each risk level one of the five states, a string of seven characters is obtained for each of the sixteen channels of each epoch. A sample output with actual patient readings is shown in fig. 2 for eight channels over three epochs. It can be seen that the Channel 1 shows low risk levels while channel 7 shows high risk levels. Also, the risk level classification varies between adjacent epochs**.** There are sixteen different channels for input to the system at three epochs.

This gives a total of forty-eight input output pairs. Since we deal with known cases of patients, it is necessary to find the exact level of squint risk in the patient. This will also aid towards the development of automated systems that can precisely classify the risk level of the squint patient under observation. Hence an optimization is necessary. This will improve the classification of the patient and can provide the P-VEP with a clear picture [15].

The outputs from each epoch are not identical and are varying in condition such as [HHVMMMM] to [LHVHHHH] to [HHVVHHH]. In this case energy factor is predominant and this results in the high risk level for two epochs and low risk level for middle epoch. Channel five and six settles at high risk level. Due to this type of mixed state output we cannot come to proper conclusion, therefore we group four adjacent channels and optimize the risk level. The frequently repeated patterns show the average risk level of the group channels. Same individual patterns depict the constant risk level associated in a particular epoch. Whether a group of channel is at the high risk level or not is identified by the occurrences of at least one V pattern in an epoch.

| Epoch 1 | Epoch 2 | Epoch 3 |
|---|---|---|
| LHHLHHH | LHHLHHH | LVHHLLL |
| HVVHMMM | HHHHMMM | HHHMHHH |
| HHVMHHH | HHHHHHH | HHHHHHH |
| HVVHMHH | MVVMHHH | HHHHHHH |
| VVVHHHH | LHHHMMM | HHHMHHH |
| HHVMMMM | LHVHHHH | HVVHHHH |
| VVVHHHH | HHHHHHH | VVVHHHH |
| HHHHMMM | HHHHMMM | HHHMVHH |

**Figure. 2.** Code Converters Output

The Code converter‘s classification efficiency is evaluated from the following parameters. The Performance of the Code converter is defined as follows [6],

$$PI = \frac{PC - MC - FA}{PC} \times 100 \qquad (6)$$

Where PC – Perfect Classification, MC – Missed Classification, FA – False Alarm. The Performance of code converter is 40%. The perfect classification represents when the physician agrees with the epilepsy risk level. Missed classification represents a High level as Low level. False alarm represents a Low level as High level with respect to physician's diagnosis. The sensitivity *Se* and specificity *Sp* are defined as [17],

Se= [PC/(PC+FA)]*100 ……(7)

(0.5/0.6)*100=83.33%

Sp= [PC/(PC+MC)] *100…..(8)

(0.5/0.7)*100=71.42%

Due to the low values of performance index, sensitivity and specificity it is essential to optimize the out put of the code converter. In the following section we discuss about the GA based optimization of squint risk levels.

## 3. Genetic Algorithm

Genetic algorithm is a population-based search method. The general scheme of a GA can be given as follows:

```
begin
  INITIALIZE population with random candidate solutions;
  EVALUATE each candidate;
  repeat
    SELECT parents;
    RECOMBINE pairs of parents;
    MUTATE the resulting children;
    EVALUATE children;
    SELECT individuals for the next generation
    until TERMINATION-CONDITION is satisfied
end
```

It's clear that this scheme falls in the category of generate-and-test algorithms. The evaluation function represents a heuristic estimation of solution quality and the search process is driven by the variation and the selection operator. GA has a number of features:

- GA is population-based
- GA uses recombination to mix information of candidate solutions into a new one.
- GA is stochastic.

The most important components in a GA consist of:

- representation (definition of individuals)
- evaluation function (or fitness function)
- population
- parent selection mechanism
- variation operators (crossover and mutation)
- survivor selection mechanism (replacement)

The complex and conflicting problems that required simultaneous solutions, which in past were considered deadlocked problems, can now be obtained with GA. However, the GA is not considered a mathematically guided algorithm. The optima obtained are evolved from generation to generation without stringent mathematical formulation such as the traditional gradient–type of optimizing procedure. In fact GA is much different in that context. It is merely a stochastic, discrete event and a non linear process. The obtained optima are an end product containing the best elements of previous generations where the attributes of a stronger individual tend to be carried forward into the following generation. The rule of the game is "survival of the fittest will win" [3].

A simple genetic algorithm can be summed up in seven steps as follows [16]:

1. Start with a randomly generated population of n chromosomes

2. Calculate fitness of each chromosome

3. Select a pair of parent chromosomes from the initial population

4. With a probability *Pcross* (the 'crossover probability' of the 'crossover rate'), perform crossover to produce two offspring

5. Mutate the two offspring with a probability *Pmut* (the mutation probability)

6. Replace the offspring in the population

7. Check for termination or go to step 2

Each iteration of the above steps is called a generation. The termination condition is usually a fixed number of generations typically anywhere from 50 to 500 or more. Under certain other circumstances, a check for global minimum is done after each generation and the algorithm is terminated as and when it is reached [4].The encoded genetic algorithm is a type of genetic algorithm that works with a finite parameter space. This characteristic makes it ideal in optimizing a cost due to parameters that assume only finite number of values. In case of optimizing parameters that are continuous, quantization is applied. The chief aspect of this method is the representation of the parameter as strings of binary digits of 0 and 1. This composition allows simple crossover and mutation functions that can operate on the chromosomes.

### 3.1 Encoding

The five risk levels are encoded as V>H>M>L>N in binary strings of length five bits using weighted positional representation as shown in table IV. Encoding each output risk level gives us a string of seven chromosomes, the fitness of which is calculated as the sum of probabilities of the individual genes. For example, if the output of an epoch is encoded as VVHMLVV, its fitness would be 0.419352.

**Table IV.** Encoded Risk Levels

| **Risk Level** | **Code** | **Encoded String** | **Weight** | **Probability** |
|---|---|---|---|---|
| Very High | V | 10000 | 16/31 =0.51612 | 0.086021 |
| High | H | 01000 | 8/31=0.25806 | 0.043011 |
| Medium | M | 00100 | 4/31=0.12903 | 0.021505 |
| Low | L | 00010 | 2/31=0.06451 | 0.010752 |
| Normal | N | 00001 | 1/31=0.03225 | 0.005376 |
| | | 11111 =31 | ∑ =1 | |

### 3.2 Operation on Data:

Using the above representation, we have developed a genetic algorithm that optimizes the output of the code converter and gives four risk level patterns in the five categories for each patient. This is obtained by the following procedure [16]

- Open three files having 16 strings each and process stage 1
- Divide into sets of 4 strings and iterate

1) Maximum of 128 generations
2) Two strings selected randomly
3) Single point crossover after 3rd position with probability Pcross = 0.75
4) Random mutation of any position to any state in the offspring with lower fitness and probability Pmut = 0.150535 which is the probability of XXXXXXX
5) Best two strings with higher fitness get selected for next stage

- Stage 2 operates on 24 chromosomes with 8 from each epoch
- Divide into sets of 4 strings and iterate in same way as stage 1
- Output of stage 2 is 4 best strings in each epoch
- Final stage is row-wise optimization in which each row of the epochs are iterated and one best output is taken
- Last iteration involving string of each row gives the final 4 output strings

By the application of the above procedure, the 48 risk level patterns obtained by the code converter are reduced to 4 risk level patterns, which define that of the patient. This process for a single patient is shown in table V.

**Table V.** Optimization by Encoding GA

| EPOCH 3 | | |
|---|---|---|
| HHHMHVV | | |
| HHHMHVV | | |
| VHHHVVH | | |
| HHHMMVH | HHHHVVH | |
| HHHHHVH | HHHMHVV | |
| HHHMHHH | HHHVVVH | VVHVHVH |
| HHHHHHH | VVHLHHH | HHHVVVH |
| VVHVVVV | VHHVHHH | HHHVVVH |
| HHHHHHH | HHHVHHH | VVHHHHV |
| HHHMHHH | VVHVVVV | |
| HHHHHHH | HHHVVVH | |
| VHHHHHH | | |
| HHHHHHH | | |
| HHHMMM | | |
| VVHVVVV | | |
| VVHVVVH | | |

**Final String for all epochs**

| Epoch 1 | Epoch 2 | Epoch 3 | Epoch 4 |
|---|---|---|---|
| VHHVHVV | VHHVVVV | VVHVHVH | VVVHVVV |
| HHHVVHH | HHHVVVV | HHHVVVH | HHHHVVV |
| HHHMVVV | VVHMVVV | HHHVVVH | HVHHLLV |
| HHHHHVV | VVHHMVV | VVHHHHV | VVVHHLL |

From the table V, each epoch is first reduced to 4 strings, which give the optimized risk levels of the epoch. An operation on the 12 strings in the final stage by a row-wise optimization gives the final 4 strings, representing the risk levels of the patient.

The drawback in this optimization as evident from the table V is that even though there are lower risk level states in the intermediate stage, they get omitted while proceeding to the final stage. This is because the algorithm takes only the higher fitness strings, which are the strings that represent the higher risk levels. Since we deal with only known cases of epilepsy, it can be stated that this is not a disadvantage, as those states will result in false alarms, which are defined later. It can also be inferred from the table IV that the mutation taking place in the initial stages affects the final result in only a small extent. Also, the final four strings which are obtained as the risk levels of the patient matches with the initial strings to a large extent. These advantages of the algorithm outline its use for the optimization of the risk levels of squint. The optimization of squint risk levels using ANFIS network is analyzed in the following section of the paper.

## 4. Adaptive Neuro Fuzzy Inference System for Risk Level Optimization

In this paper, for the classification method the ANFIS algorithm used in order to classify the trial signals into signals coming out. ANFIS's network organizes two parts like fuzzy systems. The first part is the antecedent part and the second part is the conclusion part, which are connected to each other by rules in network form. If ANFIS in network structure is shown, that is demonstrated in five layers. It can be described as a multi-layered neural network as shown in Figure (4). Where, the first layer executes a fuzzification process, the second layer executes the fuzzy AND of the antecedent part of the fuzzy rules, the third layer normalizes the membership functions (MFs), the fourth layer executes the consequent part of the fuzzy rules, and finally the last layer computes the output of fuzzy system by summing up the outputs of layer fourth. Here for ANFIS structure (fig. (4)) two inputs and two labels for each input are considered. The feed forward equations of ANFIS are as follows [27]:

$$w_i = \mu_{A_i}(x) \times \mu_{B_i}(y), \quad i = 1,2$$

$$\overline{w_i} = \frac{w_i}{w_1 + w_2}, \quad i = 1,2$$

$$f = \frac{w_1 f_1 + w_2 f_2}{w_1 + w_2} = \overline{w_1} f_1 + \overline{w_2} f_2$$

Where $f_1 = p_1, \quad f_2 = p_2$

(10).

In order to model complex nonlinear systems, the ANFIS model carries out input space partitioning that splits the input space into many local regions from which simple local models (linear functions or even adjustable coefficients) are employed. The ANFIS uses fuzzy MFs for splitting each input dimension; the input space is covered by MFs with overlapping that means several local regions can be activated simultaneously by a single input. As simple local models are adopted in ANFIS model, the ANFIS approximation ability will depend on the resolution of the input space partitioning, which is determined by the number of MFs in ANFIS and the number of layers. Usually MFs are used as bells shaped with maximum equal to 1 and minimum equal to 0 such as [27]:

$$\mu_{A_i}(x) = \frac{1}{1 + \left[\left(\frac{x - c_i}{a_i}\right)^2\right]^{b_i}}$$

$$\mu_{A_i}(x) = \exp\left\{-\left[\left(\frac{x - c_i}{a_i}\right)^2\right]^{b_i}\right\} \qquad \text{....(11)}$$

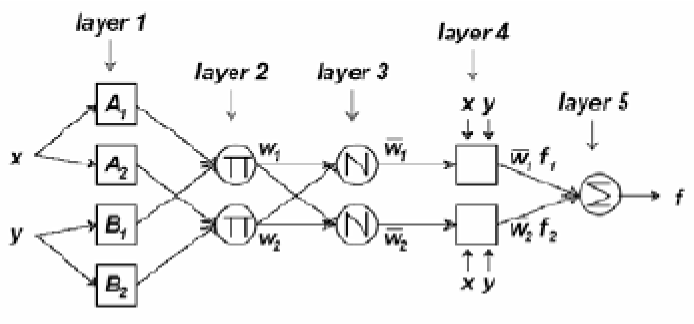


**Figure (3):** The equivalent ANFIS (type-3 ANFIS)

After applying the methodology and running the classification algorithm for 3 iterations, it reached the minimum RMSE value after the second epoch. The classification algorithm task is to classify and distinguish between the signals that are coming out. The primary aim of developing an ANN is to generalize the features (squint risk level) of the processed code converters outputs. We have applied different architectures of ANFIS networks for optimization. The simulations were realized by employing Neural Simulator 4.0 of Matlab v.7.0 [24]. Since our neural network model is patient specific in nature, we are applying 48 (3x16) patterns for each ANFIS model. There are ten models for ten patients. As the number of patterns in each database for training is limited, each model is trained with one set of patterns (16) for zero mean square error condition and tested with other two sets of patterns (2x16).

After network is trained using these, the classification performance of test set is recorded. The testing process is monitored by the Mean Square Error (MSE) which is defined as [19]

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(O_i - T_j)^2 \qquad \text{.....(12)}$$

Where Oi is the observed value at time i, Tj is the target value at model j; j=1-10, and N is the total number of observations per epoch and in our case, it is 16. As the number of hidden units is gradually increased from its initial value, the minimum MSE on the testing set begins to decrease.

The optimal number of hidden units is that number for which the lowest MSE is achieved. If the number of hidden units is increased beyond this performance does not improve and soon begins to deteriorate as the complexity of the neural network model is increased beyond that which is required for the problem. Multilayer perceptrons (MLPs) are feed forward neural networks trained with the standard back propagation algorithm [20].

To reduce the training time, an advanced NN training algorithm, called Levenberg-Marquardt (LM) is used. This training algorithm is based on the Gauss-Newton method, and it reduces the training time dramatically. It provides a fast convergence, it is robust and simple to implement, and it is not necessary for the user to initialize any strange design parameters. It out performs simple gradient descent and other conjugate gradient methods in a wide variety of problems [21].

## 5. Results and Discussions:

The outputs are obtained for three epochs for every patient in classifying the squint risk level by the code converter, Genetic algorithm, and ANFIS approaches. To study the relative performance of these systems, we measure two parameters, the Performance Index and the Quality Value. These parameters are calculated for each set of the patient and compared.

### 5.1 Performance Index

The PI calculated for the classification methods are illustrated in table VII using (5)

**Table VII.** Performance Index

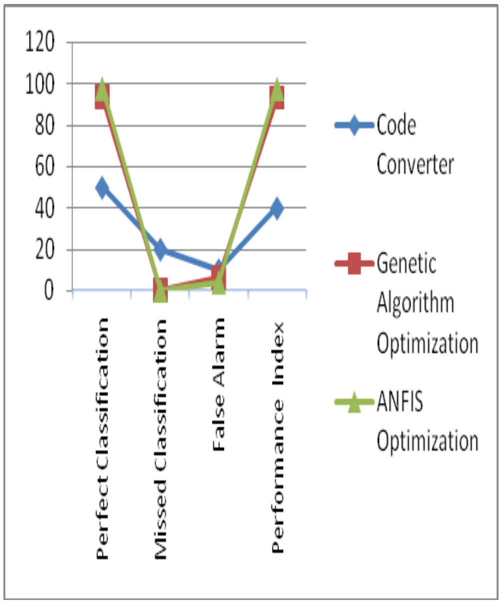


**Figure 4:** Performance Index

It is evident that the optimizations give a better performance than the code converter techniques due to its lower false alarms and missed classifications. For code converter classifier we have max detection of 50% with false alarm of 10% .Similarly for GA and ANFIS optimizations we obtained perfect detections of 93.75%and 97.83% with false alarms of 6.25% and 4.16%. This shows that the GA and ANFIS classifiers are performing better than the single code converter classifier.

### 5.2 Quality Value

The goal of this paper is to classify the squint risk level with as many perfect classifications and as few false alarms as possible. In Order to compare different classifiers we need a measure that reflects the overall quality of the classifier [15].Their quality is determined by three factors, Classification rate, Classification delay, and False Alarm rate. The quality value QV is defined as [5],

$$Q_V = \frac{C}{\left(R_{fa}+0.2\right)*\left(T_{dly}*P_{dct}+6*P_{msd}\right)}$$

Where, C is the scaling constant,
$R_{fa}$ is the number of false alarm per set
$T_{dly}$ is the average delay of the on set classification in seconds
$P_{dct}$ is the percentage of perfect classification and
$P_{msd}$ is the percentage of perfect risk level missed.

A constant C is empirically set to 10 because this scale is the value of QV to an easy reading range. The higher value of QV, the better the classifier among the different classifier, the classifier with the highest QV should be the best. The two different approaches give different results. Hence a comparative study is needed whereby the advantages of one over the other can be easily validated and the best method found out. A study of code converter method without and with GA optimization was performed and their results were taken as the average of all ten known patients was tabulated in table VIII.

| Methods | Perfect Classification | Missed Classification | False Alarm | Performance Index |
|---|---|---|---|---|
| Code Converter | 50 | 20 | 10 | 40 |
| Genetic Algorithm Optimization | 93.75 | 0 | 6.25 | 93.33 |
| ANFIS Optimization | 97.83 | 0 | 4.16 | 97.65 |

**Table VIII.** Results of Classifiers Taken As Average of All Ten Patients

| Parameters | Code Converter Method | Genetic Algorithm | ANFIS |
|---|---|---|---|
| Risk level Classification rate (%) | 50 | 92.75 | 97.83 |
| Weighted delay(s) | 4 | 0.482 | 0.463 |
| False-alarm rate/set | 0.4 | 0.0635 | 0.0416 |
| Performance Index(%) | 30 | 94.33 | 97.65 |
| Quality value | 5.25 | 19.14 | 24.59 |

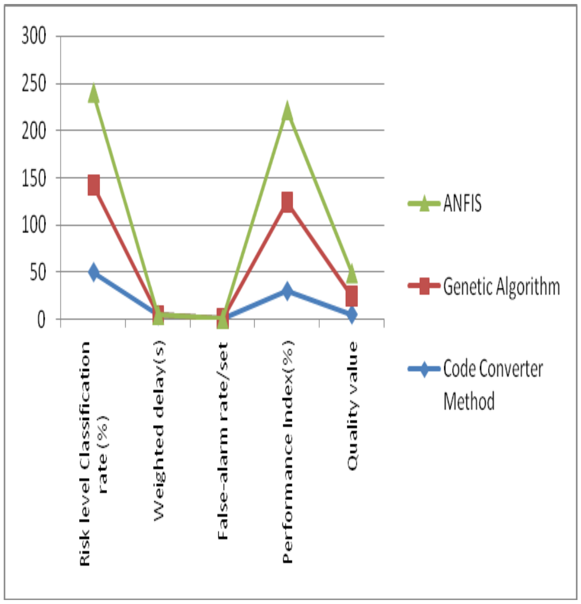


**Figure 5:** Classifiers Output

## 6. Conclusion:

This paper aims at classifying the squint risk level of squint patients from P-VEP signals. The parameters derived from the P-VEP signal are stored as data sets. Then the code converter technique is used to obtain the risk level from each epoch at every P-VEP channel. The goal was to classify perfect risk levels with high rate of classification, a short delay from onset, and a low false alarm rate. Though it is impossible to obtain a perfect performance in all these conditions, some compromises have been made. As a high false alarm rate ruins the effectiveness of the system, a low false-alarm rate is most important. Genetic algorithm and Adaptive neuro fuzzy inference system (ANFIS) optimization techniques are used to optimize the risk level by incorporating the above goals. The spatial region of normal P-VEP is easily identified in this classification method. The major limitation of GA method is that if one channel has a high-risk level, then the entire group will be maximized to that risk level. This will affect the non-squint spike region in the groups and for ANFIS its additional training cost involves in the learning procedures of the network. However, the classification rate of squint risk level of above 90% is possible in this method. The missed classification is almost 0% for a short delay of 2 seconds. From this method the inference is occurrence of High-risk level frequency and the possible medication to the patients. Also optimizing each region's data separately can solve the focal problem.

## References

[1] K.F. Man , "Genetic Algorithms: Concepts and Applications",*IEEE Transactions on Industrial Electronics*, vol,43no,5 October1996, pp 519 – 533.

[2] Randy L. Haupt and Sue Ellen Haupt, *"Practical GeneticAlgorithms"*, John Wisely and Sons, 1998.

[3] D.E.Goldberg, *"Genetic Algorithms in Search, Optimization andMachine Learning"*, Reading, MA: Addison-Wesley, 1989.

[4] Melanie Mitchell, *"An Introduction to Genetic Algorithms"*, ABradford Book MIT Press, 1997.

[5] Alison A Dingle, "A Multistage system to detect epileptic formactivity in the EEG", *IEEE Transactions on BiomedicalEngineering, vol,* 40, no, 12, December 1993, pp 1260-1268.

[6] Mark van Gils, 'Signal processing in prolonged EEG recordingsduring intensive care", *IEEE EMB Magazine*, vol, 16,no,6,November/December 1997, pp 56-63.

[7] Haoqu and Jean Gotman, "A patient specific algorithm fordetection onset in long-term EEG monitoring-possible use aswarning device", *IEEE Transactions on Biomedical Engineering*
vol,, 44,no2, February 1997,pp 115-122.

[8] Arthur C Gayton, *"Text Book of Medical Physiology"*, PrismBooks Pvt. Ltd., Bangalore, 9th Edition, 1996.

[9] Seunghan Park , "TDAT Domain Analysis Tool for EEG Analysis",*IEEE Transactions on Biomedical Engineering*, vol,37,no,8,August 1990,pp 803-811.

[10] Yuhui Shi , "Implementation of Evolutionary Fuzzy systems", *IEEETransactions on Fuzzy Systems*, vol,7,no,2, April 1999, pp109-119.

[11] R.Neelaveni and G.Gurusamy, "EEG Signal Analysis Methods – AReview*", Proceedings of National System Conference NSC-98*, ,1998, pp. 355-361.

[12] R.Harikumar and B.Sabarish Narayanan, "Fuzzy Techniques for Classification of Epilepsy risk level from EEG Signals", *Proceedings of IEEE TENCON– 2003,* Bangalore, India, 14-17
October 2003, pp 209-213.

[13] R.Harikumar, "Epilepsy Detection From EEG Signals – A Statistical Approach", *Proceedings of ICECON'03*, NIT Trichy, 5 – 6 December 2003.

[14] P.Aravindan Bharathi and V.Beena, "Analysis of Statistical Tests and Fuzzy Techniques in Diagnosis of Epilepsy from EEG signals", *Proceedings of ICSCI 2004, Hyderabad*, 12 – 15 February 2004, pp 131 – 137.

[15] Ching-Hung Wang, "Integrating Fuzzy Knowledge by Genetic Algorithms", *IEEE Transactions on Evolutionary Computations*, vol,2,no,4, November 1998, pp 138 – 149.

[16] Marco Russo, "Fu Ge Ne Sys – A Fuzzy Genetic Neural System for Fuzzy Modeling", *IEEE Transactions on Fuzzy Systems*, vol,6,no,3 ,August 1998, pp 373 – 387.

[17] Rangaraj M. Rangayyan, *"Bio- Medical Signal Analysis A Case Study Approach"*, IEEE Press-John Wiley &sons Inc New York 2002.

[18] Horn, J.Goldberg, D.E and Deb,K. "*Long path problems"*. In Y.Davidor, H.P.Schwefel and R.Männer (Eds.), *Lecture Notes in Computer Science 866-Parallel Problem Solving from*
*Nature-PPSNIII, International Conference on Evolutionary Computation, The Third Conference on Parallel Problem Solving from Nature*, 1994,pp 149-158.

[19] Nurettin Acir etal., " Automatic Detection Of Epileptiform Events In EEG By A Three- Stage Procedure Based on Artificial Neural Networks ",*IEEE transaction on Bio Medical Engineering* ,vol.52,no.1,January 2005, pp 30-40.

[20] Drazen.S.etal., "Estimation of difficult –to- Measure process variables using neural networks", *Proceedings of IEEE MELECON-2004*, 2004, pp 37-390.

[21]. Moreno.L.etal., "Brain Maturation Estimation Using Neural Classifier," *IEEE Transaction of Bio Medical Engineering*, vol.42, no.2,1995, pp 428-432.

[22]. Tarassenko.L, Y.U.Khan, M.R.G.Holt, "Identification Of Inter-Ictal Spikes in the EEG Using Neural Network Analysis," *IEE Proceedings–Science Measurement Technology*,vol.145,no.6,1998, pp 270-278.

[23]. Yuan-chu Cheng, Wei-Min Qi,WeiYou Cai, "Dynamic Properties of Elman And Modified Elman Neural Network," *Proceedings of IEEE, First International Conference on Machine Learning And Cybernetics*, Beijing,2002,pp.637-640.

[24] H.Demuth and M.Beale, "*Neural Network Tool Box: User's Guide, Version 3.0*,"The Math Works, Inc., Natick, MA, 1998.

[25] Li Gang etal, "An Artificial –Intelligence Approach to ECG Analysis," *IEEE EMB Magazine*,vol 20, 2000, pp 95-100.

[26] Guoqiang Peter Zhang , "Neural Networks for Classification: A Survey," *IEEE Transaction on Systems, Man And Cybernetics-Part C*, vol.30 No.4, November 2000, pp 451- 462.

[27].Kim S., Kim Y., Sim K., Jeon H.,"On Developing an adaptive neural-fuzzy control system",
Proc.IEEE/RSJ, Conference on intelligent robots and systems Yokohama, Japan, pp. 950-957, 1993.

[28]. Iliakis E, Moschos M, Hontos N, et al. The prognostic value of visual evoked response latency in the treatment of amblyopia caused by strabismus. Doc Ophthalmol. 1996-1997;92:223–228. [PubMed]

[29] Ridder WH, 3rd, Rouse MW. Predicting potential acuities in amblyopes: predicting post-therapy acuity in amblyopse. Doc Ophthalmol. 2007;114:135–145. [PubMed]

[30] Oner A, Coskun M, Evereklioglu C, Dogan H. Pattern VEP is a useful technique in monitoring the effectiveness of occlusion therapy in amblyopic eyes under occlusion therapy. Doc Ophthalmol. 2004;109:223–227. [PubMed]

[31] Johansson B, Jakobsson P. Fourier analysis steady-state VEPs in pre-school children with and without normal binocularity. Doc Ophthalmol. 2006;112:13–22. [PubMed]

[32] Simon JW, Siegfried JB, Mills MD, et al. A new visual evoked potential system for vision screening in infants and young children. J AAPOS. 2004;8:549–554. [PubMed]

[33] Ohn YH, Katsumi O, Matsui Y, et al. Snellen visual acuity versus pattern reversal visual-evoked response acuity in clinical applications. Ophthalmic Res. 1994;26:240–252. [PubMed]

[34]. Jenkins TC, Douthwaite WA, Peedle JE. The VER as a predictor of normal visual acuity in the adult human eye. Ophthalmic Physiol Opt. 1985;5:441–449. [PubMed]

## Author Biographies

**First Author R.Kalaivaazhi** receives her B.E (ECE) degree from REC Trichy,Tamilnadu,India in 1996. She obtained her M.Tech (Computer Science and Engineering ) degree from SASTRA University,Thanjavur, Tamilnadu,India in 2002. She has 15 years of teaching experience at college level. She worked as faculty in the Department of ECE at oxford Engineering College, Trichy. Currently she is Assistant Professor in the department of IT at AAMEC, Kovilvenni. and she ia pursuing doctrol programme in Computer Vision under guidance of Dr.D.Kumar

**Second Author Dr.D.Kumar** receives his Master's degree M.Tech(Bio-Medical Engineering and Instrumentation) from IIT, Chennai, Tamilnadu, India. He obtained his PhD in Biomedical Engineering from IIT, Chennai, Tamilnadu, India. He visited many foreign countries like London, Russia, Singapore, Germany and USA. He has published many International and National Journals in his area. Presently he in Dean of Research at Periyar Maniyammai University, Vallam, Thanjavur, Tamilnadu, India

# Prime Number Cost Estimation Criterion

Kavita Agrawal [(1)], Dr. S. K. Bajpai[(2)]
[1] Deptt of CSE, IU, UP, India
[2] UPTU, UP, INDIA
B-64 sector H, Aliganj, Lucknow, UP, India kavitalucknow@gmail.com
H-27, sector I, Jankipuram, Lucknow, UP, India skbajpaiiet@hotmail.com

***Abstract**: Estimating* cost (effort) of software accurately in the beginning of the software development lifecycle is a difficult task. Function points can be calculated apriori and are independent of the development techniques and tools used. In this paper we have proposed an idea based on cost being a prime number. In our idea we are mapping cost (effort) to prime numbers using actual function points as the count for prime numbers i.e. given function points we calculate actual function points then map actual function points to prime numbers to get the cost and conversely given cost in prime number we can find the function points. Our paper will be useful for software development industry in general.

***Keywords:*** Cost estimation, Function point, Prime numbers, Software effort estimation.

## 1. Introduction

Apriori cost estimation for any software development is a difficult task [10], [12]. Software cost estimation has been proposed traditionally by many developers based on their experience such as Algorithmic model [6], Expert judgement, Analogy, Parkinson's view, Price to win, or on the basis of top down or bottom up considerations on one hand and using Function point on the other [7],[9]. All the above methods are better from one another depending upon the ground that has been used. Our aim in writing this paper is to establish a mathematical criterion for estimating cost of any project no matter what environment has been chosen except for traditional views used in Function points and reduce the cost error from approximately 102.4% [4] to 42%.

What we have observed is that Function point criterion can be mapped to prime numbers and vice versa remembering that Function point (FP) determines cost at requirement phase. In fact cost of any software developed ultimately turns out to be unique. What we find that most often cost of software is uniquely determined except for the point of view of the difference that one finds in Function points.

We tried to map cost on several other numbers[8] [11] which diverge to infinity and hold some parametric development such as Fibonacci numbers and others but found that they are not suitable for cost determination and we found that best suited mapping is done on prime numbers.

Before we propose the following preposition we explain the difference between FP and Actual FP which we use through out this paper without further explanation.

FP (denoted by X) is the Function points calculated at the requirement phase

Actual FP gives us the number of prime numbers less than equal to the cost taken as a prime number i.e. if cost is 7units (unit refers to man hours) then Actual FP is the number of prime numbers less than equal to 7 i.e. equal to 5 unit (unit refers to count) as there are 5 prime numbers less than equal to 7.

Now we propose the following preposition:

Given X as FP point at the requirement phase we calculate new FP point Y( we call this as Actual FP point) by the following equation:-

$$526.8776+2.237601X = Y \qquad \dots (1)$$

Then mapping this to $Y^{th}$ prime number which corresponds to the cost of the software e.g. suppose Y is 5 then cost of software will be $5^{th}$ prime number which will be 7. Thus $Y^{th}$ prime number is known.

Conversely, given cost as prime number we calculate Actual FP by counting number of prime numbers less than equal to the cost (for this we used a program in Java) . Having computed Actual FP (denoted by X) we compute FP point (denoted by Y) by the following equation:-

$$319.6800037 + 0.226114097X = Y \qquad \dots (2)$$

Our preposition above thus determines cost of software and having obtained cost we compute FP as well.

Our method apriori need the knowledge of how function points criterion has been established for evaluating cost estimation. We give brief account of this method and then used function points on the basis of the equations illustrated in the above preposition.

We have used most of the available data on cost estimation and found that above preposition develops the best way of finding cost estimation at the requirement phase itself. It is interesting to look into civil engineering project which estimate cost of any project at the requirement phase based on some mechanical engineering methods (stress & strain) for their completion and view that project after completion will be usable for certain number of years without any problem. We also view software cost estimation on the same background of civil engineering project.

Our method will be appreciated if judgement is drawn on some worked out project and their cost estimation. We will illustrate our view on this strong point only. We considered the data available and cost available and compare the evaluation based on our method, remembering that our cost will be unique as prime numbers are unique. We have given cost of software project apriori without the knowledge of the software, by simply converting prime number cost to match it to function points.

## 2- Function Points

We will illustrate function points as our method depends totally on this concept [1],[3]. Since theory of Function point is well known to the reader, we briefly account the necessary parts in the following.

The function point measure is done in three steps:

Count and classify the five user function points: external input types, external output types, logical internal files, external interface file types, external inquiry types. Each FP is classified and a Weight is associated with it [1] and Total unadjusted function point (UFP) is calculated.
adjust for processing complexity. The degree of influence of each of 14 general characteristics namely: Data communication, Distributed functions, Performance, Heavily used configuration, Transaction rate, Online data entry, End user efficiency, Online data update, Complex processing, Reusability, Installation ease, Operational ease, Multiple sites and Facilitate change, is taken on a scale of 0 to 5. Where 0 is no influence and 5 is maximum influence. All influences are summed (PC, processing complexity) and an adjustment factor is developed. Where Processing complexity adjustment (PCA)= 0.65+(0.01 * PC) [1].
Make the function point calculation.Thus Function Point (FP) = UFP*PCA

## 3. The Idea

**A.** We compute equation (1) using following algorithm 1 and then calculate the cost of the software using algorithm 2 on certain specific data given in figure 2 [4].
Using data [4]

Figure 2

| Project number | Actual MM | Function point |
|---|---|---|
| 1 | 287 | 1217 |
| 2 | 82.5 | 507 |
| 3 | 1107.31 | 2306 |
| 4 | 86.9 | 788 |
| 5 | 336.3 | 1337 |
| 6 | 84 | 421 |
| 7 | 23.2 | 100 |
| 8 | 130.3 | 993 |
| 9 | 116 | 1592 |
| 10 | 72 | 240 |
| 11 | 258.7 | 1611 |
| 12 | 230.7 | 789 |
| 13 | 157 | 690 |
| 14 | 246.9 | 1347 |
| 15 | 69.9 | 1044 |

MM=number of man months
(=152 working hours) [4]

Using data given in figure 2 for deriving equation 1 we use the following Algorithm 1:

Step 1: consider actual MM*152. Take its nearest prime and find the number of primes less than it.

Step 2: this number becomes the reverse FP.

Step 3: using linear regression we relate FP (denoted by X) and reverse prime (denoted by Y). In linear regression[13] Straight line equation is taken to be: ao+a1*X=Y;
Where ao= $(\Sigma Yi\Sigma Xi^2-\Sigma Xi\Sigma(XiYi))/(n\Sigma Xi^2-(\Sigma Xi)^2)$
And a1= $(n\Sigma XiYi-\Sigma Xi\Sigma Yi)/(n\Sigma Xi^2-(\Sigma Xi)^2)$
For our data ao=526.8776 and
a1=2.237601

Step 4: resulting equation is the equation (1) ie 526.8776+2.237601X = Y

Now we obtain cost estimation using Algorithm 2 as follows:

Step 1: Function point is calculated at the requirement phase.

Step 2: Actual function point(denoted by Y) is calculated using function points (denoted by X) calculated in step 1, in equation (1).

Step 3: Effort (cost) is calculated (estimated MM) as a prime number corresponding to actual function point (Yth prime number).

Step 4: Error % is calculate as
((estimated MM-actual MM*152)/actual MM*152)*100

Step 5: Average error is calculated taking absolute values.

Figure 3

| FP (X) | REVERSE FP | ACTUAL FP (cal Y USING EQUATION) | ESTIMATED MM (TAKING ACTUAL FP'S CORRESPONDING PRIME) |
|---|---|---|---|
| 1217 | 4546 | 3250.04 | 30059.00 |
| 507 | 1499 | 1661.34 | 14083.00 |
| 788 | 1571 | 2290.11 | 20261.00 |
| 1337 | 5231 | 3518.55 | 32801.00 |
| 421 | 1524 | 1468.91 | 12281.00 |
| 100 | 493 | 750.64 | 5693.00 |
| 993 | 2242 | 2748.82 | 24851.00 |
| 1592 | 2028 | 4089.14 | 38803.00 |
| 240 | 1330 | 1063.90 | 8527.00 |
| 1611 | 4141 | 4131.65 | 39229.00 |
| 789 | 3738 | 2292.34 | 20287.00 |
| 690 | 2654 | 2070.82 | 18059.00 |
| 1347 | 3974 | 3540.93 | 33023.00 |
| 1044 | 1296 | 2862.93 | 26021.00 |

FP(denoted by X) in figure 3 corresponds to the function point data in figure 2.

Reverse FP is the number of prime numbers less than equal to the nearest prime number to actual MM.

Actual FP (denoted by Y) is calculated using the equation (1) and using FP(denoted by X) as x value.

Estimated MM is a prime number where actual FP is the number of prime numbers less than equal to Estimated MM. Actual MM corresponds to Actual MM in figure 2

MM*152 is Actual MM multiplied to 152 to get effort in hours.

Error % is calculate as ((estimated MM-actual MM*152)/actual MM*152)*100 [figure 4]

**B .** We compute equation (2) using following algorithm 3 and then calculate the FP of the software using algorithm 4 on certain specific data given in figure 2 .Results are shown in figure 5 and figure 6.

Using data given in figure 2 for deriving equation 2 we use the following Algorithm 3:

Step 1: consider actual MM*152. Take its nearest prime and find the number of primes less than equal to it.

Figure 4

| Estimated MM USING CAL Y CORRESPONDING PRIME | ACTUAL MM | MM*152 | ERROR (absolute values) % |
|---|---|---|---|
| 30059.00 | 287 | 43624 | 31.10 |
| 14083.00 | 82.5 | 12540 | 12.30 |
| 20261.00 | 86.9 | 13208.8 | 53.39 |
| 32801.00 | 336.3 | 51117.6 | 35.83 |
| 12281.00 | 84 | 12768 | 3.81 |
| 5693.00 | 23.2 | 3526.4 | 61.44 |
| 24851.00 | 130.3 | 19805.6 | 25.47 |
| 38803.00 | 116 | 17632 | 120.07 |
| 8527.00 | 72 | 10944 | 22.09 |
| 39229.00 | 258.7 | 39322.4 | 0.24 |
| 20287.00 | 230.7 | 35066.4 | 42.15 |
| 18059.00 | 157 | 23864 | 24.33 |
| 33023.00 | 246.9 | 37528.8 | 12.01 |
| 26021.00 | 69.9 | 10624.8 | 144.91 |
|  |  | Average error | 42.08% |

Step 2: this number becomes the reverse FP.

Step 3: using linear regression we relate FP (denoted by Y) and reverse prime (denoted by X). In linear regression[13] Straight line equation is taken to be: bo+b1*X=Y;
Where bo= $(\Sigma Yi\Sigma Xi^2-\Sigma Xi\Sigma(XiYi))/(n\Sigma Xi^2-(\Sigma Xi)^2)$
And b1= $(n\Sigma XiYi-\Sigma Xi\Sigma Yi)/(n\Sigma Xi^2-(\Sigma Xi)^2)$
For our data bo=319.6800037 and
b1=0.226114097

Step 4: resulting equation is the equation (2) ie
319.6800037 + 0.226114097X = Y … (2)

Now we obtain Function point (FP) using Algorithm 4 as follows:

Step 1:Given cost as a prime number we compute actual FP i.e. we calculate Actual FP by counting number of prime numbers less than equal to the cost (for this we used a program in Java).

Step 2: calculate FP (denoted by Y) using equation (2) using Actual FP as X.

Step 3: Error percentage is calculated as
((cal FP - FP)/FP)*100

Step 4: Average error is calculated taking absolute values.

Figure 5

| FP | MM*152 | Reverse FP(actual FP) denoted by X |
|---|---|---|
| 1217 | 43624 | 4546 |
| 507 | 12540 | 1499 |
| 788 | 13208.8 | 1571 |
| 1337 | 51117.6 | 5231 |
| 421 | 12768 | 1524 |
| 100 | 3526.4 | 493 |
| 993 | 19805.6 | 2242 |
| 1592 | 17632 | 2028 |
| 240 | 10944 | 1330 |
| 1611 | 39322.4 | 4141 |
| 789 | 35066.4 | 3738 |
| 690 | 23864 | 2654 |
| 1347 | 37528.8 | 3974 |
| 1044 | 10624.8 | 1296 |

Figure 6

| FP (Y) | cal FP using equation (2) | error % |
|---|---|---|
| 1217 | 1347.594688 | 10.73 |
| 507 | 658.6250348 | 29.91 |
| 788 | 674.9052497 | 14.35 |
| 1337 | 1502.482844 | 12.38 |
| 421 | 664.2778872 | 57.79 |
| 100 | 431.1542534 | 331.15 |
| 993 | 826.6278087 | 16.75 |
| 1592 | 778.239392 | 51.12 |
| 240 | 620.4117524 | 158.50 |
| 1611 | 1256.018479 | 22.03 |
| 789 | 1164.894498 | 47.64 |
| 690 | 919.7868166 | 33.30 |
| 1347 | 1218.257424 | 9.56 |
| 1044 | 612.7238731 | 41.31 |
|  | avg error % | 59.75 |

## 4. Conclusion

The results obtained are quite encouraging. For the same data average error % using function points is 102.74%, for COCOMO basic it is 610.09%, for COCOMO intermediate it is 583.82% and for COCOMO detailed it is 607.85% [4]. Whereas in our idea we mapped cost to prime numbers and the average error % is only 42.08 which is better than in most of the commonly used techniques [3], [4]. Results based on our idea are better than the previously used techniques.

Our point of view is straight forward and will be advantageous in most of the cases. However Function points used may vary from one person to another performing the analysis, thus it contributes to the variations in the results [2], [5]. Which result will be optimal is still a point of consideration but we feel that if larger set of data is used the above result will be better.

Most interesting point is that no matter what project is taken its cost can be estimated using our method. Thus it will be very useful for software development industry.

## References:

[1] Allan J. Albrecht and J.E. Gaffney, “Software function, source lines of code, and development effort prediction: a software science validation”, IEEE transactions on software engineering, SE-9, 6, 639-648, 1983.

[2] B.A. Kitchenham, E. Mendes, G.H. Travassos, “Cross versus within-company cost estimation studies: A systematic review”, IEEE transactions on software engineering , vol. 33 , no. 5, May2007, pg 316-329.

[3] Chris F Kemerer, “Software project management, readings and cases” McGraw hill company, Irwin book team-1997.

[4] Chris F Kemerer, “An empirical validation of software cost estimation models,” Communications of ACM, v-30, no-5, pp 416-429, 1987.

[5] Charles R Symons, “Function point analysis: difficulties and improvement”, IEEE transactions of software engineering, 14, 1, 2-11; Jan 1988.

[6] Ian Sommerville, “Software Engineering”, 5th edition, Addison-Wesley

[7] Bob Hughes and M. Cotterell, “Software project management”, Tata Mcgraw-Hill, 3rd edition.

[8] K. Chandrasekharan, “Introduction to analytic number theory”, Springer-Verlag New York Inc. 1968.

[9] Richard Fairley, “Software engineering concepts’, Tata McGraw Hill

[10] S. Grimstad, M. Jorgensen, “A framework for analysis of software cost estimation accuracy”, ISESE’06, Brazil, copyright 2006 ACM , pg 58-65

[11]S.P.Tripathi, Kavita Agrawal, S.K. Bajpai, “A note on cost estimation based on prime numbers”, IJITKM , July-December 2010, vol III, no.2, pp 241-245.

[12] Walker Royce, “Software project management A unified framework”, Addison-Wesley.

[13] V. Rajaraman, “Computer oriented numerical methods”, PHI, second edition.

-----------------------------------------------------------

## Author Biographies

**Kavita Agrawal** is Mtech in computer science. She is assistant professor, department of computer science and engineering, Integral University, Lucknow, UP, India.

**Dr.S.K.Bajpai** is Ex. HOD and Professor, department of computer science and engineering, IET, Lucknow, UP, India

# Image Fusion Technique: For Restoration of Images

Prof. Vanisha. P. Vaidya, Prof. V.K. Shandilya

Sipna's college of Engg. And Technology, Amravati.
Corresponding Addresses

vanishavaidya_2007@rediffmail.com, vkshandilya@rediffmail.com

***Abstract**: With the recent rapid developments in the field of sensing technologies, multisensory systems have become a reality in a growing number of fields such as remote sensing, medical imaging, machine vision and the military applications for which they were first developed. The result of the use of these techniques is a great increase of the amount of data available.*

*Image fusion provides an effective way of reducing this increased volume of information while at the same time extracting and increasing all the useful information from the source images. The underlying idea used here is to fuse different views of the same image .For achieving this; first the image is segmented and then fused into a complete image. The fused image provides better information for human or machine perception as compared to any of the input images. A total variation norm based approach has been adopted to fuse the pixels of the noisy input images.*

*Better results can be obtained on several test images. The goal of image fusion hence achieved and gives better human perception.*

***Keywords**:* Image fusion, pixel-level fusion, total variation.

## 1. Introduction

With the recent rapid developments in the field of sensing technologies multisensory systems have become a reality in a growing number of fields such as remote sensing, medical imaging, machine vision and the military applications for which they were first developed. The result of the use of these techniques is a great increase of the amount of data available.

Image fusion provides an effective way of reducing this increased volume of information while at the same time extracting and increasing all the useful information from the source images. Fusion integrates redundant as well as complementary information present in input image in such a manner that the fused image describes the true source better than any of the individual images. The exploitation of redundant information improves accuracy and the reliability whereas integration of complementary information improves the interpretability of the image. Image fusion has been used extensively in various areas of image processing such as remote sensing, biomedical imaging, nondestructive evaluation etc. For example, in optical remote sensing, due to physical and technical constraints, some sensors provide excellent spectral information but inadequate spatial information about the scene. On the other hand, there are sensors that are good at capturing spatial information but which fail to capture spectral information reliably. Fusing these two types of data provides an image that has both the spatial and the spectral information. Therefore, only the fused image needs to be stored for subsequent analysis of the scene. Multi-sensor data often presents complementary information about the region surveyed, so image fusion provides an effective method to enable comparison and analysis of such data. The aim of image fusion, apart from reducing the amount of useless data, is to create new images that are more suitable for the purposes of human/machine perception, and for further image-processing tasks such as segmentation, object detection or target recognition in applications such as remote sensing and medical imaging.

The underlying idea used here is to fuse different views of same image. For achieving this; first the image is segmented and then fused into a complete image.

Segmentation is done by minimizing a convex energy functional based on weighted total variation leading to a global optimal solution. Each salient region provides an accurate figure, ground segmentation highlighting different parts of the image. These highly redundant results are combined into one composite segment by analyzing local segmentation certainty.

From the perspective of fusion, features of the observed images that are to be fused can be broadly categorized in the following three classes.

1) Common features: These are features that are present in all the images.

2) Complementary features: Features that are present only in one of the images are called complementary features.

3) Noise: Features that are random in nature and do not contain any relevant information are termed as noise.

The goal of image fusion is to extract information from input images and fused it such that the fused image provides better information for human or machine perception as compared to any of the input images.

## 2. Related work

There are several approaches to the pixel level fusion of spatially registered input images. Most of these methods have been developed for the fusion of stationary input images (such as multispectral satellite imagery). Due to the static nature of the input data, temporal aspects arising in the fusion process of image sequences, e.g. stability and consistency are not addressed.

A generic categorization of image fusion methods is the following:

- **Linear superposition** : This is most straightforward way to build a fused image of several input frames is performing the fusion as a weighted superposition of all input frames. the linear combination of all inputs in a pre-chosen color space (eg. R-G-B or H-

S-V), leading to a false color representation of the fused image.

- **Nonlinear methods** : An approach to image fusion is to build the fused image by the application of a simple nonlinear operator such as max or min. If in all input images the bright objects are of interest, a good choice is to compute the fused image by an pixel-by-pixel application of the maximum operator.
- **Optimization approaches** : In this approach to image fusion, the fusion task is expressed as a bayesian optimization problem.
- **Artificial neural networks** : By the fusion of different sensor signals in biological systems, many researchers have employed artificial neural networks in the process of pixel-level image fusion. Several researchers modeled this fusion process for the combination of multispectral imagery by a combination of several neural networks.
- **Image pyramids** : Image pyramids consist of multiresolution image analysis and as a model for the binocular fusion in human vision. A generic image pyramid is a sequence of images where each image is constructed by low pass filtering and subsampling from its predecessor.
- **Wavelet transform** : A signal analysis method similar to image pyramids is the discrete wavelet transform. The main difference is that while image pyramids lead to an over complete set of transform coefficients, the wavelet transform results in a nonredundant image representation.
- **Generic multi resolution fusion scheme** : The basic idea of the generic multiresolution fusion scheme is motivated by the fact that the human visual system is primary sensitive to local contrast changes, i.e. edges. The above methods doesn't gives satisfactory result so I proposed a new technique for image fusion.

## 3. **Proposed work and objectives**:

In this project we are proposing a system for pixel level fusion to fuse images acquired using multiple sensors. The goal of our theme provides an effective way of reducing this increased volume of information while at the same time extracting and increasing all the useful information from the source images. The aim of image fusion, apart from reducing the amount of data, is to create new images that are more suitable for the purposes of human / machine perception, and for further image-processing. A total variation norm based approach has been adopted to fuse the pixels of the noisy input images. The underlying idea is to fuse different views of same image.

The entire process of fusing an image is proposed as follows:

1. Step1: First take an image as an input.
2. Step2: Perform segmentation over the captured or input image.
3. Step3: While performing segmentation focus on different salient feature of the image.
4. Step4: Finally, fuse all the segments to form one composite image.

**Step1:** First take an image as an input. This input can be acquired with the help of image acquisition model:

**3.1** Image acquisition model

Let $f_0(x,y)$ be the true image, which is inspected by *n* different sensors and *f*1(*x*,*y*), *f*2(*x*,*y*), ..... ,$f_n(x,y)$ are the corresponding *n* measurements for *x, y* ∈ Ω. The local affine transform that relates the input pixel and the corresponding pixel in the measured images is given by

$fi(x,\ y)=\beta i(x,\ y)fo(x,\ y)\ +\ \eta_i(x;\ y);\ 1\ \le\ i\ \le\ n$ (1)

Here, $\beta i\ (x,\ y)$ and $\eta_i(x,\ y)$ are the gain and sensor noise, respectively, of the i$^{th}$ sensor at location (*x* , *y*).

The goal of fusion is to estimate *fo*(*x, y*) from

$fi(x, y),\ 1 \le i \le n.$

In many applications such as radar imaging and visual and IR imaging, the complementary as well as redundant information are available at the local level in the measured images. The main advantage of the local affine transform model is that it can relate this local information content in a mathematically convenient manner. For example, as an extreme case, two sensors *i* and *j* ($i \neq j$; $1 \le i,\ j \le n$) have complementary information at location (*x*,*y*) if $\beta i\ (x,\ y) \neq \beta j\ (x,\ y)$ and $\beta i\ (x,\ y)\ ,\beta j\ (x,\ y)$ € *{0,1}*. Similarly, these two sensors have redundant information if $\beta i\ (x, y) = \beta j\ (x, y)$

**Step2:** Perform segmentation over the captured or input image by using total variation algorithm.

3,2 Total variation norm for image fusion

In order to estimate *fo*(*x, y*) from eq. (1), we assume that $fo(x, y);\ fi\ (x, y) \ge 0$ ($1 \le i \le n$). This assumption is valid for many imaging devices such as digital cameras, IR cameras, etc. and does not limit the proposed algorithm in any way since data not satisfying this requirement (*i.e.*, with negative pixel values) can always be transformed using a simple linear transformation to make the pixel values positive. Furthermore, we also assume that sensor noise $\eta_1(x,\ y),\ \eta_i(x,\ y),\ldots,\ \eta_n(x,\ y)$ are zero mean random variables and are independent of each other. The standard deviation of $\eta_i(x,\ y)$ $\sigma_i$ is denoted as $\sigma_i$, and $\sigma_i$ is assumed to be known *a priori* and independent of spatial location (*x, y*).

$$\begin{pmatrix} f_1 \\ \vdots \\ f_n \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_n \end{pmatrix} f_o + \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_n \end{pmatrix}$$

$$\Rightarrow \mathbf{f} \ = \ \beta f_o + \eta$$

**Step3:** While performing segmentation focus on different salient feature of the image such as denoising and deconvolution.

**3.3** The Matlab package implements total variation (TV) based image denoising, deconvolution etc.

**Denoising**

The problem of image noise removal or denoising is, given a noisy image f: Ω→ R to estimate the clean underlying image

u. For (additive white) Gaussian noise, the degradation model describing the relationship between f and u is
$f= u+\eta$ ,
where $\eta$ is i.i.d. Gaussian distributed.
The tvdenoise command implements TV-based denoising:
u = tvdenoise(f, lambda)

**Deblurring (deconvolution)** The image deblurring problem is to recover u from a given blurry and noisy image f. For Gaussian noise, the degradation model is
$f=Ku+\eta$
where K is the blur operator. For simplicity, the tvreg package is limited to the easier case of deconvolution, where $Ku=\varphi*u$ with some point-spread function $\varphi$, and $\varphi$ is assumed to be known exactly.
The tvdeconv command implements TV-based deconvolution:
u = tvdeconv(f, lambda, psf)
It solves for u approximately equal to f*psf. Parameter lambda balances between deblurring accuracy and denoising, where smaller lambda implies stronger denoising (but at the cost of deblurring accuracy).

**Step4:** Finally, fuse all the segments to form one composite image.The total variation norm has been used in several image processing applications. In this project we propose to use total variation norm for image fusion. Better results can be obtained on several test images, and the performance assessment of the final fusion results also are evaluated by using several classical evaluation methods like Root Mean Square Error or Peak Signal to Noise Ratio.
The proposed fusion algorithm was applied to two different datasets: (i) medical imaging and (ii) aircraft navigation. For each dataset, only two input images were considered for the fusion process and these two inputs were co-registered.
The sensor noise will be simulated by adding zero mean white Gaussian noise to the input images. For ease of quantitative analysis of the fusion performance, the variance of the noise for each input image was selected appropriately to get the same level of signal-to-noise (SNR) ratio for all the input images, where the SNR will be computed using the following expression:
$$SNR=10\log_{10}\frac{Signal\ Variance}{Noise\ Variance}dB$$

## 4. Conclusion

The goal of image fusion is to compare the information content in the fused image and the corresponding input images. Therefore, the similarity index will be computed by comparing the fused images with the noiseless versions of the corresponding input images.
A total variation algorithm has been used for fusing an image. An output that will be generated by a fused image will generate much better results as compared to the original image.

## References

1. X. Bresson, S. Esedoglu, P. Vandergheynst, J. P. Thiran, and S. J. Osher. Fast global minimization of the active contour/snake model. *Journal of Mathematical Imaging and Vision*,
28(2):151–167, 2007.

2. Z. Zhang and R. Blum, "A categorization and study of multiscale-decomposition-based image fusion schemes," *Proceedings of the IEEE*, vol. 87, no. 8, pp. 1315–1326, August 1999.

3. A. Chambolle. An algorithm for total variation minimization and application. *Journal of Mathematical Imaging and Vision*, 20(1-2):89–97, 2004.

4. D. L. Hall and J. Llinas, "An introduction to multisensor data fusion," *Proc. IEEE*, vol. 85, no. 1, pp. 6–23, 1997.

5. T. Pock, M. Unger, D. Cremers, and H. Bischof. Fast and exact solution of total variation models on the gpu. In *CVPR Workshop on Visual Computer Vision on GPUs, 2008.*

6 . G. Simone, A. Farina, F. C. Morabito, S. B. Serpico and L. Bruzzone, Image Fusion Techniques for Remote Sensing Applications**,** *Information Fusion*, Volume 3, Issue 1, Pages 3-15, March 2002

7. A. Chambolle and P.L. Lions. /Image Recovery via Total Variation-Based Restoration." *SIAM J. Sci. Compute. 20,* pp. 1964{1977, 1999.

8. Eduardo Fernández Canga,Image Fusion, University of Batch, Signal & Image Processing Group Department Of Electronic & Electrical Engineering ,June 2002.

9. Peter J. Burt and Edward H. Adelson, The Laplacian Pyramid as a Compact Image Code**,** *IEEE Transactions On Communications*, Vol. Com-3l, No. 4, Pages 532-540, April 1983.

10. Mrityunjay Kumar , Pradeep Ramuhalli, " A Total Variation Based Algorithm for Pixel Level Image Fusion", *IEEE Transactions On Communications.*

## Author Biographies

**First Author** :Prof. V.P.Vaidya received her Bachelor of Computer Science & Engg in 2008 from Sipna's C.O.E.T Amravati (MS),Currently pursuing her M.E in Information Technology and working as a lecturer in Sipna's C.O.E.T,Amravati.

**Second Author:** Prof. V.K.Shandilya received her Bachelor of Computer Science & Engg in 1997 from P.R.M.I.T & R, Badnera (MS), Masters degree in Computer Science & Engg from PRMIT & R Badnera, Amravati in 2006, currently working as a Asst. Prof in Sipna's COET Amravati.

# Implementation of Advanced Encryption Standard using FPGA

Prof. Shilpa.A.Ingole[1,] Prof.V.T.Gaikwad[2]

Sipna's college of Engg. And Technology ,Amravati.
Corresponding Addresses
shilpa.ingole@gmail.com, vtgaikwad@rediffmail.com

***Abstract**: Advanced Encryption Standard (AES) algorithm was developed by Vincent Rijmen and Joan Daeman and named as Rijndael cipher algorithm. AES consists of 128 block length of bits and supports 128,192 and 256 key length bits and consists of 10, 12 or 14 iteration rounds, respectively. Each round mixes the data with a round key, which is generated from the encryption key. For security and fast transmission of data over an insecure path, cryptography methods have been used so far. For this reason, we are trying to implement secure, fast and efficient cryptographic algorithms in hardware. In this project we are implementing advanced encryption standard (AES) algorithm using a Field Programmable Gate Array (FPGA).*

*The objective is to implement an efficient realization of AES using very high speed integrated circuit hardware description language (VHDL). A fast and area efficient composite field implementation of the byte substitution phase is designed using an optimum number of stages for FPGA implementation. This seminar proposes an efficient solution to combine Rijndael encryption in one FPGA design, with a strong focus on low area constraints.*

*To ensure that our implementation gives a better result in terms of area and speed, we compare the two encryption codes, original and modified. This comparison is done by considering two criteria: chip speed and area utilization. The design will be implemented by using VHDL frontend and backend tools.*

***Keywords**: Cryptography, AES, DES, FPGA, compact encryption/decryption implementation, embedded Systems.*

## 1. Introduction

### 1.1 The AES algorithm

The Advanced Encryption Standard (AES, Rijndael) algorithm is a symmetric block cipher that processes data block of 128, 192 and 256 bits using, respectively,keys of the same length[3]. In this paper,only the 128 bit encryption version (AES-128) is considered. The 128-bit data block and key are considered as a byte array, respectively called State and RoundKey, with four rows and four columns.Let a 128-bit data block in the ith round be defined as:

**data_block** $^{I}$ $=d^{i}15|d^{i}14|d^{i}13|d^{i}12|d^{i}11|d^{i}10|d^{i}9|d^{i}8|d^{i}7|d^{i}6|d^{i}5|d^{i}4|d^{i}3|d^{i}2|d^{i}1|d^{i}\ 0$

where di 15 represents the most significant byte of the data block of the round i. The corresponding State is:

$$State^i = \begin{bmatrix} d^i_{15} & d^i_{11} & d^i_7 & d^i_3 \\ d^i_{14} & d^i_{10} & d^i_6 & d^i_2 \\ d^i_{13} & d^i_9 & d^i_5 & d^i_1 \\ d^i_{12} & d^i_8 & d^i_4 & d^i_0 \end{bmatrix}$$

AES-128 consists of ten rounds. One AES encryption round includes four transformations: SubByte,ShiftRow, MixColumn and AddRoundKey. The first and last rounds differ from the other ones Indeed there is an additional AddRoundKey transformation at the beginning of the first round and noMixColumn transformation is performed in the last round. This is done to facilitate the decryption process. SubByte (SB) is a non-linear byte substitution. It operates with every byte of the State separately. The substitution box (S-box) is invertible and consists of two transformations:

1. Multiplicative inverse in GF($2^8$). The zero element is mapped to itself.
2. An affine transform over GF($2^8$).

The SubByte transformation applied to the State can be represented as follows:

$$SB(State^i) = \begin{bmatrix} SB(d^i_{15}) & SB(d^i_{11}) & SB(d^i_7) & SB(d^i_3) \\ SB(d^i_{14}) & SB(d^i_{10}) & SB(d^i_6) & SB(d^i_2) \\ SB(d^i_{13}) & SB(d^i_9) & SB(d^i_5) & SB(d^i_1) \\ SB(d^i_{12}) & SB(d^i_8) & SB(d^i_4) & SB(d^i_0) \end{bmatrix}$$

The inverse transformation is defined InvSubByte (ISB). ShiftRow (SR) performs a cyclical left shift on the last three rows of the State. The second row is shifted of one byte, the third row is shifted of two bytes and the fourth row is shifted of three bytes. Thus, the ShiftRow transformation proceeds as follows:

$$SR(SB(State^i)) = \begin{bmatrix} SB(d^i_{15}) & SB(d^i_{11}) & SB(d^i_7) & SB(d^i_3) \\ SB(d^i_{10}) & SB(d^i_6) & SB(d^i_2) & SB(d^i_{14}) \\ SB(d^i_5) & SB(d^i_1) & SB(d^i_{13}) & SB(d^i_9) \\ SB(d^i_0) & SB(d^i_{12}) & SB(d^i_8) & SB(d^i_4) \end{bmatrix}$$

The inverse ShiftRow operation (InvShiftRow (ISR)) is trivial.MixColumn (MC) operates separately on every column of the State. A column is considered as a polynomial over GF($2^8$) and multiplied modulo x4+1 with the xored polynomial c(x): c(x) ='03'x3 +'01'x2 +'01'x +'02'

As an illustration, the multiplication by 0020 corresponds to a multiplication by two, modulo the irreductible polynomial m(x) = x8 + x4 + x3 + x + 1.

This can be represented as a matrix multiplication

$$R^i = MC(SR(SB(State^i))) = \begin{bmatrix} '02' & '03' & '01' & '01' \\ '01' & '02' & '03' & '01' \\ '01' & '01' & '02' & '03' \\ '03' & '01' & '01' & '02' \end{bmatrix} \otimes \begin{bmatrix} SB(d^i_{15}) & SB(d^i_{11}) & SB(d^i_7) & SB(d^i_3) \\ SB(d^i_{10}) & SB(d^i_6) & SB(d^i_2) & SB(d^i_{14}) \\ SB(d^i_5) & SB(d^i_1) & SB(d^i_{13}) & SB(d^i_9) \\ SB(d^i_0) & SB(d^i_{12}) & SB(d^i_8) & SB(d^i_4) \end{bmatrix}$$

To achieve the inverse operation (InvMixColumn (IMC)), every column is transformed by multiplying it with a specific multiplication polynomial d(x), de-fined by
c(x)  d(x) =`01`
d(x) =`0B`x3 +` 0D`x2 +’09`x +` 0E`
AddRoundKey (AK) performs an addition (bitwise XOR) of the State with the RoundKey:

$$AK(R^i) = \begin{bmatrix} R^i_{15} & R^i_{11} & R^i_7 & R^i_3 \\ R^i_{14} & R^i_{10} & R^i_6 & R^i_2 \\ R^i_{13} & R^i_9 & R^i_5 & R^i_1 \\ R^i_{12} & R^i_8 & R^i_4 & R^i_0 \end{bmatrix} \oplus$$

$$\begin{bmatrix} rk^i_{15} & rk^i_{11} & rk^i_7 & rk^i_3 \\ rk^i_{14} & rk^i_{10} & rk^i_6 & rk^i_2 \\ rk^i_{13} & rk^i_9 & rk^i_5 & rk^i_1 \\ rk^i_{12} & rk^i_8 & rk^i_4 & rk^i_0 \end{bmatrix}$$

The inverse operation (InvAddRoundKey
(IAK)) is trivial.
RoundKeys are calculated with the key schedule
for every AddRoundKey transformation. In AES-128, the original cipher key is the first RoundKey` (rk0) used in the additional AddRoundKey at the beginning of the first round. RoundKey, where 0 < i _ 10, is calculated from the previous
RoundKey 1. Let p(j) (0 _ j _ 3) be the column
j of the RoundKey 1 and let w(j) be the column
j of the RoundKey. Then the new RoundKey is
calculated as follows:
w(0) = p(0) _ (Rot(Sub(p(3))) _ rcon;
w(1) = p(1) _ w(0)
w(2) = p(2) _ w(1)
w(3) = p(3) _ w(2)
Rot is a function that takes a four byte input
[a0; a1; a2; a3] and rotates them as [a1; a2; a3; a0].The function Sub applies the substitution box (S-box) to four bytes. The round constant rconi contains values [(`02`)i1;`00`;`00`;`00`].

### 1.2 AES Decryption Algorithm

The Cipher transformation can be inverted to produce the Inverse Cipher which is the Decryption. The key Expansion remains the same for Encryption and Decryption if the data is decrypted by the Inverse Cipher. The AES encrypted data can be decrypted by an equivalent Inverse Cipher in which the transformation used is in the same sequence of the cipher whereas the sequence of transformation in the Key expansion changes. The Decryption generates the Plain text from the Cipher text. It also operates on 128 bits of Cipher text which is the output of the Cipher and uses the Key which is generated at the last iteration of the Cipher. The transformations used are

1. Inverse shift rows
2. Inverse Sub bytes Transformation
3. Inverse Mix Column transformation
4. Add round key Transformation

## 2. Related work:

In embedded applications, it is required to minimize the area rather than to maximize the throughput. Pramstaller et al presented a compact implementation of AES encryption and decryption with all key lengths using a novel State representation, which solves the problem of accessing both rows and columns of the State. Therefore, FPGAs are considered one of the integral part of the cryptographic hardware implementation in order to achieve further acceleration in throughput and efficient utilization of memory and other hardware resources.

Uinversity of Lethbridge [1] proposed a three stage sub pipelined architecture which is the first entry. The forte for this architecture is that it includes both encryption and decryption. The architecture is divided into the data module, the key generator module and the input/output module. The important block which implements the AES algorithm is the data module.

The S – box implemented is based on the combinational circuit proposed by J. Wolkerstorfer et al [2]. The column mixing is based on their own architecture put forth in The Key generator module also has 3 stages of pipelining and generates the same key for three consecutive cycles. The same architecture can also be extended to 192 and 256 bit key AES. The encoding rate is 1.57 Gbps and the latency is 30 cycles for 3 blocks of 128 bit key.

## 3. Proposed work and objectives:

The goal of this project is to compare c code with the FPGA output.

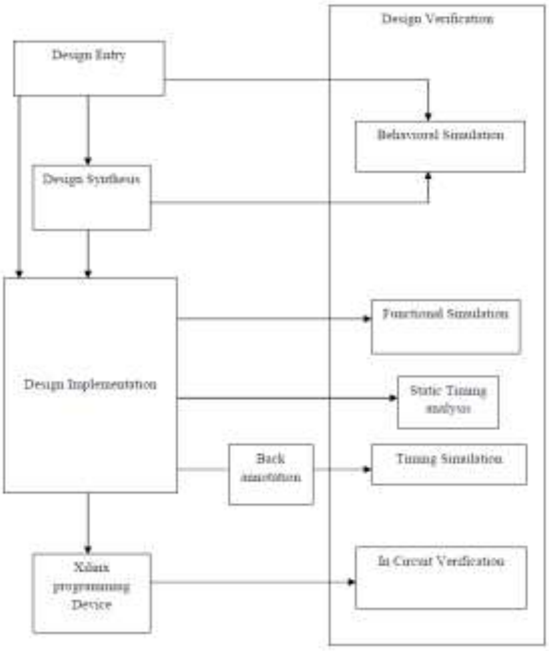


*Figure 1 Block diagram*

**3.1 Design Entry** The design entry for this project is basically the HDL codes for the AES

### 3.2 Design Synthesis and Implementation

The design is synthesized and implemented using Xilinx ISE. Once the design is placed and routed, the Post placed and routed design is simulated and the Output is cross checked with outputs after the simulation and the output from the C code. The same process is repeated for all the architectures of both the encryption and the Decryption module. The implementation is constrained for the maximum speed.

### 3.3 In Circuit verification

After the implementation of the Design a bit stream is generated. This bit stream programs the FPGA and makes the FPGA work as the AES. The circuit is synthesized and the bit stream is generated. This bit stream is used to program the FPGA. The Output is viewed verified using the C- code.

## 4. Performance Analysis:

The data input to the encryption module is (0000_0000_0000_0000_0000_0000_0000_0000) h and the key is also (0000_0000_0000_0000_0000_0000_0000_0000)h. The above data is encrypted to (66E9_4BD4_EF8A_2C3B_884C_FA59_CA34_2B2B)h. The encrypted data is also verified by the AES algorithm implemented in C. Figure 2: Online verification of basic AES (Encryption). Similarly the same data is sent as the input to the decryption module and the basic AES (decryption) is also verified. Figure 3 shows the online verification of Basic AES (decryption).

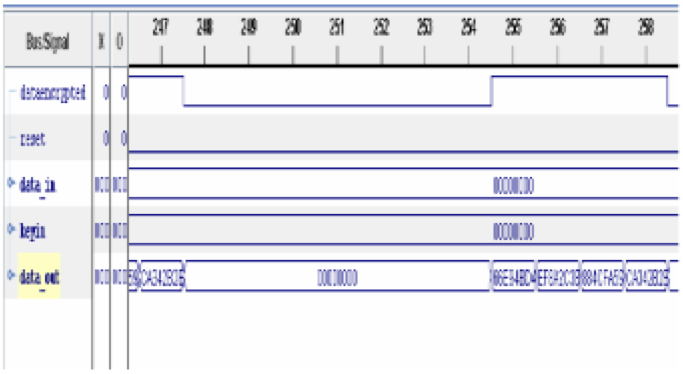


*Figure 2: Online verification of basic AES (Encryption).*

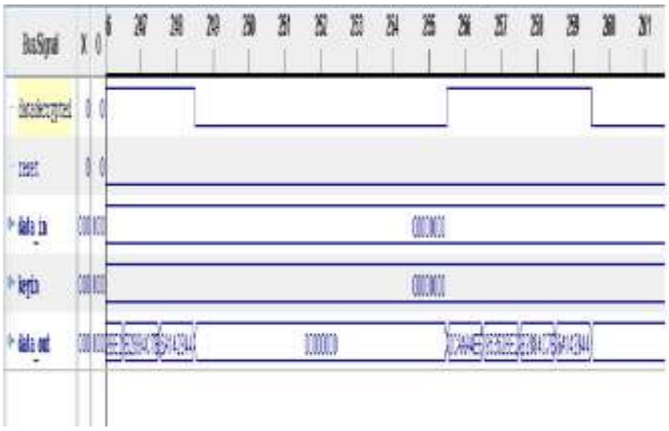


*Figure 3: The online verification of Basic AES (decryption).*

## 5.Conclusion

For the AES design, a software model of the AES algorithm was initially developed in C which would read a binary text file and then output the encoded bit stream into another binary text file. The same piece of code also decoded the encrypted data to plain text. This C code serves as a check for validating the behavioral output generated from the Simulation of the design. Once the results from the C code matches with the results in the behavioral simulation the design is further synthesized and checked for behavioral simulation again.. The Design is then implemented in a FPGA. The Xilinx development kit was used for verifying the designs. The design is then placed and routed on a FPGA.

## 6.References:

[1] "A High Performance Sub-Pipelined Architecture for AES" Hua Li and Jianzhou Li Department of Mathematics and Computer Science, University of Lethbridge, Canada T1K 3M4

[2] J. Wolkerstorfer, E. Oswald and M. Lamberger, "An ASIC Implementation of the AES SBoxes," CT-SA 2002, LNCS2271, pp. 67-78, 2002.

[3] National Institute of Standards and Technology: Specification for the Advanced Encryption Standard (AES). http://csrc.nist.gov/publications/fips/fips197/ fips-197.pdf, (2001).

[4] The Mathworks: Galois Field Computations. http://www.mathworks.com/ access/helpdesk/help/toolbox/comm/tutor3.shtml, Communications Toolbox,(2001)

## Author Biographies

**First Author** :Prof.S.A.Ingole received her Bachelor of Information technology in 2008 from GCOE0 Amravati,Currently pursuing her M.E in Information technology and working as a lecturer in Sipna's C.O.E.T, Amravati.

**Second Author:** Prof.V.T.Gaikwad received his Bachelor of Electronics & telecommunication in 1994 from SSGM College of engg, shegaon. Masters degree in Electronics from GCOE ,Amravati in 2001 ,Currently working as a Asst.Prof in sipna's COET Amravati.

# Security Constraints for Sequence Diagram and Code Generation

Abdeslam Jakimi[1,2] and Mohammed Elkoutbi[1]

[1]Mohammed V University, SIME, ENSIAS, Rabat, Morocco ,
[2]My Ismail University, FSTE, B.P 509, Boutalamin, Errachidia, Morocco
{ajakim@yahoo.fr}

***Abstract***: The Unified Modeling Language, which has become a standard notation for object-oriented modeling, provides a suitable framework for scenario acquisition using use case diagrams and sequence diagrams. A sequence diagrams shows the interactions among the objects participating in a scenario in temporal order. It depicts the objects by their lifelines and shows the messages they exchange in time sequence. In this paper, we suggest to offer the extension of scenarios that describe a given system in a natural way based directly on sequence diagrams. We developed algorithm and tool support that can automatically produce a code of sequence diagram with security constraints.

***Keywords***: UML, sequence diagrams, security constraints, code generation.

## 1. Introduction

The Unified Modeling Language (UML) [1,2,3] is an expressive language that can be used for problem conceptualization, software system specification as well as implementation. UML is a graphical language for specifying the analysis and design of object-oriented software systems [2]. It provides several diagram types that can be used to view and model the software system from different perspectives and/or at different levels of abstraction. UML defines thirteen types of graphical diagrams. The four diagrams which become important in the design phase are a class diagram, use case diagram, state chart diagram and sequence diagram.

The emergence of UML as a standard for modeling systems has encouraged the use of automated software tools that facilitate the development process from analysis through coding. In UML, the static structure of classes in a system is represented by a class diagram while the dynamic behavior of the classes is represented by a set of usecase, sequence and statechart diagrams. To facilitate the software development process, it would be ideal to have tools that automatically generate or help to generate executable code from the models.

In the present study, an effort has been made to find methods to automatically generate executable security code from the UML sequence diagram. We propose to extend the UML with the following message constraints: security constraints.

## 2. Scenarios and UML

### 2.1. Scenarios

Scenarios have been evolved according to several aspects, and their interpretation seems to depend on the context of use and the way in which they were acquired or generated. In a survey, Rolland [4] proposed a framework for the classification of scenarios according to four aspects: the form, contents, the goal and the cycle of development

The form view deals with the expression mode of a scenario. Are scenarios formally or informally described, in a static, animated or interactive form?

The contents view concerns the kind of knowledge which is expressed in a scenario. Scenarios can, for instance, focus on the description of the system functionality or they can describe a broader view in which the functionality is embedded into a larger business process with various stakeholders and resources bound to it.

The purpose view is used to capture the role that a scenario is aiming to play in the requirement's engineering process. Describing the functionality of a system, exploring design alternatives or explaining drawbacks or inefficiencies of a system are examples of roles that can be assigned to a scenario.

The lifecycle view considers scenarios as artefacts existing and evolving in time through the execution of operations during the requirement's engineering process. Creation, refinement or deletion are examples of such operations.

### 2.2. Scenarios in UML

Object-oriented analysis and design methods offer a good framework for scenarios. In our work, we have adopted the UML, which is a unified notation for Object-oriented analysis and design. It directly unifies the methods of Booch, Rumbaugh and Jacobson.

Scenarios and use cases have been used interchangeably in several works meaning partial descriptions. UML distinguishes between these terms and gives them a more precise definition. A use case is a generic description of an entire transaction involving several objects of the system. A use case diagram is more concerned with the interaction between the system and actors (objects outside the system that interact directly with it). It presents a collection of use cases and their corresponding external actors. A scenario shows a particular series of interactions among objects in a single execution of a use case of a system (execution instance of a use case). A scenario is defined as an instance of a given use case. Scenarios can be viewed in two different ways through sequence diagrams or communication diagrams. Both types of diagrams rely on the same underlying semantics. Conversion from one to the other is possible.

### 2.2.1 Class diagram

A class diagram is a graphic view of the static structural model. It shows a set of classes, interfaces and their relationships. The main focus is on the description of the classes. Class diagrams are important for constructing systems through forward engineering. The ClassD is the central diagram of a UML model. The translation of class diagrams to an Object-oriented programming language is easy and provided by most CASE tools.

### 2.2.2 Sequence diagram

For our work, we chose to use sequence diagrams because of their wide use in different domains. A sequence diagram shows interactions among a set of objects in temporal order, which is good for understanding timing and interaction issues. It depicts the objects by their lifelines and shows the messages they exchange in time sequence. However, it does not capture the associations among the objects.

A sequence diagrams has two dimensions: the vertical dimension represents the time, and the horizontal dimension represents the objects. Messages are shown as horizontal solid arrows from the lifeline of the object sender to the lifeline of the object receiver (Figure 2). A message may be guarded by a condition, annotated by iteration or concurrency information, and/or constrained by an expression. Each message can be labeled by a sequence number representing the nested procedural calling sequence throughout the scenario, and the message signature. Sequence numbers contain a list of sequence elements separated by dots.

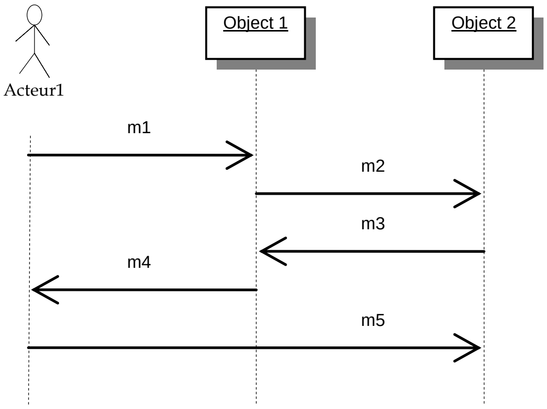


**Figure 2.** Example of a SequenceD.

### 2.2.3 Constraints for sequence diagram

UML defines two standard constraints for messages: vote and broadcast. The vote constraint restricts a collection of return messages, and the broadcast constraint specifies that the constrained messages are not invoked in any particular order.

Beyond the UML standard message constraints found in sequence diagrams, elkoutbi et al. [5,6] define the two additional constraints input Data and output Data. The input Data constraint indicates that the corresponding message holds input information from the user. The outputData constraint specifies that the corresponding message carries information for display. Both input Data and output Data constraints have a parameter that indicates the kind of user action. This parameter normally represents the dependency between the message and the elements of the underlying class diagram.

## 3. Description of the Approach

In this section, we describe the overall approach to add security constraints and generate code from scenarios (sequence diagram). The approach consists of three activities (see Figure 3), which are detailed below:

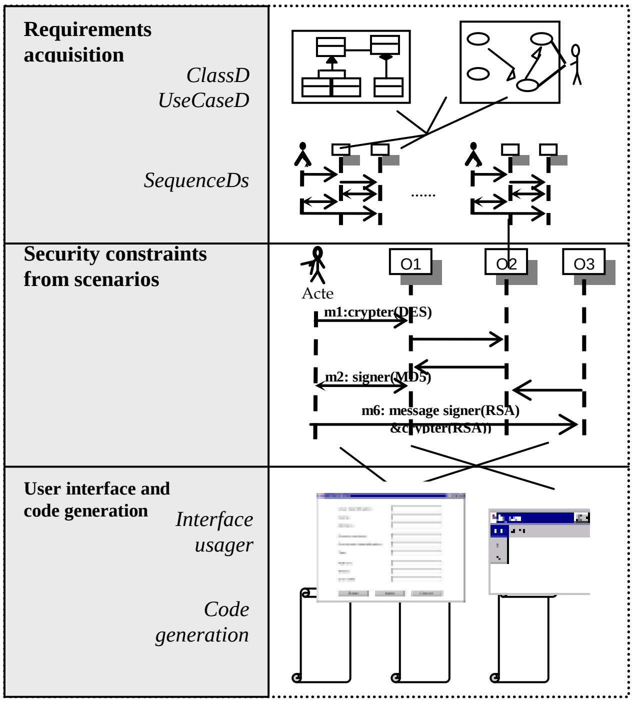


**Figure 3.** Overview of the proposed process

In the *Requirements Acquisition* activity, the analyst elaborates the UsecaseD, and for each use case, he or she elaborates several SequenceDs corresponding to the scenarios of the use case at hand.

The *Security Constraints activity* consists of extending interaction diagrams. We propose to extend the UML with the following message constraints: security constraints.

In the *User interface prototype and code generation activity*, we generate code from a sequence diagram and derive user interface prototypes for all the interface objects found in the system.

In the following, we will discuss these activities of the proposed process.

### 3.1 Security constraints for sequence diagram

Today, security has become a major issue for information systems (e-business, e-trade, etc). It will be convenient to be able to define and represent these constraints in the step of requirement engineering. We were interested in the major security aspects: authenticity and confidentiality. Authenticity means the proof of identity and confidentiality relates to the privacy of information. Using UML, when a message is sent from a source to a target object, it can carry some information (message parameters). We aim to express that the exchange is private using some encryption algorithms (RSA, AES, 3DES, etc). This can be specified as a parameter of the constraint. The two constraints defined to model security aspects are given below:

m{Auth}: The message m must be signed by the sender object to proof its identity to the receiver object.

m{Crypt(algo)}: The message content (message parameters) must be encrypted using the algorithm (algo).

Figures 4 give an example a sequence diagram with security constraints for ATM (Automatic Teller Machine)

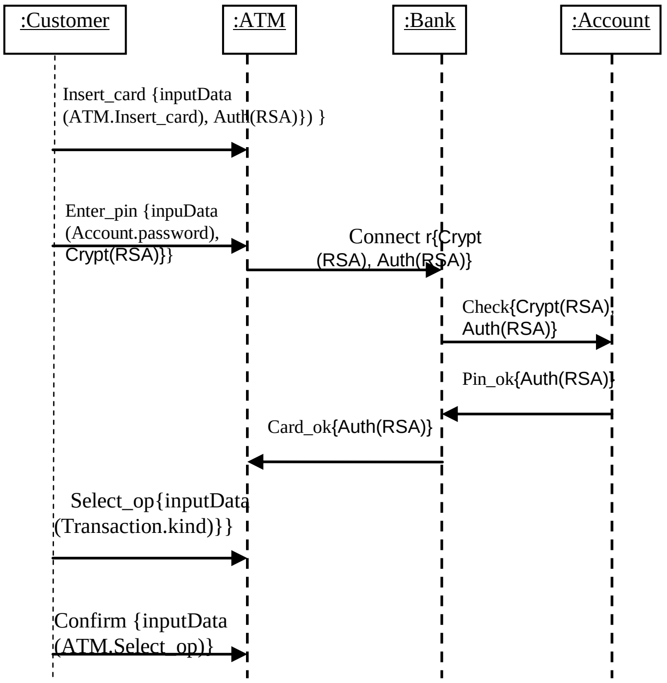

:Customer
:ATM
:Bank
:Account
Insert_card {inputData (ATM.Insert_card), Auth(RSA)}) }
Enter_pin {inpuData (Account.password), Crypt(RSA)}}
Connect r{Crypt (RSA), Auth(RSA)}
Check{Crypt(RSA) Auth(RSA)}
Pin_ok{Auth(RSA)}
Card_ok{Auth(RSA)}
Select_op{inputData (Transaction.kind)}}
Confirm {inputData (ATM.Select_op)}

**Figure 4.** Sequence diagram with security constraints

### 3.2 Code generation

UML and Java [7], which is an OO model and design notation and an OO programming language respectively, are some of the tools widely used in many software development projects. However, these modeling and programming activities are mostly separated. And a gap exists between these models and programs.

This paper proposes an approach to narrow the gap between multiple UML models and an implemented system. The narrowing of a gap is achieved by generating Java source code directly from multiple UML model diagrams. The code generation is achieved by creating a mapping between UML and the Java programming language.

Many current OO CASE tools [8, 9, 10, 11, 12,13] generate limited skeleton code from such models. The main drawback of this approach is that there is no code generation for object behavior and thus the code generated is not complete. Chow et al. [14] developed two main steps in translating code from dynamic behaviour of the system. Translate an object's state diagram into Java code and Generate method body based on the pre/post condition of an operation and specify the order of language statements based on the message passing sequence in the interaction diagram. Jakimi et al. [15] generated automatically implementation code from the UML sequence diagrams in an object-oriented programming language such as Java.

The generation of Java code from the UML sequence diagram has been met with some degree of success. In addition to the generation of skeleton class code from class diagram, Java codes have been generated from the statechart, the sequence diagram and the component diagram. However, generation of Java source code from all UML diagrams is not yet achievable.

Figure 5 presents the types of the security constraints which we can associate an exchange of messages between the objects.

- Simple message
- Encrypted message
- Encrypted and signed message
- Signed message

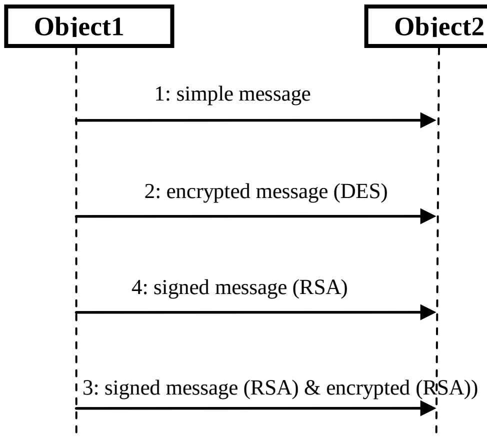

Object1
Object2
1: simple message
2: encrypted message (DES)
4: signed message (RSA)
3: signed message (RSA) & encrypted (RSA))

**Figure 5.** Sequence diagram with type's security constraints

The code generated relating to the diagram of figure 5 is arising according to the type of message sent (simple, signed, encrypted or signed and encrypted). The code generated by this approach for the figure 6 is:

```
class mySystem {
          public void service(){
      // simple message
             Object2.m1();
    //encrypted  messageé
        Object2.m2(encrypted(DES/RSA));
  // signed message
             Object2.m3(signed(MD5/RSA));
   //signed/encrypted message
              Object2.m3(signeg(RSA) & encrypted(RSA)));
          }
}
```

**Figure 6.** Code generation from figure 5

### 3.3 Tool support

For create the tool support for code generated from the sequence diagram with security constraints. We have used the Eclipse environment, the TogetherJ plug-in for UML modeling and the application programming interface (API) JDOM for XML manipulation.

We used the plug-in for UML diagrams (from Together) which makes it possible for us to create sequence diagrams. This diagram is first acquired to throw the UML diagram plug-in, and then there is transformed in form XML files.

This XML file can also be imported via the UML diagram plug-in for purposes of visualization and annotation. Finally we develop a code generator for automatic Java code security generation from sequence diagram.

## 4. Conclusion

In conclusion, we have proposed in this paper an UML-based code generation approach. In this work, we have presented a new approach that produces automatically code from the sequence diagram with security constraints. This approach helps developers to transit from the design to implementation phase and to shorten the software development cycle.

The future works of this research include the following areas: generate code from UML diagrams that describe dynamic and non-functional aspects of a system while remaining platform independent and optimize generated code, find a rigorous method to lower the abstraction level.

## References

[1] Object Management Group (OMG), Unified Modeling Language (UML) specifications version 1.5, 2003. http://www.omg.org/

[2] G. Booch, J. Rumbaugh, and I. Jacobson, "The Unified Modeling Language: User Guide", Massachusetts: Addison-Wesley, 1999.

[3] J. Rumbaugh, I. Jacobson, and G. Booch, "The Unified Modeling Language: Reference Manual Guide", Massachusetts: Addison-Wesley, 1999.

[4] C. Rolland, C. Ben Achour, C. Cauvet, J. Ralyté, A. Sutcliffe, N.A.M. Maiden, M. Jarke, P. Haumer, K. Pohl, E. Dubois and P. Heymans. "A Proposal for a Scenario Classification Framework". The Requirements Engineering Journal, Volume 3, Number 1, 1998.

[5] M. Bennani., M. Elkoutbi., and K. Nafil; Modelling Real-time Aspects using UML Scenarios proceedings of the 3rd International Conference on Software Methodologies, Tools and Techniques, pp. 200-213, Leipzig, Germany, September 2004.

[6] M. Elkoutbi, Khriss I., R.K. Keller. "Automated Prototyping of User Interfaces Based on UML Scenarios". The Automated Software Engineering Journal, 13, 5-40, 2006.

[7] Sun Microsystems Inc., Java Technology, http://java.sun.com

[8] J. Ali, and J. Tanaka, "Converting Statecharts into Java Code", in Proc. IDPT'00, Dallas, Texas, USA, 2000.

[9] I-Logix Inc., Rhapsody, http://www.ilogix.com.

[10] D. Harel, and E. Grey, "Executable Object Modeling with Statecharts", in Proc. of 18th Inter. Conf. on Software Engineering, IEEE, March 1996, pp. 246-257.

[11] D. Harel, and E. Grey, "Executable Object Modeling with Statecharts",Computer, vol. 30, no. 7, 1997, pp. 31-42.

[12] I. Azim Niaz and J. Tanaka, An Object-Oriented Approach To Generate Java Code From UML Statecharts, International Journal of Computer & Information Science, Vol. 6, No. 2, June 2005

[13] J. Ali, and J. Tanaka, "Implementing the Dynamic Behavior Represented as Multiple State Diagrams and Activity Diagrams", Journal of Computer Science & Information Management , vol. 2, no. 1, 2001, pp.

[14] K.O. Chow, W. Jia, V.C.P. Chan and J. Cao, Modelbased generation of Java code, Proc. International Conf. on Parallel and Distributed Processing Techniques and Applications (PDPTA 2000), Las Vegas, USA, 2000.

[15] A. Jakimi and M. El Koutbi, "An Object-Oriented Approach to UML Scenarios Engineering and Code Generation" International Journal of Computer Theory and Engineering , vol.1, No 1, pp 35-41, April 2009.

## Author Biographies

**Abdeslam Jakimi** is a professor at Faculty of Science and technology in myIismail University, he received his Masters degree in software engineering in 2004. His current research interests include requirements engineering, user interface prototyping, design transformations, scenario engineering and code generation.

**Mohammed Elkoutbi** is a professor at National School of Computer and Systems Analysis in Agdal, Rabat, Morocco. His current research interests include requirements engineering, user interface prototyping and design, and formal methods in analysis and design. He earned a PhD in Computer Science from University of Montreal in 2000.

# Interaction and Integration of Agent Mining in Distributed Data Environment

K.Raghava Rao[1], M. Nagabhushnam Rao[2], Y.Siva Prasad [3] & K.Ruth Ramya

Professor & Head[1], MCA Department, Asst. Prof[3], CSE Department, Asst. Professor[4], Dept. of CSE,

KLUniversity,Vaddeswaram,AP State ,Professor & Head[2], BV Raju Engg. Colleg Bhimavaram,AP

{raghavarao1@yahoo.com},{mnraosir,sivaprasady,ruthmoses.mathi}@gmail.com

***Abstract:*** In recent years, more and more researchers have been involved in research on both agent technology and distributed data mining. A clear disciplinary effort has been activated toward removing the boundary between them,that is the interaction and integration between agent technology and distributed data mining. We refer this to *agent mining* as a new area. The marriage of agents and distributed data mining is driven by challenges faced by both communities, and the need of developing more advanced intelligence, information processing and systems. In this paper presents an overall picture of agent mining from the perspective of positioning it as an emerging area. We summarize the main distributed data mining, driving forces, disciplinary framework, applications, and trends and directions, data mining-driven agents, and mutual issues in agent mining. Arguably, we draw the following conclusions: (1) agent mining emerges as a new area in the scientific family, (2) both agent technology and distributed data mining can greatly benefit from agent mining, (3) it is very promising to result in additional advancement in intelligent information processing and systems. However, as a new open area, there are many issues waiting for research and development from theoretical, technological and practical perspectives.

***Keywords:*** Agent mining, distributed data mining& environment, KDD, AAMAS, DDM Algorithm, Adaptive Learning

## 1. Introduction

Autonomous agent and multi-agent systems (AAMAS, refer to here as *agents*) [44] and knowledge discovery from data (KDD, or otherwise known as *data mining*)[10] have emerged and developed separately in the last twenty years. Both areas are currently very active. Agents primarily focus on issues from many aspects, from theoretical, methodological, and experimental to practical issues in developing agent-based computing and agent-oriented intelligent systems, which are a powerful technology for autonomous intelligent system analysis, design and implementation. The major topics of interest consist of research on individual agents, multi-agent systems (MAS), methodology and techniques, tools and applications. The agent technology contributes to many diverse domains such as software engineering, user interfaces, ecommerce, information retrieval, robotics, computer games, education and training, ubiquitous computing, and social simulation.

Currently, agent studies have been spread from programming to organizational and societal factors to study agents and agent-based systems. The research on agents has far exceeded the original community scope of artificial intelligence and software. Researchers from many other areas have started to discuss, develop, wrap and use the concept of agents, covering almost all aspects of the social sciences such as law, business, organizational, behavior sciences, finance and economics, tourism, not to mention the extensive family of natural science and technology. The benefits from agents are expected to be very comprehensive and diverse, from academic disciplines, to the sciences, the social sciences and the humanities.

Similarly, *data mining* originally focused on knowledge discovery in databases, but it has experienced a migration from data-centered pattern discovery, to knowledge discovery, actionable knowledge discovery, and currently to domain-oriented decision delivery [11]. Data mining and its tools is becoming a ubiquitous information processing field and tools, involving techniques and researchers from many areas such as statistics, information retrieval, machine learning, artificial intelligence, pattern recognition, and database technologies. Data mining is increasingly widely tested in varying applications and domains, for instance, web mining and services, text mining, telecommunications, retail, governmental service, fraud, security, business intelligence studies. Besides the emphasis of in-depth data intelligence, recent efforts in data mining cover many additional areas and domain problems. Data mining researchers recognize the need to involve the environment, human intelligence, domain intelligence, organizational intelligence, and social intelligence in the mining process, models, the findings and deliverables.

This will trigger another wave of migration from the discovery of knowledge to the delivery of deep knowledge-based problem-solving systems and services. The above analysis of trends and directions of both areas shows that these two independent research streams have been created and originally evolved with separate aims and objectives. They used to target individual methodologies and techniques to cope with domain-specific problems and challenges in respective areas. However, both are concerned with many similar aspects and factors, such as human roles, user system interaction, dynamic modeling, domain factors, organizational and social factors. In fact, both areas contribute to the advancement of intelligence, and intelligent information processing, services and systems. In fact, they need each other, as evidenced by typical topics of agent-based data mining in the middle 1990s. Consequently, we see a clear trend of the interaction and integration between agents and data mining. Its development has reached the level of a new and promising area, and is moving towards becoming a first-class citizen in the science and technology family [12, 5, 6]. In this paper presents an overall picture of this emerging field, distributed data mining and multi-agent integration. We first analyze the respective and common challenges in agents and distributed data mining areas. These challenges motivate and drive the need and emergence of agent mining. A scientific framework and theoretical underpinnings are presented, which illustrate the synergy methods and foundations of agents and data mining. Further, we briefly summarize the research on three major directions in agent mining, namely agent-driven distributed data mining, mining-driven agents, and mutual issues in agent mining and applications are discussed. Finally, we discuss the development of agent mining community. Information provided here can benefit new researchers, and enable them to quickly step into this field.

## 2. Distributed Data Mining

Traditional warehouse-based architectures for data mining suppose to have centralized data repository. Such a centralized approach is fundamentally inappropriate for most of the distributed and ubiquitous data mining applications. In fact, the long response time, lack of proper use of distributed resource, and the Fundamental characteristic of centralized data mining algorithms do not work well in distributed environments. A scalable solution for distributed applications calls for distributed processing of data, controlled by the available resources and human factors. For example, let us suppose an ad hoc wireless sensor network where the different sensor nodes are monitoring some time-critical events. Central collection of data from every sensor node may create traffic over the limited bandwidth wireless channels and this may also drain a lot of power from the devices. A distributed architecture for data mining is likely aimed to reduce the communication load and also to reduce the battery power more evenly across the different nodes in the sensor network. One can easily imagine similar needs for distributed computation of data mining primitives in ad hoc wireless networks of mobile devices like PDAs, cell phones, and wearable computers. The wireless domain is not the only example. In fact, most of the applications that deal with time-critical distributed data are likely to benefit by paying careful attention to the distributed resources for computation, storage, and the cost of communication. As another example, let us consider the World Wide Web: it contains distributed data and computing resources. An increasing number of databases (e.g., weather databases, oceanographic data, etc.) and data streams (e.g., financial data, emerging disease information, etc.) are currently made on-line, and many of them change frequently. It is easy to think of many applications that require regular r monitoring of these diverse and distributed sources of data. A distributed approach to analyze this data is likely to be more scalable and practical particularly when the application involves a large number of data sites. Hence, in this case we need data mining architectures that pay careful attention to the distribution of data, computing and communication, in order to access and use them in a near optimal fashion. Distributed Data Mining (sometimes referred by the acronym DDM) considers data mining in this broader context. DDM may also be useful in environments with multiple compute nodes connected over high speed networks. Even if the data can be quickly centralized using the relatively fast network, proper balancing of computational load among a cluster of nodes may require a distributed approach. The privacy issue is playing an increasingly important role in the emerging data mining applications. For example, let us suppose a consortium of different banks collaborating for detecting frauds. If a centralized solution was adopted, all the data from every bank should be collected in a single location, to be processed by a data mining system. Nevertheless, in such a case a distributed data mining system should be the natural technological choice: both it is able to learn models from distributed data without exchanging the raw data between different repository, and it allows detection of fraud by preserving the privacy of every bank's customer transaction data. For what concerns techniques and architecture, it is worth noticing that many several other fields influence Distributed Data Mining systems concepts. First, many DDM systems adopt the Multi-Agent System (MAS) architecture, which finds its root in the Distributed Artificial Intelligence (DAI). Second, although Parallel Data Mining often assumes the presence of high speed network connections among the computing nodes, the development of DDM has also been influenced by the PDM literature. Most DDM algorithms are designed upon the potential parallelism they can apply over the given distributed data.. In figure 1 a general Distributed Data Mining framework is presented. In essence, the success of DDM algorithms lies in the aggregation. Each local model represents locally coherent patterns, but lacks details that may be required to induce globally meaningful knowledge. For this reason, many DDM algorithms require a centralization of a subset of local data to compensate it. The ensemble approach has been applied in various domains to increase the accuracy of the predictive model to be learnt. It produces multiple models and combines them to enhance accuracy. Typically, voting (weighted or un-weighted) schema are employed to aggregate base model for obtaining a global model. As we have discussed above, minimum data transfer is another key attribute of the successful DDM algorithm.

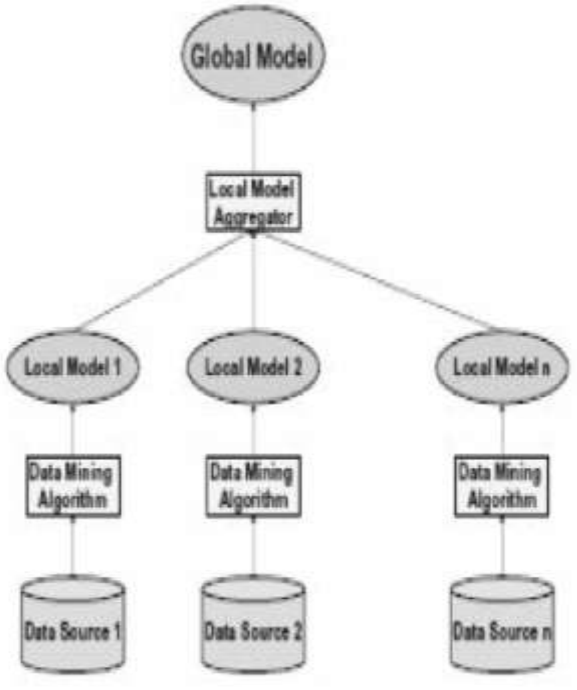


**Figure 1:** General Distributed data mining Frame work

## 3. The Challenges of Distributed Data Mining

Data mining and machine learning currently forms a mature field of artificial intelligence supported by many various approaches, algorithms and software tools. However, modern requirements in data mining and machine learning inspired by emerging applications and information technologies and the peculiarities of data sources are becoming increasingly tough. The critical features of data sources determining such requirements are as follows:

In enterprise applications, data is distributed over many heterogeneous sources coupling in either a tight or loose manner. Distributed data sources associated with a business line are often complex, for instance, some is of high frequency or density, mixing static and dynamic data, mixing multiple structures of data; Data integration and data matching are difficult to conduct; it is not possible to store them in centralized storage and it is not feasible to process them in a centralized manner; In some cases, multiple sources of data are stored in parallel storage systems; Local data sources can be of restricted availability due to privacy, their commercial value, etc., which in many cases also prevents its centralized processing, even in a collaborative mode; In many cases, distributed data spread across global storage systems is often associated with time difference; Availability of data sources in a mobile environment depends on time; The infrastructure and architecture weaknesses of existing distributed data mining systems requires more flexible, intelligent and scalable support. These and some other peculiarities require the development of new approaches and technologies of data mining to identify patterns in distributed data. Distributed data mining (DDM), in particular, Peer- to-Peer (P2P) data mining, and multi-agent technology are two responses to the above challenges.

## 4. Challenges in Distributed Data Mining Disciplines

Data mining faces many challenges when it is deployed to real world problem solving, in particular, in handling complex data and applications. We list here a few aspects that can be improved by agent technology. These include enterprise data mining infrastructure, involving domain and human intelligence, supporting parallel and distributed mining, data fusion and preparation, adaptive learning, and interactive mining.

*(a) Enterprise data mining infrastructure:* The development of data mining systems supporting real-world enterprise applications are challenging. The challenge may arise from many aspects, for instance, integrating or mining multiple data sources, accessing distributed applications, interacting with varying business users, and communicating with multiple applications. In particular, it has been a grand challenge and a longstanding issue to build up a distributed, flexible, adaptive and efficient platform supporting interactive mining in real-world data.

*(b) Involving domain and human intelligence:* Another grand challenge of existing data mining methodologies and techniques are the roles and involvement of domain intelligence and human intelligence in data mining. With respect to domain intelligence, how to involve, represent, link and confirm to components such as domain knowledge, prior knowledge, business process, and business logics in data mining systems is a research problem. Regarding human intelligence, we need to distinguish the role of humans in specific applications, and further build up system support to model human behavior, interact with humans, bridge the communication gap between data mining systems and humans, and most importantly incorporate human knowledge and supervision into the system.

*(c) Data fusion and preparation:* In the real world, data is getting more and more complex, in particular, sparse and heterogeneous data distributed in multiple places. To access and fuse such data needs intelligent techniques and methods. On the other hand, today's data preparation research is facing new challenges such as processing high frequency time series data stream, unbalanced data distribution, rare but significant evidence extraction from dispersed data sets, linking multiple data sources, accessing dynamic data. Such situations expect new data preparation techniques.

*(d) Adaptive learning:* In general, data mining algorithms are predefined to scan data sets. In real-world cases, it is expected that data mining models and algorithms can adapt to dynamic situations in changing data based on their self-learning and self-organizing capability. As a result, models and algorithms can automatically extract patterns in changing data. However, this is a very challenging area, since existing data mining methodologies and techniques are basically non-automatic and inadaptable. To enhance the automated and adaptive capability of data mining algorithms and methods, we need to search for support from external disciplines that are related to automate and adaptive intelligent techniques.

*(e) Interactive mining:* Controversies regarding either automatic or interactive data mining have been raised in the past. A clear trend for this problem is that interaction between humans and data mining systems plays an irreplaceable role in domain-driven data mining situations. In developing interactive mining, one should study issues such as user modeling, behavior simulation, situation analysis, user interface design, user knowledge management, algorithm/model input setting by users, mining process control and monitor, outcome refinement and tuning. However, many of these tasks cannot be handled by existing data mining approaches.

## 5. Driving Forces of Agent Mining Interaction and Integration

The emergence of agent mining results from the following driving forces: The critical challenges in agents and data mining respectively, the critical common challenges troubling agents and data mining the complementary essence of agents and data mining in dealing with their challenges, and the great add-on potential resulting from the interaction and integration of agents and data mining. Agents and data mining are facing critical challenges from respective areas. Many of these challenges can be tackled by involving advances in other areas.

In this section, we specify both individual and mutual challenges in agent and mining disciplines that may be complemented by the interaction with the other disciplines.

## 6. Challenges in Agent Disciplines

As addressed in some retrospective publications, traditional agent technology has been challenged in many aspects such as developing organizational and social intelligence.

Figure **2:** Challenges in agents and data mining.

In the following analysis, we explain this from the following aspects: agent awareness, agent learning, agent action ability, and agent distributed processing, agent in-depth services, and agent constraint processing.

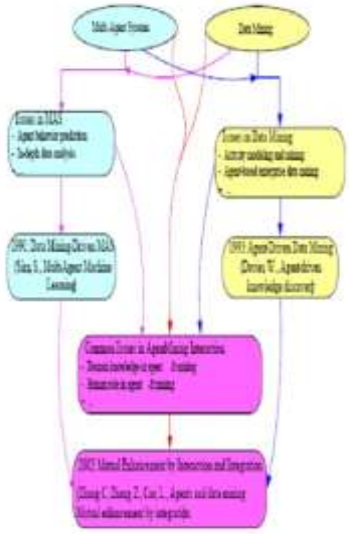


Fig.2. illustrates these challenges.

*(a) Agent awareness:* Agent awareness refers to the capability of an agent to recognize internal and/or external environment change, and analyze situation change. In contrast to normal sensing and perception as conducted in reactive agents, here agent awareness specifically refers to situation analysis and environment modeling driven by agent learning and discovery. Agents with such a capability should self-recognize, compare and reason the changes taking place in the environment. To this end, it is necessary for agents to accumulate learning capability.

*(b) Agent learning:* In open multi-agent organizations, interaction widely exists between agent and environment, and between an agent and the other agent(s). Agents are expected to learn from other agents, their environment, and from the interaction and dynamics. In addition, agents may be expected to learn from users and interaction with humans. To foster such learning capability, agents need to be fed with learning and reasoning algorithms that can support them to discover reason or simulate interesting information from

interactive and situational data. Learning capability is widely recognized to be significant for enhancing agent intelligence. On the basis of the varying objectives, agent learning has been paid unprecedented attention. Multiple forms of agent learning capability are being studied. Agent learning may be conducted in terms of agent architectures such as cognitive learning, deductive learning, distributed learning, and cooperative learning. With respect to learning objectives, agent learning may also be classified into procedural learning, action learning, rule and pattern learning, and decision-making learning. From the learning process aspect, agent learning can be categorized into reinforcement learning, discovery learning, single-trial learning, reasoning learning, and random learning. The implementation of agent learning presents either a passive or active manner. In a broad sense, learning can be in a supervised, unsupervised or hybrid manner.

*(c) Agent action ability:* Agent action ability refers to the capability of an agent to take actions to its advantage on the basis of the knowledge obtained through in-depth analysis, reasoning and discovery. Unlike general action taken by agents, we are specifically interested in actions for recommendation, servicing, searching, discovery, conflict resolution, etc. with great benefits but low costs. To this end, agents need to balance benefits and costs, and maximize their return while minimizing the risk before taking an action or a sequence of actions.

*(d) Agent distributed processing:* In middle to large scale multi-agent systems, agents need to deal with distributed processing tasks such as learning from agents across multiple organizations, applications or data, conducting decentralized coordination, cooperation and negotiation among agents crossing resources, and implementing information gathering, dispatching and transport among agents located in distributed applications. To tackle the above tasks in distributed conditions, agents need to make decisions after analyzing and utilizing relevant information from multiple sources. Information analysis and utilization is not a trivial job. Agents may need to develop capabilities such as data analysis and discovery, procedural learning, goal adjustment, and information fusion.

*(e) Agent in-depth services:* Agents are often developed for providing varied services, for instance, network services such as web recommender systems, mobile agents for information searching and passing, and user services such as for user interaction and user modeling. Smart service providing relies on in-depth analysis of the service request-related data and information, as well as service historical data and service performance, in order to deeply understand service data and select the best service solutions. However, the agent community often does not work on such kinds of capabilities.

*(f) Agent constraint processing:* Open complex agent systems often involve many types of constraints from many aspects, for instance, temporal and spatial constraints, or execution constraints from organizational aspects. Such constraints form conditions in improving agent capabilities such as learning, adaptation, action ability, and services. There is a need to understand such constraints, and to involve and best treat such constraints in an agent system and solution generation.

## 7. Mutual Challenges in Agent and Distributed Data Mining

As addressed in [5, 6, 7], agents can enhance data mining through involving agent intelligence in data mining systems, while an agent system can benefit from data mining via extending agents' knowledge discovery capability. Nevertheless, the agent mining interaction symbiosis cannot be established if mutual issues are not solved. These mutual issues involve fundamental challenges hidden on both sides and particularly within the interaction and integration. Figure 2 presents a view of issues in agent-mining interaction highlighting the existence of mutual issues. Mutual issues constraining agent-mining interaction and integration consist of many aspects such as architecture and infrastructure, constraint and environment, domain intelligence, human intelligence, knowledge engineering and management, and nonfunctional requirements.

*(a) Architecture and infrastructure:* Data mining always faces a problem in how to implement a system that can support those brilliant functions and algorithms studied in academia. The design of the system architecture conducting enterprise mining applications and emerging research challenges needs to provide (1) functional support such as crossing source data management and preparation, interactive mining and the involvement of domain and human intelligence, distributed, parallel and adaptive learning, and plug-and-play of algorithms and system components, as well as (2) nonfunctional support for instance adaptability, being user and business friendly and flexibility. On the other hand, middle to large scales of agent systems are not easily built due to the essence of distribution, interaction, human and domain involvement, and openness. In fact, many challenging factors in agent and mining systems are similar or complementary.

*(b) Constraint and environment:* Both agent and mining systems need to interact with the environment, and tackle the constraints surrounding a system. In agent communities, environment could present characters such as openness, accessibility, uncertainty, diversity, temporality, spatiality, and/or evolutionary and dynamic processes. These factors form varying constraints on agents and agent systems. Similar issues can also be found from real-world data mining, for instance, temporal and spatial data mining. The dynamic business process and logics surrounding data mining make the mining very domain-specific and sensitive to its environment.*2 Domain intelligence* Domain intelligence widely surrounds agent and mining systems. Both areas need to understand, define, represent, and involve the roles and components of domain intelligence. In particular, it is essential in agent mining interaction to model domain and prior knowledge, and to involve it to enhance agent-mining intelligence and actionable capability.

*(c) Human intelligence:* Both agent and mining need to consider the roles and components of human intelligence. Many roles may be better played by humans in agent-mining interaction. To this end, it is necessary to study the definition and major components of human intelligence, and how to involve them in agent mining systems. For instance, mechanisms should be researched on user modeling, user and business friendly interaction interfaces, and communication languages for agent-mining system dialogue.

*(d) Knowledge engineering and management:* To support the involvement of domain and human intelligence, proper mechanisms of knowledge engineering and management are substantially important. Tasks such as the management, representation, semantic relationships, transformation and mapping between multiple domains, and meta-data and meta-knowledge are essential for involving roles and data/knowledge intelligence in building up agent-mining simians.

*(e) Nonfunctional requirements* Nonfunctional requests are essential in real-world mining and agent systems. The agent-mining simians may more or less address nonfunctional requirements such as efficiency, effectiveness, action ability, and user and business friendliness.

## 8. Disciplinary Framework of Agent and Mining Interaction and Integration

This section aims to draw a concept map of agent mining as a scientific field. We observe this from the following perspectives: evolution process and characteristics, agent-mining interaction framework.

### 8.1 Evolution process and characteristics

As an emerging research area, agent mining experiences the following evolution process, and presents the following unprecedented characteristics.

*(a) From one-way interaction to wo-way interaction*: The area was originally initiated by incorporating data mining into agent to enhance agent learning [20, 40].Recently, issues in two-way interaction and integration have been broadly studied in different groups.

*(b) From single need-driven to mutual needs-driven*: Original research work started on the single need to integrate one into the other, whereas it it is now driven by both needs from both parties. As discussed in [12, 8], people have found many issues in each of the related communities. These issues cannot be tackled by simply developing internal techniques. Rather, techniques from other disciplines can greatly complement the problem-solving when they are combined with existing techniques and approaches. This greatly drives the development of agent-driven data mining and data mining-driven agents.

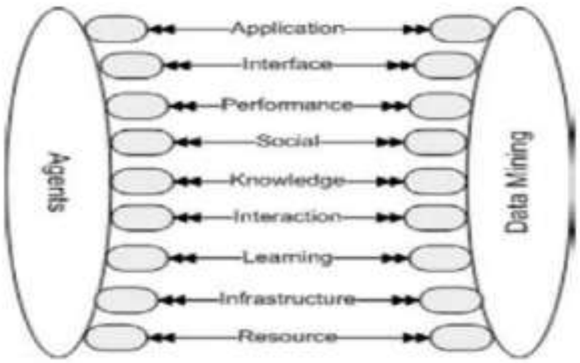
*(c ) Intrinsic associations and utilities*: The interaction and integration between agents and data mining is also driven and connected by intrinsic overlap, associations, complementation and utilities of both parties, as discussed in [5, 6]. This drives the research on mutual issues, and the synergetic research and systems coupling both technologies, into a more advanced form.

*(d) Application drives*: Application request is one of the key driving forces of this new trend. We present some major application domains and problems that may be better handled by both agent and mining techniques.

*(e) Major research groups and researchers* [6] :in respective communities tend to undertake both sides of research. Some of them are trying to link them together to solve problems that cannot be tackled by one of them alone, for instance, agent-based distributed learning [30, 31, 32, 25, 26], agent-based data mining infrastructure [4, 5, 26], or data mining driven agent intelligence enhancement [4, 35].

### 8.2 Agent-mining interaction framework

The interaction and integration between agents and data mining are comprehensive, multiple dimensional, and inter-disciplinary. As an emerging scientific field, *agent mining* studies the methodologies, principles, techniques and applications of the integration and interaction between agents and data mining, as well as the community that focuses on the study of agent mining. On the basis of complementation between agents and data mining, agent mining fosters a synergy between them from different dimensions, for instance, *resource*, *infrastructure*, *learning*, *knowledge*, *interaction*, *interface*, *social*, *application* and *performance*. As shown in Figure 3, we briefly discuss these dimensions.



*(a) Resource layer:* Interaction and integration may happen on data and information levels;
Infrastructure layer: Interaction and integration may be on infrastructure, architecture and process sides;
*(b) Knowledge layer:* Interaction and integration may be based on knowledge, including domain knowledge, human expert knowledge, meta-knowledge, and knowledge retrieved, extracted or discovered in resources;

*(c) Learning layer:* Interaction and integration may be on learning methods, learning capabilities and performance perspectives;

*(d) Interaction layer:* Interaction and integration may be on coordination, cooperation, negotiation, communication perspectives;

*(e) Interface layer:* Interaction and integration may be on human-system interface, user modeling and interface design;
*(f) Social layer:* Interaction and integration may be on social and organizational factors, for instance, human roles;
*(g) Application layer:* Interaction and integration may be on applications and domain problems;

*(h) Performance layer:* Interaction and integration may be on the performance enhancement of one side of the technologies or the coupling system.

From these dimensions, many fundamental research issues/problems in agent mining emerge. Correspondingly, we can generate a high-level research map of agent mining as a disciplinary area. Figure .4 shows such a framework, which consists of the following research components: *agent mining foundations*, *agent-driven data processing*, *agent-driven knowledge discovery*, *mining-driven multi-agent systems*, *agent-driven information processing*, *mutual issues in agent mining, agent mining systems*, *agent mining applications*, *agent mining knowledge management*, and *agent mining performance evaluation*. We briefly discuss them below.

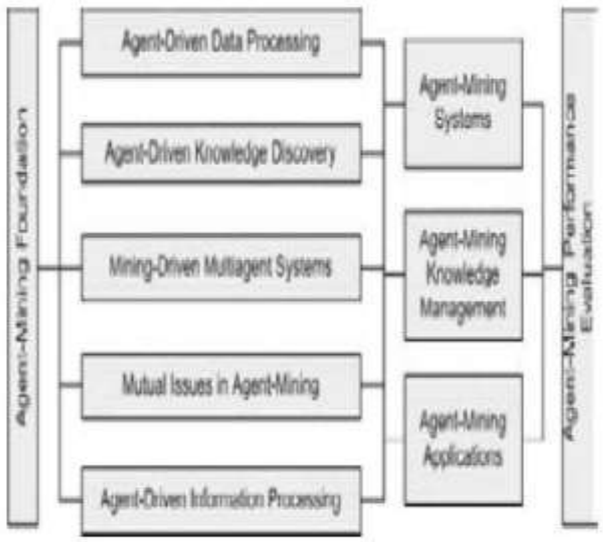

Agent-Mining Foundation
Agent-Driven Data Processing
Agent-Driven Knowledge Discovery
Mining-Driven Multiagent Systems
Mutual Issues in Agent-Mining
Agent-Driven Information Processing
Agent-Mining Systems
Agent-Mining Knowledge Management
Agent-Mining Applications
Agent-Mining Performance Evaluation

**Figure 4:** Agent-Mining Disciplinary Framework.

(1) Agent mining foundations: Studies issues such as the challenges and prospects, research map and theoretical underpinnings, theoretical foundations, formal methods, and frameworks, approaches and tools.

(2) Agent-driven data processing: Studies issues including multi-agent data coordination ,multi-agent data extraction, multi-agent data integration, multi-agent data management, multi-agent data monitoring, multi-agent data processing and preparation, multi-agent data query and multi-agent data warehousing;

(3) Agent-driven knowledge discovery: Studies problems like multi-agent data mining infrastructure and architecture, multi-agent data mining process modeling and management, multi-agent data mining project management, multi-agent interactive data mining infrastructure, multi-agent automated data learning, multi-agent cloud computing, multi-agent distributed data mining, multi-agent dynamic mining, multi-agent grid computing, multi-agent interactive data mining, multi-agent online mining, multi-agent mobility mining, multi-agent multiple data source mining, multi-agent

ontology mining, multi-agent parallel data mining, multi agent peer-to-peer mining, multi-agent self-organizing mining, multi-agent text mining, multi-agent visual data mining, and multi-agent web mining; *Mining-driven multi-agent systems (MAS)* studies issues such as data mining driven MAS adaptation, data mining-driven MAS behavior analysis, data mining driven MAS communication, data mining-driven MAS coordination, data mining driven MAS dispatching, data mining-driven MAS distributed learning, data mining-driven MAS evolution, data mining-driven MAS learning, data mining driven MAS negotiation, data mining-driven MAS optimization, data mining driven MAS planning, data mining-driven MAS reasoning, data mining-driven MAS recommendation, data mining-driven MAS reputation/risk/trust analysis, data mining-driven self-organized and self-learning MAS, data mining-driven user modeling and servicing, and semi-supervised MAS learning;

(4) Agent-driven information processing: Multi-agent domain intelligence involvement, multi-agent human-mining cooperation, multi-agent enterprise application integration, multi-agent information gathering/retrieval, multi-agent message passing and sharing, multi-agent pattern analysis, and multi-agent service oriented computing.

(5) Mutual issues in agent mining: Including issues such as actionable capability, constraints, domain knowledge and intelligence, dynamic, online and ad-hoc issues, human role and intelligence, human-system interaction, infrastructure and architecture problems, intelligence meta synthesis, knowledge management, lifecycle and process management, networking and connection, nonfunctional issues, ontology and semantic issues, organizational factors, reliability, reputation, risk, privacy, security and trust, services, social factors, and ubiquitous intelligence; *Agent mining knowledge management:* knowledge management is essential for both agents and data mining, as well as for agent mining. This involves the representation, management and use of ontologies, domain knowledge, human empirical knowledge, meta-data and meta-knowledge, organizational and social factors, and resources in the agent-mining symbionts. In this, formal methods and tools are necessary for modeling, representing and managing knowledge. Such techniques also need to cater for identifying and distributing knowledge, knowledge evolution in agents, and enabling knowledge use.

(6) Agent mining performance evaluation: *R*esearches on methodologies, frameworks, tools and test beds for evaluating the performance of agent mining, and performance benchmarking and metrics. Besides technical performance such as accuracy and statistical significance, business-oriented performance such as cost, benefit and risk are also important in evaluating agent mining. Other aspects such as mobility, reliability, dependability, trust, privacy and reputation, etc., are also important in agent mining.

(7) Agent mining systems: this research component studies the formation of systems, including techniques for the frameworks, modeling, design and software eng.

## 9. Applications

As we can see from many references, the proposal of agent mining is actually driven by broad and increasing applications. Many researchers are developing agent mining systems and applications dealing with specific business problems and for intelligent information processing. For instance, we summarize the following application domains.

- Artificial immune systems
- Artificial and electronic markets
- Auction
- Business intelligence
- Customer relationship management
- Distributed data extraction and preparation
- E-commerce
- Finance data mining
- Grid computing
- Healthcare
- Internet and network services, e.g., recommendation, personal assistant, searching retrieval, extraction services
- Knowledge management Marketing
- Network intrusion detection
- Parallel computing, e.g., parallel genetic algorithm
- Peer-to-peer computing and service
- Semantic web
- Text mining
- Web mining.

## 11. Conclusions

Agent and distributed data mining interaction and integration has emerged as a prominent and promising area in recent years. The dialogue between agent technology and data mining can not only handle issues that are hardly coped with in each of the interacted parties, but can also result in innovative and super-intelligent techniques and symbionts much beyond the individual communities.

This chapter presents a high-level overview of the development and major directions in the area. The investigation highlights the following findings: (1) agent mining interaction is emerging as a new area in the scientific family, (2) the interaction is increasingly promoting the progress of agent and mining communities, (3) it results in ever-increasing development of innovative and significant techniques and systems towards super-intelligent symbionts. As a new and emerging area, it has many ope issues waiting for the significant involvement of research resources, in particular practical and research projects from both communities. We believe the research and development on agent mining is very promising and worthy of substantial efforts by both established and new researchers.

## 12. References

[1]. Aciar, S., Zhang, D., Simoff, S., and Debenham, J.: Informed Recommender Agent: Utilizing Consumer Product Reviews through Text Mining. Proceedings of IADM2006. IEEE Computer Society (2006)

[2]. Batik's., Cho, J., and Bala, J.: Performance Evaluation of an Agent Based Distributed Data Mining System. Advances in Artificial Intelligence, Volume 3501/2005 (2005)

[3]. Cory, J., Butz, Nguyen, N., Takama, Y., Cheung, W., and Cheung, Y.: Proceedings of IADM2006 (Chaired by Longbing Cao, Zili Zhang, Vladimir Samoilov) in WI-IAT2006 Workshop Proceedings. IEEE Computer Society (2006)

[4]. Cao, L., Wang, J., Lin, l., and Zhang, C.: Agent Services-Based Infrastructure for Online Assessment of Trading Strategies. Proceedings of IAT'04, 345-349 (2004).

[5]. Cao, L.: Integration of Agents and Data Mining. Technical report, 25 June 2005. http://wwwstaff.it.uts.edu.au/ lbcao/publication/publications.htm.

[6]. Cao, L., Luo, C. and Zhang, C.: Agent-Mining Interaction: An Emerging Area. AIS-ADM, 60-73 (2007).

[7]. Cao, L., Luo, D., Xiao, Y. and Zheng, Z. Agent Collaboration for Multiple Trading Strategy Integration. KES-AMSTA, 361-370 (2008).

[8]. Cao, L.: Agent-Mining Interaction and Integration – Topics of Research and Development. http://www.agentmining.org/

[9]. Cao, L.: Data Mining and Multiagent Integration. Springer (2009).

[10]. Cao, L. and Zhang, C. F-trade: An Agent-Mining Symbiont for Financial Services. AAMAS 262 (2007).

[11]. Cao, L., Yu, P., Zhang, C. and Zhao, Y. Domain Driven Data Mining. Springer (2009).

[12]. Cao, L., Gorodetsky, V. and Mitkas, P. Agent Mining: The Synergy of Agents and Data Mining. IEEE Intelligent Systems (2009).

[13]. Cao, L. Integrating Agent, Service and Organizational Computing. International Journal of Software Engineering and Knowledge Engineering, 18(5): 573-596 (2008)

[14]. Cao, L. and He, T. Developing Actionable Trading Agents. Knowledge and Information Systems: An International Journal, 18(2): 183-198 (2009).

[15]. Cao, L. Developing Actionable Trading Strategies, Knowledge Processing and Decision Making in Agent-Based Systems, 193-215, Springer (2008).

[16]. Cao, L., Zhang, Z., Gorodetsky, V. and Zhang, C.. Editor's Introduction: Interaction between Agents and Data Mining, International Journal of Intelligent Information and Database Systems, Inderscience, 2(1): 1-5 (2008).

[17]. Cao, L., Gorodetsky, V. and Mitkas, P. Editorial: Agents and Data Mining. IEEE Intelligent Systems (2009).

[18]. Cao, L. Agent & Data Mining Interaction, Tutorial for 2007 IEEE/WIC/ACM Joint Conferences on Web Intelligence and Intelligent Agent Technology (2007).

[19]. Cao, L., Zhang, C. and Zhang, Z. Agents and Data Mining: Interaction and Integration, Taylor & Francis (2010).

[20]. Brazdil, P., and Muggleton, S.: Learning to Relate Terms in a Multiple Agent Environment.EWSL91 (1991)

[21]. Davies, W.: ANIMALS: A Distributed, Heterogeneous Multi-Agent Learning System. MSc Thesis, University of Aberdeen (1993)

[22]. Davies, W.: Agent-Based Data-Mining (1994)

[23]. Edwards, P., and Davies, W.: A Heterogeneous Multi-Agent Learning System. In Deen, S.M. (ed) Proceedings of the Special Interest Group on Cooperating Knowledge Based Systems. University of Keele (1993) 163-184.

[24]. Gorodetsky, V., Liu, J., Skormin, V. A.: Autonomous Intelligent Systems: Agents and Data Mining book. Lecture Notes in Computer Science Volume 3505 (2005)

[25]. Gorodetsky, V.; Karsaev, O.and Samoilov, V.: Multi-Agent Technology for Distributed Data Mining and Classification. IAT 2003. (2003) 438 - 441

[26]. Gorodetsky, V., Karsaev, O. and Samoilov, V.: Infrastructural Issues for Agent-Based Distributed Learning. Proceedings of IADM2006, IEEE Computer Society Press

[27]. Han, J., and Kamber, M.: Data Mining: Concepts and Techniques (2nd version). Morgan Kaufmann (2006)

[28]. Kaya, M. and Alhajj, R.: A Novel Approach to Multi-Agent Reinforcement Learning: Utilizing OLAP Mining in the Learning Process. IEEE Transactions on Systems, Man and Cybernetics, Part C, Volume 35, Issue 4 (2005) 582 - 590

[29]. Kaya, M. and Alhajj, R.: Fuzzy OLAP Association Rules Mining-Based Modular Reinforcement Learning Approach for Multi-Agent Systems. IEEE Transactions on Systems, Man and Cybernetics, Part B, Volume 35, Issue 2, (2005) 326 - 338

[30]. Klusch, M., Lodi, S. and Gianluca, M.: The Role of Agents in Distributed Data Mining: Issues and Benefits. Intelligent Agent Technology (2003): 211 - 217

[31]. Klusch, M., Lodi, S. and Moro,G.: Agent-Based Distributed Data Mining: The KDEC Scheme. Intelligent Information Agents: The AgentLink Perspective Volume 2586 (2003) Lecture Notes in Computer Science

[32]. Klusch, M., Lodi, S. and Moro, G.: Issues of Agent-Based Distributed Data Mining. Proceedings of AAMAS, ACM Press (2003)

# Semi-Centralized Multi-Authenticated RSSI Based Solution to Sybil Attack

Himadri Nath Saha # [1] , Dr. Debika Bhattacharyya # [2] , Dr. P. K.Banerjee *[3]

Assistant Professor # [1], Professor # [2], 3 Professor *[3]
Department of Computer Science and Engineering, Institute of Engineering and
Management, West Bengal, India # [1], # [2]
Department of Electronics and Communication Engineering, Jadavpur University,
West Bengal, India *[3]
him_shree_2004@yahoo.com # [1], bdebika@yahoo.com# [2]

**Abstract:** *Sybil attack is a serious threat for today's wireless adhoc networks. In this attack a single node impersonates several other nodes using various malicious means. In this paper we attempt to provide a hybrid solution using a combination of two already proposed methods. According to this newly proposed method the total network will be dynamically divided into several subgroups, as more and more nodes will enter the network. Each subgroup will be under the super vision of a single node, a central authority. Each subgroup will also contain RSSI detector nodes.*

**Keywords:** *Sybil ttack,MANET, RSSI authentication.*

## 1. Introduction

A mobile adhoc network is a collection of wireless mobile nodes that are dynamically and arbitrarily located in such a manner that the interconnections between nodes are capable of changing on a continual basis. This particular nature of the network makes it vulnerable to various, sybil attack being one of them. With no logically central, trusted authority to vouch for a one-to-one correspondence between entity and identity, it is always possible for an unfamiliar entity to present more than one identity, except under conditions that are not practically realizable for large-scale distributed systems. Peer-to-peer systems commonly rely on the existence of multiple, independent remote entities to mitigate the threat of hostile peers. Many systems replicate computational or storage tasks among several remote sites to protect against integrity violations (eg. data loss). Others fragment tasks among several remote sites to protect against privacy violations (data leakage). In either case, exploiting the redundancy in the system requires the ability to determine whether two ostensibly different remote entities are actually different.Firstly all the messages in the network are of broadcast nature; secondly the network has no fixed infrastructure. If a good number of nodes are compromised then the network may totally collapse. Trusted Certification is one of the proposed solutions to sybil attack which requires a central trusted authority. Another proposed solution to this problem is an RSSI (Received Signal Strength Indication) based solution in which the physical location of the nodes are calculated. In this paper we attempt to combine the two proposed methods into a more efficient and practical solution to thwart the sybil attack.

## 2. Related Work

### *A. Trusted Certification*

One solution to the sybil attack is to assign unique node-Ids to each node in the network with the help of a central trusted authority. We use a set of trusted certification authorities (CAs) to assign node-Ids to principals and to sign node-Id certificates, which bind a random node-Id to the public key that speaks for its principal and an IP address. The CA's ensure that node-Ids are chosen randomly from the id space, and prevent nodes from forging node-Ids. Furthermore, these certificates give the overlay a public key infrastructure, suitable for establishing encrypted and authenticated channels between nodes. None of the known solutions to node-Id assignment are effective when the overlay network is very small.

For small overlay networks, we must require that all members of the network are trusted not to cheat. Only when a network reaches a critical mass, where it becomes sufficiently hard for an attacker to muster enough resources to control a significant fraction of the overlay, should mistrusted nodes be allowed to join.

## 3. B. RSSI Based solution

In this solution there is a detector node that calculates the RSSI ratio for each pair of nodes in the network. Suppose D1, D2, D3, D4 be the detector nodes and let a compromised node have 2 IDs S1 and S2. At time t1, a sybil node broadcasts a message with its forged ID as S1. Monitoring nodes record the RSSI and the forged ID. Each monitoring node sends a message to D1 containing the received RSSI from S1. Let Rki denote the RSSI value when a message from a sender k is received at i. Then, accumulating the messages from the monitors, D1 computes each ratio

*(Rs1d1/Rs1d2)*, *(Rs1d1/Rs1d3)* and *(Rs1d1/Rs1d4)*

and stores them locally. At time t2, the sybil node broadcasts a message again with a different ID, S2. The monitoring nodes record the RSSI from S2 and report to D1. D1 computes each ratio as before:

*(Rs2d1/Rs2d2)*, *(Rs2d1/Rs2d3)* and *(Rs2d1/Rs2d4)*

Now, D1 can detect the sybil node by comparing the ratio at time t1 and t2. If the difference between two ratios is very close to zero, D1 concludes that a sybil attack occurred in the region. Since RSSI ratios are same, the location is in fact the same for the alleged multiple IDs. Otherwise, D1 concludes that there is no sybil node. That is, if

*((Rs1d1 / Rs1d2) = ( Rs2d1 / Rs2d2))*
*((Rs1d1 / Rs1d3) = (Rs2d1 / Rs2d3))*
*and ((s1d1 / Rs1d4) = (Rs2d1 / Rs2d4))*

is true, then D1 detects a sybil attack.

## 4. The Proposed semi - centralized Solution

### Description of Notations:

**Inputs provided**:

- V →Velocity of the central authority C
- R → Approximate radius of area occupied by a single node in the subgroup

### Output:

- N → approximate threshold value of the subgroup

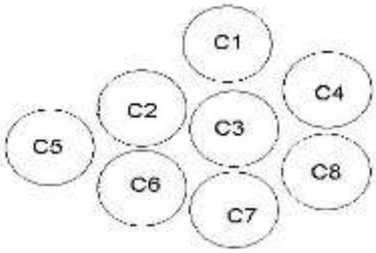


Fig.1 → Total network divided into logical subgroups
Ci = Central authority of each subgroup

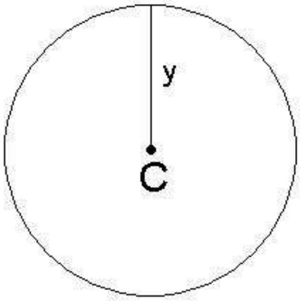


Fig.2 → Each subgroup
C = Central authority of the subgroup
y = distance of C from farthest node in subgroup

We have tried to combine the above two solutions in generating a new solution to detect Sybil nodes in a network. The main disadvantage of the central authority based solution is that it is a centralized solution which is not entertained in ad hoc networks and will be a serious bottleneck when network size increases. The second solution which is a hardware based solution is approximate in nature and must be verified before any node can be removed from the network, thus requiring something like a central authority.

The biggest disadvantage of any central authority based scheme is that if the central authority is compromised, the whole network falls apart and all nodes become vulnerable to the malicious nodes.

As we know that the only way to somehow eliminate Sybil attack is to implement a central authority based scheme, we have tried to distribute the authority of this central node as far as possible. We assume that the total network will be dynamically divided into several subgroups as more and more nodes will enter the network. Each subgroup will be under the super vision of a single node, a central authority. Each subgroup will also contain RSSI detector nodes. The number of nodes for each subgroup is dynamically calculated taking in consideration the mobility of the central authority and the terrain where the subgroup is present. Whenever the number of nodes exceeds this threshold value, a new subgroup will be created and a new trusted node will be assigned as the central authority of that subgroup. The solution may be algorithmically stated as follows:

1. The network starts with n number of nodes. One trusted node assumes the responsibility of central authority C.

2. C calculates the threshold value, which determines the maximum number of nodes that can be present in that subgroup. C assigns a suitable number of nodes as RSSI detectors $R_i$. The number of RSSI detectors required for the subgroup is calculated from the threshold value of the subgroup by the central authority.

3. He manually assigns a unique identity to each node present in the network, but does not monitor the nodes once an identity is assigned to them.

4. The $R_i$'s take over at this point of time. They constantly monitor all nodes in the network, calculating and comparing the ratios of the RSSI values obtained for each node by at least 2 $R_i$'s.

5. C has the responsibility to constantly monitor the $R_i$'s manually such that they are not compromised. If the $R_i$'s declare Sybil attack has occurred at a particular location, C manually checks whether the node is indeed a malicious node or not. Threshold value n is calculated in such a way that maximum time taken by C to travel across the subgroup is optimal (around 1 min).

6. If C finds out that accused node is indeed malicious, the node is removed and the identities that were being used by the node are marked as available.

7. A point of time will come when each subgroup gets saturated. At this point the central authority will appoint a new node as the central authority of a new node. Then it will redirect all new requests to join its own subgroup to the newly created subgroup. Each central authority will be synchronized with the new subgroup it has created. In this way the initial central authority will redirect a new node to the next subgroup. If that too is saturated then it will be redirected to the next and so on.

8. When a new node will enter the network and request to be registered with the network, it might happen that two central authorities with unsaturated subgroups will receive the request simultaneously and respond. At this point of time the new node will have the liberty to choose any subgroup arbitrarily.

## Derivation of Threshold Value:

### Inputs provided:

- V →Velocity of the central authority C
- R → Approximate radius of area occupied by a single node in the subgroup

**Output:**

- N → Approximate threshold value of the subgroup

The calculations are as follows:

$$V_1 = V*k_2 \quad (1)$$
$$y = V_1*k_1 \quad (2)$$

Where $k_1$ = maximum time taken by C to travel across the node.

$k_2$ = **Terrain constant** < 1 and is dynamically determined by the surrounding conditions and terrain of the subgroup. This is done as the velocity will reduce in those extreme conditions.
Y is maximum possible radius of the subgroup.

$$N = y^2 / R^2 \quad (3)$$

**Explanation:**

Since velocity of C is V, and $k_1$ is the maximum time to travel across the subgroup, value of y is $V_1*k_1$. Thus the area is лy$^2$. Area occupied by each node is лR$^2$. So number of nodes is given is given by dividing them and result is given in (3).

## 5. Conclusion

Our solution combines two robust solutions and hence is robust. But there are a few points of concern. Firstly if the adhoc network finally has n number of subnets then initially there must be at least n trusted nodes. Otherwise there is a chance that one of the certifiers become compromised, disrupting the entire group. Secondly if one of the detectors of a group is compromised there might be some trouble. The detector may send false RSSI readings, thus creating chaos in a group. But here the advantage is that even if this happens, the problem would be bounded within the particular group only. Hence the situation would never get out of hand. So we hope this solution would make adhoc networks more secure and efficient at the same time.

## References

1. M.Mohsin and R.Prakash, *ip address assignment in mobile ad hoc networks*.
2. C.E.Parkins and P.Bhagwat, *highly dynamic DSDV routing for mobile computers.*
3. C.Karlof and D.Wagner, *secure routing in wireless sensor networks: attacks and counter measures.*
4. M.Demirbas and Y.Song, *an RSSI-based scheme for sybil attack detection in Wireless Sensor Networks.*
5. A.Ghaffari, *vulnerability and security of mobile ad hoc networks.*
6. J.R.Douceur, *the sybil attack.*
7. B.N.Levine, C.Shield and N.B.Margolin, *a survey of solutions to the sybil attack*
8. M.Castro, P.Druschel, A.Ganesh, A.Rowstron and D.S.Wallach, *secure routing for structured peer-to-peer overlay networks*
9. J.Newsome, E.Shi, D.Song and A.Perrig, *the sybil attack in sensor networks: analysis & defenses.*
10. H.Yu, P.B.Gibbons and M.Kaminsky, *brief announcement: toward an optimal social network defense against sybil attacks*
11. Issa Khalil, Saurabh Bagchi, Ness B. Shroff, *MOBIWORP: Mitigation of the wormhole attack in mobile multihop wireless networks*
12. W. Zhang, G.Cao, *Defending against cache consistency attacks in wireless adhoc networks*

## Author Biographies

**Prof Himadri Nath Saha :**Prof. Saha is graduated from Jadavpur University.He did his post graduate degree from Bengal Engineering and Science university.He is Assistant Professor of Institute of Engg and Management .His research interest is security in MANET.

**Prof.(Dr)Debika Bhattacharyya**:

Prof.Bhattacharyya did Phd. From Jadavpur University in the dept. of ETCE. She is HOD in the Dept of CSE.Her research Interest is security in MANET

**Prof.(Dr) P. K. Banerjee:** Prof. Banerjee is retired professor of Jadavpur University in the dept. of ETCE. His research interest is security in MANET.

# A Mathematical Formula for the Search Engine Ranking Efficiency Evaluation Tool S.E.R.E.E.T

Moiez Tapia[1], Wadee S. Alhalabi[2], Miroslav Kubat[3]
[1,3]Electrical and Computer Engineering Department, University of Miami, USA
[2]Computer Technology Department, College of Technology, KSA
[1]mtapia@mimai.edu,[2]w.alhalabi@umiami.edu,[3]mkubat@miami.edu

*Abstract*- The rapid developments in the field of internet search engines underline the need for reliable method to evaluate its performance. So far, the vast majority of researchers have relied on the "precision" and "recall" measures known from the field of Information Retrieval Unfortunately, both of them fail to assess how successfully they rank the returned documents according to their relevance. In this paper, we discuss this issue in some detail, and then propose a new mechanism for the evaluation of the quality of search-engine rankings.

***Index Terms—SEREET, Search Engine Ranking, Information Retrieval Evaluation***

## I. INTRODUCTION

THE Internet revolution gave rise to the search engine, the only tool capable of identifying among the billions of web sites those that are relevant to the user's needs. Starting from mid 1990s, hundreds of companies specializing on these tools have appeared. Many of them have gone out of business; others have merged, and yet others have joined this thriving market only recently, seeking either to outperform their predecessors, or to fit previously unexplored niches.

The principle of this tool is simple. Upon the entry of user's query, the search engine analyzes its repository of stored web sites and returns a list of relevant hyperlinks ordered by the relevance of the web sites to what the user needs. Many mechanisms to assess this relevance have been exploited, among them keyword frequency, page usage, link analysis, and various combinations of these three. Each of the multitude of alternative ranking algorithms leads to a different hyperlink ordering. Hence it becomes necessary to determine as to which of these algorithms yields the best results in terms of offering the most realistic set of hyperlinks to an average user query.

A two-pronged strategy is necessary if the question is to be answered in a satisfactory manner. First, we need appropriate experimental procedures that submit to the machine well-selected testing queries to which the relevant answers are known. Second, we need performance criteria to evaluate the quality of the search engine responses to the testing queries.

In this paper, we focus on the latter aspect. As discussed in the next section, the previous research has predominantly used the current classical performance metrics of precision and recall that are commonly used in the field of Information Retrieval. However, the utility of these metrics for search engine evaluation is limited: precision and recall establish whether the returned list contains the predominantly relevant links, and how many relevant links are missing. What they ignore is whether more relevant links find themselves high up in the list.

We begin by an extensive survey of related work in Section 2. Section 3 addresses our method accompanied by examples and comparison with existing algorithms. In section 4, we present the discussion and conclusions of our work.

## II. RELATED STUDIES

Precision and Recall is the most widely used tool to evaluate an information retrieval system. It is used by scientists to evaluate retrieval information systems. Zhang and Dong [8] present a review of many ranking algorithms and discuss the deficiencies in the existing techniques. The authors propose an algorithm with a multidimensional technique and claim an improvement in the ranking result. Their algorithm produces more relevant documents and better precision. Shafi and Rather [20] use Precision and Recall to evaluate the performance of five different search engines. Chu and Rosenthal [11] present the same evaluation criteria for retrieval performance as the work proposed by Shafi and Rather [20]. However, they use precision and response time instead of precision and recall. Li and Danzig [23] introduce a new ranking algorithm. They argue that their technique is much better in space and time complexity. The authors claim that their system has a better precision and recall than the existing algorithms.

New approaches evolved in ranking algorithms with new ideas, but Precision and Recall was used to evaluate the retrieval system. Eastman and Jansen [5] explore the impact of query operators on web search engine results. The authors use coverage, relative precision and ranking as questions trying to answer in their research. Goncalves et al. [15] present an algorithm to measure the effectiveness of a retrieval system as an overall. It measures how much a document is relevant to the query, but it does not compare two retrieval systems. It does not show if a rank of a retrieval system is efficient enough. The authors use Precision and Recall as an evaluation tool. Yuwono et al. [4] explore the relevance feedback in effecting the retrieved documents. The authors use Precision and Recall as a tool to evaluate the ranking efficiency. Hawking et al. [7] tried to answer the question "Can link information result in better PageRank?" The authors discuss the effectiveness of a search engine and its performance by measuring its precision and recall. Yuwono and Lee [3] provide four different ranking algorithms: Boolean Spread Activation, Most-cited, TFxIDF and Vector Spread Activation. The authors use different queries to compare these four algorithms with each other. Their ranking evaluation was based on Precision and Recall.

The hypertext algorithm was a new approach proposed by Brin and Page [22] to improve the ranking of retrieved web pages. The authors claim that this approach would improve the search result by having high precision rank. Baeza-Yates and Davis [16] show that link attribute of a Web Page can improve the ranking by improving the precision of the system. Trotman and O'Keefe [2] use precision to evaluate the ranking algorithm. They depict how a weight is awarded to each document.

Pay per performance (PPP) search engine is a different approach in search engine ranking. Goh and Ang [9] discuss this approach and use precision and recall to evaluate the ranking performance. Ljosland [14] presents a comparison between three search engines: Atavista, Google and Alltheweb. The author uses precision to evaluate the performance of each engine. Bifet and Catillo [1] explore the top web pages appearing in the rank. They also explore the shifted ones. The authors use precision to calculate the efficiency of the rank.

Precision and Recall was used in most ranking evaluation as we saw in previous works. However, many scientists use different evaluation tools. Precision and Recall can evaluate the retrieval system, but they cannot precisely evaluate the efficiency of the rank. Any change in the order of the retrieved documents does not necessarily affect the precision and the recall. This variable (the order of the retrieved documents) cannot be measured using Precision and Recall method.

Clarke and Cormack [6] introduced a new approach toward ranking evaluation. Their work was to evaluate each document and give a specific weight to the document according to all other retrieved documents. They are interested in documents' weight according to other documents. Their method would change the order of the retrieved documents. But it does not evaluate the rank itself. Algorithms for ranking retrieved documents such as these introduced in [22], [12] and [19] were used to rank web pages; however, they still do not measure the ranking algorithm and its efficiency. Kamvar et al. [21] explore many PageRank scheme and provide two algorithms. They present the Adaptive PageRank and the Modified Adaptive PageRank. The authors have not discussed the ranking evaluation in their work. White et al. [18] present an evaluation to encourage user to interact with the search result. They showed how their approach improves the PageRank. However, their paper does not show any tool to numerically evaluate a PageRank.

New evaluation tool other than Precision and Recall were introduced. Losee and Paris [17] oppose the use of Precision and Recall as a measure to evaluate search engine ranking performance. The authors suggest a probability method and proved that their proposed solution would result in a much better evaluation. The authors present the Average Search Length (ASL). ASL finds the average position of the retrieved document. This method is much better than precision in evaluating the ranking performance. However, as the authors mention, a small number of relevant documents in the top of the rank may represent a superior performance. They present the Expected Search Length (ESL) as an alternative approach. This method counts only the non-relevant documents. In this evaluation the system must minimize the ESL value for better performance. The authors [17] advocate our approach in finding an alternative method to measure the performance of a ranking system. Haveliwala [24] compares two different ranking by measuring the degree of similarity. He calculates the degree of overlap between the top URLs of the two ranking lists. Our approach is to find a numerical evaluation for each ranking list rather than comparing the two different ranks.

## III. Proposed Mechanism for Ranking

Precision and Recall is used to evaluate the efficiency of a retrieval system. A large number of relevant documents and a few irrelevant ones give a high system precision. Precision is calculated according to formula (1). It is the ratio of the relevant documents retrieved to the total number of retrieved documents. The recall of a retrieval system is the ratio of the relevant documents retrieved to the relevant documents in the database of the system. It is infeasible to accurately calculate the number of documents in a database of a search engine. It appears that Precision and Recall has some limitations in calculating ranking efficiency.

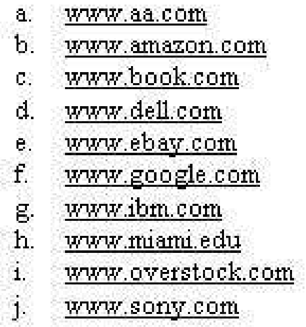
$$precision = \frac{\#\text{of r.d.r}}{(\#\text{of r.d.r} + \#\text{of i.d.r})} * 100\% \qquad (1)$$

where r.d.r: relevant documents retrieved
i.d.r: irrelevant documents retrieved

$$recall = \frac{\#\text{of r.d.r}}{\#\text{of r.d.db}} * 100\% \qquad (2)$$

where r.d.r: relevant documents retrieved
r.d.db: relevant documents in database

In all cases, we might have the tool to find the number of relevant documents retrieved, and the number of irrelevant documents retrieved. However, the number of relevant documents that exist in a database cannot be found. Therefore, we are certainly unable to calculate the recall value precisely. The problem with precision is presented in example 1, 2 and 3.



a. www.aa.com
b. www.amazon.com
c. www.book.com
d. www.dell.com
e. www.ebay.com
f. www.google.com
g. www.ibm.com
h. www.miami.edu
i. www.overstock.com
j. www.sony.com

Fig1. The retrieved web pages

**Example 1**: Suppose we have the retrieved web pages in the order shown in figure 1. Suppose the following web pages are the only relevant documents to the query and the other documents (web pages) are irrelevant.

a. www.aa.com
b. www.amazon.com
e. www.ebay.com
g. www.ibm.com
i. www.overstock.com
j. www.sony.com

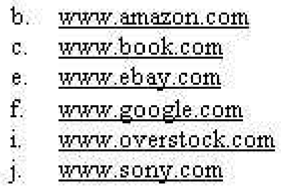
Then

$$precision = \frac{\#\text{of r.d.r}}{(\#\text{of r.d.r} + \#\text{of i.d.r})} * 100\%$$

$$precision = \frac{6}{6+4} * 100\% .$$

$$= 60\ \%$$

**Example 2:** Suppose the following documents are the only relevant documents and the other documents are irrelevant.



b. www.amazon.com
c. www.book.com
e. www.ebay.com
f. www.google.com
i. www.overstock.com
j. www.sony.com

Suppose the search engine ranks the web pages in the order shown in Figure1:

Then

$precision = \frac{6}{6+4} *100\%$ = 60%.

**Example 3**: Suppose the following web pages are the only relevant documents and the other documents are irrelevant:

a. www.aa.com
b. www.amazon.com
c. www.book.com
d. www.dell.com
e. www.ebay.com
f. www.google.com

Suppose the search engine ranks the web pages in the order shown in Figure1:
Then

$precision = \frac{6}{6+4} *100\%$

= 60 %.

With three different sequences, we found that the precision does not change. Using precision tells us that the efficiency of the three retrieval systems is the same in the three systems. However, examples 1, 2 and 3 show that we have three systems. These systems have totally different sequences and they should have different ranking efficiencies.

We propose the *Search Engine Ranking Efficiency Evaluation Tool* (S.E.R.E.E.T.) to distinguish among any ranking systems. The purpose of this algorithm is to numerically evaluate the efficiency of the search engine rank.

**<u>Definition</u>**
Let there be m≥0 hits and n≥0 misses, (m+n)≥1.

Let i, $1 \leq i \leq m + n$

represents the position of a website name on the search output list that is hit or missed, and let the position of the $j_{th}$ hit,

$1 \leq j \leq m$

be given by $h_j$ Obviously 1<= $h_j$ <= (m+n), for all j.

Let the $W_i$ denote the weight of the $i_{th}$. website name, where

$1 \leq j \leq m + n$

* Define $W_i$ as follows:
$W_i = m + n + 1 - i$, if the $i_{th}$ name is a hit.
$W_i = 0$, if the $i_{th}$ name is a miss.

Then the efficiency of ranking of search engine is given by formula (3):

$$E= \sum_{i=1}^{i=m+n} Wi * \frac{2}{(m+n)*(m+n+1)} *100\% \qquad (3)$$

**Example 4:**
Consider there are m=5 hits and n=4 misses in the order shown below:
1. h1
2. m1
3. m2
4. h2
5. h3
6. h4
7. h5
8. m3
9. m4

w1= m+n+1-1= 5+4 =9
w2=0
w3=0
w4= m+n+1-4 =6
w5= m+n+1-5=5
w6= m+n+1-6=4
w7= m+n+1-7=3
w8=0
w9=0

Then

$$E = (\sum_{i=1}^{i=m} Wi) * \frac{2}{(m+n)*(m+n+1)} * 100\ \%$$

$$= (9+6+5+4+3)*\frac{2}{(5+4)*(5+4+1)} *100\%$$

$$= 27 * \frac{2}{9*10} *100\ \%$$

$$= \frac{54}{90} *100\ \%$$

= 60 %

**<u>Lemma 1</u>**: For integer m > 0,

$$\frac{m*(m+1)}{2}$$

$1+2+...+m$ =

Proof:

Let S(m) = 1+2+ … + m (L1.1)
We can rewrite

S(m) = m+ m-1+ … +1 (L1.2)

Adding both the sides of (L1.1) and (L1.2),

we get 2*S (m) = m * (m+1)

Hence

$$S(m) = \frac{m*(m+1)}{2}$$

Hence we have proved the lemma.

**Theorem 1:**

If there are no misses, the efficiency of the search engine is 100%.

Proof:
If there are no misses, n=0.
Hence Efficiency (E) :

$$E = (\sum_{i=1}^{i=m} Wi) * [\frac{2}{(m+n)*(m+n+1)}] * 100 \%$$

$$E= (\sum_{i=1}^{i=m} Wi) * [\frac{2}{(m)*(m+1)}] * 100 \% \quad (T1.1)$$

Also, $Wi = m+1-i$

Hence from (T1.1),

Efficiency

$$E= (\sum_{i=1}^{i=m} Wi) * [\frac{2}{(m)*(m+1)}] * 100 \% \quad (T1.1)$$

$$= (m+1-1+m+1-2+...+m+1-m)*[\frac{2}{(m)*(m+1)}]*100\%$$

$$= (m+m-1+m-2+...+1)*[\frac{2}{m*(m+1)}]*100\%$$

$$= \frac{m*(m+1)}{2} * \frac{2}{(m)*(m+1)} * 100\% \quad \text{from Lemma 1.}$$

= 100%.

**Theorem2:**

Let there be m hits, m≥0, and n misses, n≥0, (m+n)≥1, for two website name lists $W_1$ and $W_2$. $W_1$ is such that it has its hits only positions i , 1 ≤ i ≤ m. $W_2$ is such that it has one miss in position M, 1≤ M ≤ m, all other hits in position i , 1 ≤ i ≤ m and one hit in position H, m < H ≤ ( m + n ). Then E1 >E2 where $E_1$ and $E_2$ are efficiencies for $W_1$ and $W_2$, respectively.

Proof:
We have

$$E_1 = \sum_{i=1}^{i=m} Wi * \frac{2*100}{(m+n)+(m+n+1)}$$

$$= K.[\sum_{\substack{i=1 \\ i\neq M}}^{i=m} W_i + W_M]$$

$$\text{where } K = \frac{2*100}{(m+n)+(m+n+1)}$$

$$\therefore E_1 = K[\sum_{\substack{i=1 \\ i\neq M}}^{i=m} Wi + (m+n+1-M)] \quad (2.1)$$

$$E2 = K * \sum_{\substack{i=1 \\ i\neq M}}^{i=m+n} Wi$$

$$= K[\sum_{\substack{i=1 \\ i\neq H}}^{i=(m+n)} W_i + W_H]$$

$$= K[\sum_{\substack{i=1 \\ i\neq H}}^{i=m} Wi + W_H]$$

$$= K[\sum_{\substack{i=1 \\ i\neq M \\ i\neq H}}^{i=m} Wi + W_M + W_H]$$

$$= K[\sum_{\substack{i=1 \\ i\neq M \\ i\neq H}}^{i=m} Wi + 0 + (m+n+1-H)]$$

$$= K[\sum_{\substack{i=1 \\ i\neq M \\ i\neq H}}^{i=m} Wi + (m+n+1-H)] \quad (2.2)$$

Observe that H > m > M by hypothesis (2.3). Hence comparing (2.1) and (2.2) in view of the fact (2.3),
we have

$E_1 > E_2$

**Theorem 3**

Let there be m hits, m ≥ 1 and n misses, n ≥ 0, (m+n)≥1 for two website name lists $W_1$ and $W_2$. $W_1$ has all the hits in $1 \le i \le m$ positions i,
,
and all the misses in positions j, $m \le j \le (m+n)$
$W_2$ has p misses in positions k , $1 \le k \le m$
and $0 \le p \le m$
and (n-p) hits in positions l ,$1 \le l \le m$

Then $E_1 > E_2$ .

Proof

We will prove this theorem by induction.

1. By theorem 2, the hypothesis is true for p=1.
2. Assume the hypothesis is true for p = e, e ≤ n

let $E_e$be the efficiency of this website name list denoted by $W_e$

then $E_1 > E_e$ (3.1)

let us exchange the positions of one hit at position s, s<m and one miss at t , t>m.

Thus now we have a new website name list $W_{e+1}$ such that it has

e +1 misses in positions K , $1 \le k \le n$

Let $E_{e+1}$ be the efficiency of $W_{e+1}$

then

$$E_{e+1} = \sum_{i=1}^{m+n} Wi \ * \ \frac{2}{(m+n)+(m+n+1)} * 100$$

$$= k \ [\sum_{\substack{i=1, \\ i \neq s, \\ i \neq t}}^{i=m+n} Wi \ + Ws \ + Wt \ ]$$

Where k = $[\frac{2*100}{(m+n)*(m+n+1)}]$

Hence

$$E_{e+1} = \text{K} \ [\sum_{\substack{i=1, \\ i \neq s, \\ i \neq t}}^{i=m+n} Wi \ + 0 + (m+n+1-t)]$$

$$= \text{K} \quad [\sum_{\substack{i=1, \\ i \neq s, \\ i \neq t}}^{i=m+n} Wi \ + Ws \ + Wt \ ]$$

Since s<m and t>m , we have s<t (3.3)

Also , $E_e$ = k $[\sum_{\substack{i=1, \\ i \neq s, \\ i \neq t}}^{i=m+n} Wi \ + Ws \ + Wt \ ]$

$$= \text{k} \ [\sum_{\substack{i=1, \\ i \neq s, \\ i \neq t}}^{i=m+n} Wi \ + Ws \ + Wt \ ]$$

$$= \text{k} \ [\sum_{\substack{i=1, \\ i \neq s, \\ i \neq t}}^{i=m+n} Wi \ + 0 + (m+n+1-s)] \qquad (3.4)$$

Comparing (3.2) and ( 3.4) and using (3.3), we have

$E_e > E_{e+1}$ (3.5)

From (3.1) we have$E_e > E_{e+1}$

Hence $E_e > E_{e+1}$

Hence the hypothesis is true for p=e+1
Hence by induction, we have proved the theorem.

## IV. 4. Discussion and conclusions

With a close look at many of the evaluation tools used in search engine ranking, we found that those tools do not address the need for ranking evaluation. They only look at document level, but not at the ranking evaluation level. Precision and Recall does not precisely measure ranking efficiency; they rather measure the percentage of good and bad documents in the retrieved bag.

Search Engine Ranking Efficiency Evaluation Tool (S.E.R.E.E.T.) introduced in our earlier work provides a unique tool to precisely measure the ranking efficiency. In this paper we show the mathematical formula behind the SEREET tool. If we have two different ranking algorithms with the same precision, we still can favor one over the other by using SEREET.

## REFERENCES

[1]Albert Bifet and Carlos Catillo. *An Analysis of Factors Used in Search Engine Ranking*. Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), 2005.

[2] Andrew Trotman and Richard A. O'Keefe. *Identifying and Ranking Relevant Document Elements*. In INEX '03, 2003.

[3] B. Yuwono and D.L. Lee. *Search and ranking algorithms for locating resources on the World Wide Web*. 12th International Conference on Data Engineering (ICDE'96) p. 164

[4] Budi Yuwono, Savio L. Y. Lam, Jerry H. Ying and Dik L. Lee. *A World Wide Web Resource Discovery System*. Proceedings of the Fourth International World Wide Web Conference, Boston, MA., Dec. 1995.

[5] Caroline M. Eastman and Bernard J. Jansen. Coverage, Relevance and Ranking: *The Impact of Query Operators on Web Search Engine Results*. ACM Transaction on Information Systems, Vol. 21, No. 4, October 2003, Page 383-411.

[6] Charles L. A. Clarke and Gordon V. Cormack. *Shortest-Substring Retrieval and Ranking*. ACM Transactions on Information Systems, Vol 18, No. 1, January 2000, Pages 44-78.

[7] David Hawking, Nick Craswell, Paul Thistlewaite and Donna Harman. *Results and Challenges in Web Search Evaluation*. Computer Networks, 31(11-16): 1321-1330, May 1999. Also in Proceedings of the 8th International World Wide Web Conference.

[8] Dell Zhang and Yisheng Dong. *An efficient algorithm to rank Web resources*. Proceedings of the 9th international World Wide Web conference on Computer networks : the international journal of computer and telecommunications networking, Amsterdam, The Netherlands, Pages: 449 – 455, 2000 .

[9] Dion H. Goh and Rebecca P. Ang. *Relevancy Rankings – Pay for Performance Search Engines in* the Hot Seat. Online Information Review, Volume 27, Number 2, 2003 , pp. 87-93(7).

[10] Helen Ashman and Paul Thistlewaite, editors. Proceedings of the Seventh International World Wide Web Conference, volume 30 of Computer Networks and ISDN Systems. The International Journal of Computer and Telecommunications Networking, Amsterdam, April 1998. Elsevier. Brisbane, Australia.

[11] Heting Chu and Marilyn Rosenthal. *Search Engines for the World Wide Web: A Comparative Study and Evaluation Methodology*. ASIS 1996 Annual Conference Proceedings, October 1996.

[12] Jon Kleinberg. *Authoritative sources in a hyperlinked environment*. Technical Report RJ 10076, IBM, May 1997.

[13] Miguel Costa and Mario J. Silva. *Optimizing Ranking Calculation in Web Search Engines:* a Case Study, 2004.

[14] Mildrid Ljosland. *Evaluation of Web Search Engine and the Search for Better Ranking Algorithms*. SIGIR99 Workshop on Evaluation of Web Retrieval (1999)

[15] Pedro Goncalves, Jacques Robin, Thiago Santos, Oscar Miranda and Silvio Meira. *Measuring the Effect of Centroid Size on Web Search Precision and Recall*. In Proceedings 8th Annual Conference of the Internet Society (INET'98). Geneva, Switzerland,July,1998. http://www.isoc.org/inet98/proceedings/1x/1x_8.htm.

[16] Ricardo Baeza-Yates and Emilio Davis. *Web Page Ranking using Link Attributes*. In Alt. track papers & posters, WWW Conf., pp. 328-329, New York, NY, USA, 2004.

[17] Robert M. Losee and Lee Anne H. Paris. *Measuring Search Engine Quality and Query Difficulty: Ranking with Target and Freestyle*. Journal of the American Society for Information Science archive Volume 50 , Issue 10 (1999), Pages: 882 – 889, 1999 .

[18] Ryen W. White, Ian Ruthven and Joemon M. Jose. *Finding Relevant Documents using Top Ranking Sentences: An Evaluation of Two Alternative Schemes*. Proceedings of the 25th Annual International ACM SIGIR Conference (SIGIR 2002). Tampere. Pages 57-64. 2002.

[19] S. Lawrence and C. L. Giles, "*Inquirus, the NECI meta search en-gine,*" in Seventh International World Wide Web Conference, (Bris-bane, Australia), pp. 95–105, 1998.

[20] S. M. Shafi and Rafiq A. Rather. *Precision and Recall of Five Search Engines for Retrieval of Scholarly Information in the Field of Biotechnology*. Webology , Volume 2, Number 2, August, 2005 .

[21] Sepandar Kamvar, Taher Haveliwala and Gene Golub. *Adaptive Methods for the Computation of PageRank*. Numerical Solution of Markov Chains, 2003, pages 31-44.

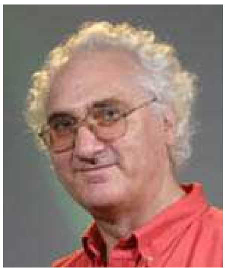
[22]Sergey Brin and Lawrence Page, "*The anatomy of a large-scale hypertextual Web search engine,*" Computer Networks and ISDN Systems archive, Volume 30 , Issue 1-7 Pages: 107 – 117, 1998.

[23] Shih-Hao Li and Peter B. Danzig. *Precision and Recall of Ranking Information Filtering Systems*. Journal of Intelligent Information Systems 1-22, Kluwer Academic Publishers, Boston. 1996.

[24]Taher H. Haveliwala. *Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search*. IEEE Transactions on Knowledge and Data Engineering, 15(4):784-796. IEEE, August 2003.

Moiez Tapia is a Professor in the Department of Electrical andComputer Engineering at the University of Miami. His research interests are in the field of Multi-valued logic and calculus, real-time systems, machine learning, fault-tolerant computing.

Wadee Alhalabi, received his MSECE and PhD from the University of Miami in 2004 and 2008. He is now the director of the Virtual Reality Therapy and Rehabilitation Research Center in KSA.



Miroslav Kubat is an Associate Professor in the Department of Electrical and Computer Engineering at the University of Miami. His research interests are in the field of machine learning, data mining and artificial neural networks.

# Application of GA and PSO to the Analysis of Digital Image Watermarking Processes

Surekha P[1], and S.Sumathi[2]

[1]Research Scholar, EEE, PSG College of Technology, Coimbatore
[2]Asst. Professor, EEE, PSG College of Technology, Coimbatore
Email: surekha_3000@yahoo.com

***Abstract*** – The increasing effect of illegal exploitation and imitation of digital images in the field of image processing has led to the urgent development in the growth of copyright protection methods. Digital watermarking has proved best in protecting illegal authentication of data. In this paper, we propose a digital image watermarking scheme based on computational intelligence paradigms like Genetic Algorithm (GA) and Particle Swarm Optimization (PSO). The input digital host images undergo a set of pre-watermarking stages like image segmentation, feature extraction, orientation assignment, and image normalization to obtain image invariance properties when subject to attacks. Expectation Maximization (EM) algorithm is used to segment the images and the features are extracted using Difference of Gaussian (DoG) technique. The feature maps from the feature extraction methods locate the magnitude by orientation assignment making the circular regions invariant. The resultant image is normalized by scaling to acquire the scaling invariance for the circular region. The watermark image is then embedded into the host image using Discrete Wavelet Transform (DWT). During the extraction process, GA, and PSO are applied to improve the robustness, and fidelity of the watermarked image by evaluating the fitness function. The perceptual transparency and the robustness of the watermarked and the extracted images are evaluated by applying filtering attacks, additive noise, rotation, scaling and JPEG compression attacks to the watermarked image. From the simulation results the performance of the Particle Swarm Optimization technique is proved best based on the computed robustness and transparency measures along with the evaluated parameters like elapsed time, computation time and fitness value. The performance of proposed scheme was evaluated with a set of 50 textures images taken from online resources of Tampere University of Technology, Finland and the entire algorithm for different stages was simulated using MATLAB R2008b.

***Keywords*** – Expectation Maximization, Difference of Gaussian, Orientation Assignment, Image Normalization, DWT, Genetic algorithm, Particle Swarm Optimization.

## 1. Introduction

Effective digital image copyright protection methods have become a vital and instantaneous necessitate in multimedia applications due to the increasing unauthorized manipulation and reproduction of original digital objects. Multimedia data protection has become one of the interesting challenge and has drawn the attention of researchers towards the development of protection approaches. Among several protection methodologies, digital watermarking is the leading approach, to protect illegal authentication of data [1]. Digital image watermarking is a technique to embed a secret message or valuable information (watermark) within an ordinary image (host image) and extract the image at the destination, therefore protecting the image from common image processing attacks during the transmission process. The watermark can either be a random signal, an organization’s trademark symbol, or a copyright message for copy control and authentication. The chosen watermark to be embedded in the host image should be resilient to standard manipulations of unintentional as well as intentional nature, should be statistically unremovable, and must be capable of withstanding multiple watermarking to facilitate traitor tracing [2]. The type of manipulations and the signal processing attacks on the watermark depend upon the specific application of digital image watermarking. The most important characteristics of digital watermarking such as imperceptibility, robustness, inseparability, security, provable, permanence, data capacity, and fidelity allow the technique applicable in owner identification, copyright protection, image authentication, broadcast monitoring, transaction tracking, and usage control.

The frequently used watermarking techniques are spatial domain watermarking and frequency domain watermarking. Spatial domain was the first watermarking scheme, in which the perceptual information about the image was obtained and used to embed the watermarking key in the predefined intensity regions of the image. Embedding an invisible watermark was more simple and effective in the spatial domain, but when subject to image alterations the robustness was poor [1]. In the frequency domain, the watermark is transformed into the frequency domain by application of Fourier, discrete cosine or the discrete wavelet transforms. The watermarks are added to the transform coefficients of the image instead of modifying the pixels, thus making it difficult to remove the embedded watermark. Compared to spatial domain technique, frequency domain techniques are more robust and have a high range of control in maintaining the perceptual quality of the watermark. Discrete Wavelet Transform (DWT) is one of the most attractive transform domain watermarking techniques since it is a computationally efficient version of the frequency models for the human visual system [3]. DWT has exceptional properties like excellent localization in time and frequency domain, symmetric spread distributions and multiresolution characteristics [4] which led to the development of various DWT based algorithms.

In this paper, initially, the host image is segmented into a number of homogeneous regions using the Expectation Maximization (EM) algorithm and the feature points are extracted based on the difference of Gaussian (DoG) algorithm. Then the circular regions based on image normalization and orientation assignment are defined for

DWT based watermark embedding or extraction process. There has been a considerable amount of research proposals on the applications of Discrete Wavelets Transform in digital image watermarking systems by virtue of its excellent and exceptional properties mentioned above, but the scope of optimization in this area is tremendously less. An optimized DWT for digital image watermarking is capable of producing perceptual transparency and robustness among the watermarked and the extracted images. During the past few years, evolutionary intelligent algorithms such as genetic algorithm (GA) and particle swarm optimization (PSO) have shown good performances in optimization problems [5] [6]. Moreover watermark techniques based on these evolutionary intelligent algorithms seemed to improve security, robustness, and quality of the watermarked images [7]. We propose a technique based on the evolutionary optimizers to choose the best geometric positions for embedding and extracting thereby preserving the information in watermarked image and preventing the loss of data due to geometric attacks, filtering attacks and JPEG compression. From the simulation results the performance of the Particle Optimization technique is proved best based on the evaluated parameters like robustness, transparency, elapsed time, CPU time and fitness value. The similarity measures of the extracted watermark and the transparency is maintained in the proposed method.

### 1.1 Related work

Dong Zheng, Sha Wang, and Jiying Zhao [8] proposed the watermark embedded and extraction scheme using the RST Invariant Image Watermarking Algorithm with Mathematical Modeling and Analysis of the Watermarking Processes. The basic experimental idea is extracted from this paper. The mathematical relationship between fidelity and robustness is established in their work. Though the experimental results show the effectiveness and accuracy in watermarking, the attacked watermark was not optimized. David G. Lowe [9], proposed a method for extracting distinctive invariant features from images that can be used to perform reliable matching between different views of an object or scene. They also explained about the steps involved in difference of Gaussian algorithm and to detect the feature points from the image without any loss of information by mathematical calculations. Several approaches have been proposed in literature to prove that application of GA to DWM provides effective robustness.

Ali Al-Haj et al. [11] proposed a technique to obtain optimal DWT-based image watermarking only if watermarking has been applied at specific wavelet sub-bands with specific watermark amplification values. This approach concentrated only on a few attacks with fixed values. *Zhicheng Wei* et al [12] proposed an algorithm that yielded a watermark that is invisible to human eyes and robust to various image manipulation, and the results showed that only some specific positions were the best choices for embedding the watermark. The authors applied GA to train the frequency set for embedding the watermark and compared their approach with the Cox's method to prove robustness. The analysis of GA was restricted to JPEG compression attack in their proposed method. In [13], Jin Cong et al proposed a scheme that does not require the original image because the informations from the shape specific points of the original image were been memorized by the neural network. This scheme applies the shape specific points technique and features point matching method by genetic algorithm for resisting geometric attacks. Simulations have confirmed that their scheme has high fidelity and is highly robust against geometric attacks and signal processing operations such as additive noise and JPEG compression.

G. Boato [14] et al. proposed a new flexible and effective evaluation tool based on genetic algorithms to test the robustness of digital image watermarking techniques. Given a set of possible attacks, the method finds the best possible un-watermarked image in terms of Weighted Peak Signal to Noise Ratio (WPSNR). Chin-Shiuh Shieh [15] proposed an innovative watermarking scheme based on genetic algorithms (GA) in the transform domain considering the watermarked image quality. Zne-Jung Lee et al. [16] proposed a hybrid technique, where the parameters of perceptual lossless ratio (PLR) were determined for two complementary watermark modulations. Furthermore, a hybrid algorithm based on genetic algorithm (GA) and particle swarm optimization (PSO) is simultaneously performed to find the optimal values of PLR instead of heuristics. In this approach, GA's crossover and mutation are performed in parallel with particle velocity update of PSO, which sometimes tends to get locked in the local optima without reaching the best solution. Ziqiang Wang et al. [17] proposed a novel blind watermark extracting scheme using the Discrete Wavelet Transform (DWT) and Particle Swarm Optimization (PSO) algorithm. In his work, the experimental results show that the proposed watermarking scheme results in an almost invisible difference between the watermarked image and the original image, and is robust only to JPEG lossy compression.

Almost all the related work concentrated Digital watermarking based on JPEG attacks, and in very few papers additional attacks like rotation and scaling are dealt. The performance of DWT based watermarking can be evaluated best for robustness by applying different attacks with varying parameters. We propose the hybrid technique for embedding and extracting watermarks thus conserving the information in watermarked image and also avoiding the hacking of data due to geometric attacks, additive noise attacks, filtering attacks and JPEG compression.

The sections of the paper are organized as follows: Section 2 deals with the details and the scheme of the digital image watermarking approach including image segmentation, feature extraction, orientation assignment, image normalization and DWT based watermarking. The optimization techniques Genetic Algorithm, Particle Swarm Optimization, and Hybrid Particle Swarm Optimization, its operators and the procedure for DWM techniques are elaborated in Section 3. The experimental analysis and results are explained in Section 4 and Section 5 deals with the conclusion.

## 2. Digital Image Watermarking Scheme

The proposed watermarking scheme is illustrated in Figure 1. In the scheme, first, the image is segmented into a number of homogeneous regions by Expectation Maximization (EM) algorithm and by applying the DoG filter the feature points are extracted [8]. Based on image normalization and orientation assignment, the circular regions are chosen for watermark embedding and extraction. In this section, the phases of the watermarking scheme are discussed.

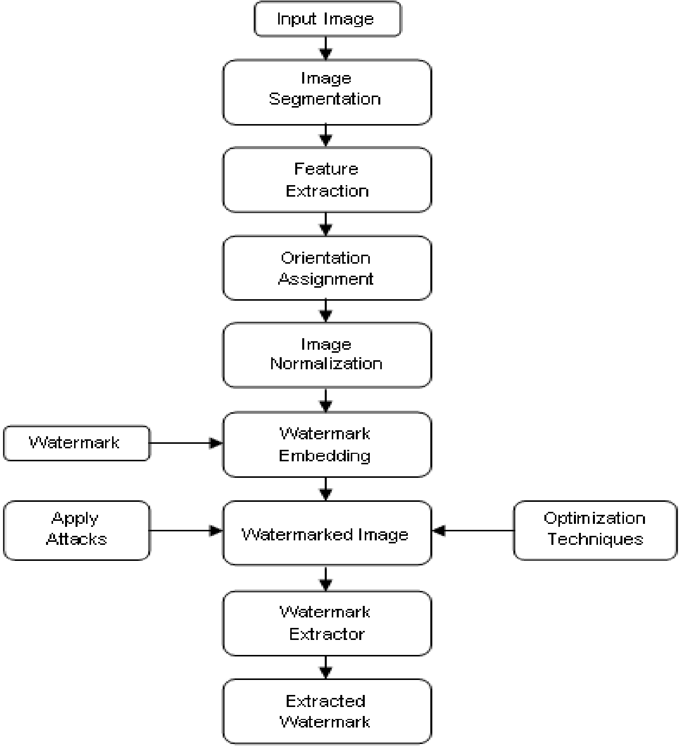


**Figure 1** Watermark Embedding and Extraction Scheme

## 2.1 Image Segmentation

Image segmentation is an elegant technique in which a signal is decomposed into segments with different time and frequency resolutions. The goal of segmentation is to simplify and change the representation of an image interms of different homogeneous regions that is more meaningful and easier to analyze. Expectation-maximization algorithm is used for image segmentation and hence to determine the embedding strength of the segmented homogenous region. The mean vector and the covariance matrix of each homogeneous area is computed and this aids in implementing the watermark embedding process [18]. The EM is an iterative procedure alternating between an expectation (E) step and a maximization (M) step. The E step computes an expectation of the log likelihood with respect to the current estimate of the distribution for the latent variables. The M step computes the parameters which maximize the expected log likelihood found in the E step [19] [20]. These parameters are then used to determine the distribution of the latent variables in the next iterative E step. Assume the observed image is y and the segmentation region is x, for each subregion $x_s$, the conditional distribution of $y_s$ (subregion of y) given $x_s$ is a Gaussian distribution with mean and variance. The mixture Gaussian distribution is used to model the observed image y by eq. 1,

$$Py_s|x_s(y_s|x_s) \sim N(\mu y_s),\sigma y_s) \quad (1)$$

Where, $\mu y_s$ is the mean of the segmented image, $\sigma y_s$ is the variance of the segmented image and N is the number of segmented regions. The density is given as,

$P_{yx}$ (y,x) = $\sum_{s\varepsilon S} m_s \, p_{ys\ xs}(\frac{y_s}{x_s})$ ,where $m_s$ is the mixture weighting factor (2)

For each distribution, the mean vector and covariance are set to predefined initial values. The covariance can be set as the identity matrix and the mean is calculated by determining the average of different regions of the image. Then the probability of the pixel falling into one of the Gaussian distribution can be calculated according to eq.2. The mixture weighting factor, mean and variance are evaluated based on eqs. 3-5.

$$m_s = \frac{1}{N} \sum_{j=1}^{N} P(s, y_j) \quad (3)$$

$$\mu y_s = \frac{\sum_{j=1}^{N} y_j P(s \mid y_j)}{\sum_{j=1}^{N} P(s \mid y_j)} \quad (4)$$

$$\sigma_{ys} = \frac{\sum_{j=1}^{N} P(s \mid y_j)(y_j - \mu_{y_S})(y_j - \mu_{y_S})}{\sum_{j=1}^{N} P(s \mid y_j)} \quad (5)$$

The EM algorithm is designed to deal with closed form solution problem but it frames the image into different homogeneous regions. Based on the intensity of the homogenous regions, the high frequency components can be evaluated. The darker the regions, the larger the variance and hence more high-frequency components are available in the segmented regions of the image. Though several image segmentation algorithms are available we used the expectation maximization since it provides a simple, easy-to-implement and efficient tool for learning parameters of a model, and it also presents a mechanism for building and training rich probabilistic models for image processing applications [19].

## 2.2 Feature Extraction

The huge set of data available in images is simplified for analysis by the technique known as feature extraction. The number of variables in the large data set often causes several problems while analyzing the complex data. The complex data variables inturn require a large amount of memory and computation power which overfits the training sample and generalizes poor samples. Thus, feature extraction is one of the efficient method of constructing combinations of the data variables maintaining the data with sufficient accuracy.

DoG algorithm [9] was proposed by David Lowe in 1999. Difference of Gaussian (DoG) is a grayscale image enhancement algorithm that involves the subtraction of one blurred version of an original grayscale image from another less blurred version of the original image. The original gray scale images are convolved with Gaussian kernels having different standard deviations to obtain the blurred images. This process of blurring suppresses only high-frequency spatial information retaining the other information. The spatial information within the range of frequencies are preserved in both the blurred images. This kind of a technique is similar to a band-pass filter that discards all but a handful of spatial frequencies that are present in the original grayscale image.

While extracting features, the difference of Gaussian technique enhances the visibility of edges in a digital image. A wide variety of alternative edge sharpening filters operate by enhancing high frequency detail, but because random noise also has a high spatial frequency, many of these sharpening filters tend to enhance noise, which can be an undesirable artifact. Usually the high frequency detail that often includes random noise is removed by the DoG approach. The cost of extracting the features determined from the original is minimized by taking a cascade filtering approach, in which more expensive operations are applied only at locations that pass an initial test. The major stages of computation to generate the set of image features are:
Scale-space extrema detection: The overall scales and the image locations are determined in the initial state. DoG algorithms is applied to identify potential interest points that are invariant to scale and orientation.
Keypoint localization: At each candidate location, a detailed model is fit to determine location and scale from which the keypoints are selected based on measures of their stability [9].

To efficiently detect stable keypoint locations, scale-space extrema in the difference-of-Gaussian function is convolved with the image, D(x, y, σ ), which can be computed from the difference of two nearby scales separated by a constant multiplicative factor k by using eq.6:

$$D(x,y,\sigma) = \left(G(x,y,k\sigma) - G(x,y,\sigma)\right) * I(x,y)$$
$$= \mathrm{L}(\mathrm{x},\mathrm{y},\mathrm{k}\sigma) - \mathrm{L}(\mathrm{x},\mathrm{y},\sigma) \qquad (6)$$

The local extrema in the DoG are found and by removing those with strong edge responses, the final results are selected as the feature points.

### 2.3 Orientation Assignment

Orientation assignment is the key step in achieving invariance to rotation as the keypoint descriptor can be represented relative to the orientation of the image [21]. Each circular region is made rotation invariant by defining a window centered at the chosen feature point. For all the pixels in the selected window, the gradients are computed and histogram of the gradient is determined. The peak of the histogram is selected as the orientation of the feature point, Ѳ(x, y) [8]. The scale of the keypoint is used to select the Gaussian smoothed image, L, with the closest scale, so that all computations are performed in a scale-invariant manner. For each image sample, L(x, y), at this scale, the gradient magnitude, m(x, y), and orientation, Ѳ(x, y), is pre-computed using pixel differences according to eq (7) and eq (8),

$$m(x,y) = \sqrt{\left(L(x+1,y) - L(x-1,y)\right)^2 + \left(L(x,y+1) - L(x,y-1)\right)^2} \qquad (7)$$

$$\theta(x,y) = \tan^{-1}\frac{\left(L(x,y+1) - L(x,y-1)\right)}{\left(L(x+1,y) - L(x-1,y)\right)} \qquad (8)$$

The magnitude and direction calculations for the gradient are done for every pixel in a neighboring region around the keypoint in the Gaussian-blurred image L. In the case of multiple orientations being assigned, an additional keypoint is created having the same location and scale as the original keypoint for each additional orientation.

### 2.4 Image Normalization

Image scaling is considered one of the fatal geometric attacks the image may undergo. Scaling can be either symmetric or nonsymmetric in which the scaling factor in *x* direction is different from the scaling factor in *y* direction [22]. The normalized image is assumed to have a predefined area and a unit aspect ratio. The aspect ratio $\gamma$ of an image f(x,y) is defined using eq. 9 as,

$$\gamma = \frac{l_y}{l_x} \qquad (9)$$

Where $l_y$ and $l_x$ are the height and the width of f(x,y), respectively. Let f ((x/a),(y/b)) be the rescaled image with $\gamma$=1 and area $\propto = (a/x)(b/y)$, where a and b are the required scaling factors (eq. 10)

$$al_x = bl_y \qquad (10)$$

where a and b, are determined using eq (11) as,

$$a = \sqrt{\frac{\beta\gamma}{m_{o,o}}}, \quad b = \sqrt{\frac{\beta}{\gamma m_{o,o}}} \qquad (11)$$

Transforming the image into its standard form requires translating the origin of the image to its centroids. By using eq. 12 the coordinates (x,y) are changed into (x’,y’),

$$\acute{x} = \frac{x-\bar{x}}{a}, \quad \acute{y} = \frac{y-\bar{y}}{b} \qquad (12)$$

The image in the new coordinates system has aspect ratio $\gamma$=1 and area α.
Several remarks need to be mentioned at this point.

- The normalizing scheme does not need the original image for implementing the normalization at the decoder which adds a great advantage to the system.
- The normalized image suffers from smoothing effect which is a direct result of the interpolation that occurs in scaling and rotation correction.

With the scaling normalization, the aligned circular regions can be transformed to its compact size. Therefore, the selected circular regions are scaling invariant and are ready for watermark embedding. Based on the above analysis, the rotation and scaling invariant regions can be located in the image for watermark embedding. Because each region is a homogeneous area and its mean vector and covariance matrix have been calculated during the image segmentation, this information can help guide the watermark embedding process.

### 2.5 Watermark Embedding and Extraction

In the process of DWT watermark embedding, a bit stream of length L is transformed into a sequence W(l)….W(L) by replacing the 0 by –1 and W(K) €{-1,1} (k=1,...,L), used as the watermark. The original image is first decomposed using Haar filter into several bands using the discrete wavelet transformation with the pyramidal structure. The watermark is added to the largest coefficients in all bands of details which represent the high and middle frequencies of the image [24]. Let f(m,n) denote the DWT coefficients which are not located at the approximation band LL of the image. The embedding procedure is performed according to eq. 13.

$$f^1(m,n) = f(m,n) + \alpha f(m,n) W(k) \qquad (13)$$

Where, α is the strength of the watermark, controlling the level of the watermark W(1)….W(L). By this embedding, DWT coefficients at the lowest resolution which are located in the approximation band are not modified. The watermarked image is obtained by applying the Inverse Discrete Wavelet Transform (IDWT). Figure 2 shows the block diagram of the embedding method.

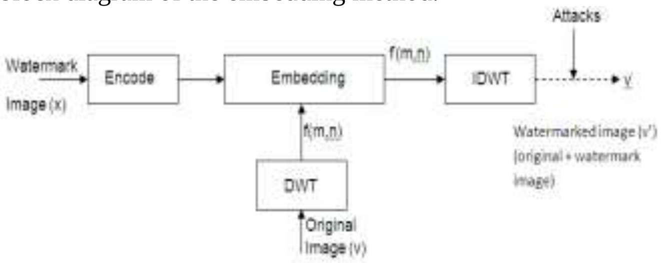


**Figure 2** Block Diagram of the embedding method

In the watermark extraction procedure, both the received image and the original image are decomposed into the two levels. It is assumed that the original image is used for extraction. The extraction procedure shown in Figure 3 is described by the formula given in eq. 14

$$W_r(k) = \left(f_r^1(m,n) - f(m,n)\right) / \left(\alpha f(m,n)\right) \qquad (14)$$

Where, fr′(m,n) are the DWT coefficients of the received image. Due to noise added to the image by attacks or transmission over the communication channel, the extracted sequence $W_r(1)$......$W_r(L)$ consists of positive and negative values. Hence, the extracted watermarks are modified according to eq. 15.

$$W_e(k) = \mathrm{sgn}(W_r(k)) \qquad (15)$$

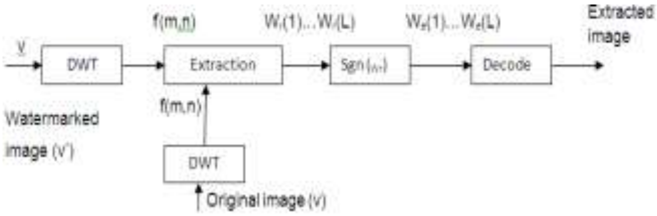


**Figure 3** Block Diagram of the extraction method

## 3. Optimization Techniques

The goal of applying optimization in watermarking is to resolve the conflicting requirements of different parameters and properties of digital images. The balance between the watermark robustness and transparency has been one of the defying task for watermarkers and as a result, there is an urgent requirement for using powerful computation and optimization techniques that guarantee the watermarking performance. Optimized watermarking methods are discussed in this section based on theoretical derivations of algorithms with the aid of evolutionary computing techniques like Genetic Algorithm (GA), Particle Swarm Optimization (PSO) and Hybrid Particle Swarm Optimization (HPSO).

### 3.1 Genetic Algorithm

Genetic Algorithm (GA) is a search technique for determining the global maximum/minimum solutions for problems in the area of evolutionary computation. Although the GA operation performs randomly, choosing candidates to avoid stranding on a local optimum solution, there is no guarantee that the global maximum/minimum will be found [25]. The probability and the possibility of obtaining the global optimal solution by using GA are based on the complexity of the problem. Inspite of these difficulties, GA have been successfully applied to obtain good solutions in several applications. Any optimization problem is modeled in GA by defining the chromosomal representation, fitness function, and application of the GA operators. The GA process begins with a few randomly selected genes as the first generation, called population. Each individual in the population corresponding to a solution in the problem is called chromosome, which consists of finite length strings. The objective of the problem, called fitness function, is used to evaluate the quality of each chromosome in the population. Chromosomes that possess good quality are said to be fit and they survive and form a new population of the next generation. The three GA operators, selection, crossover, and mutation, are applied to the chromosomes repeatedly to determine the best solution over successive generations [26].

In digital image watermarking, the population is initialized by choosing a set of random positions in the cover image and inserting the watermark image into the selected positions. The optimal solutions for digital watermarking using DWT are obtained based on two key factors: the DWT sub-band and the value of the watermark strength factor [11]. The GA algorithm searches its population for the best solution with all possible combinations of the DWT sub-bands and watermark amplification factors. The host image taken into consideration is decomposed into four sub-bands with different resolutions. The decomposition process can be performed at different DWT levels, first, second, third, or higher. The optimal subband is determined by GA as follows: The first level produces 4 sub-bands, the second level takes each sub-band of the first level and decomposes it further into four sub-bands resulting in 16 sub-bands. Similarly, the third DWT level decomposes each second level sub-band into 4 subbands, giving a total of 64 sub-bands. The genetic algorithm procedure will attempt to find the specific sub-band that will provide simultaneous perceptual transparency and robustness. Inorder to improve the robustness of the algorithm against attacks, the watermark strength or the amplification factor α should be optimized, but this factor varies on each sub-band. The fitness function is formed based on the parameters Peak Signal to Noise Ratio (PSNR) and the correlation factor ρ (α * NC) as shown is Eq. 16. Here, the correlation factor is the product of Normal Correlation (NC) and the watermark strength factor α. The fitness function increases proportionately with the PSNR value, but NC is the key factor contributing to the robustness and ultimately, the fitness value increases with the robustness measure. The correlation factor ρ has been multiplied by 100 since its normal values fall in the range 0 ~ 1, where as PSNR values may reach the value of 100.

$$Fitness\ Function = PSNR + 100 * \rho \qquad (16)$$

where, PSNR in decibels (dB) is computed as shown in eq. (17):

$$PSNR_{AB} = 10\log\left(\frac{MAX_i^2}{MSE}\right) = 20\log_{10}\frac{MAX_i}{\sqrt{MSE}} \qquad (17)$$

Here, MSE = the mean square error between the original image and the watermarked image

MAXi = the maximum pixel value of the image which is generally 255 in our experiment since pixels were represented using 8 bits per sample.

The fitness function is evaluated for all the individuals in the population and the best fit individual along with the corresponding fitness value are obtained. Genetic operators like crossover and mutation are performed on the selected parents to produce new offspring which are included in the population to form the next generation. The entire process is repeated for several generations until the best solutions are obtained. The correlation factor ρ measures the similarity between the original watermark and the watermark extracted from the attacked watermarked image (robustness). The correlation factor ρ is computed using eq. 18 shown:

$$\rho(W,\hat{W}) = \frac{\sum_{i=1}^{N} W_i,\hat{W}}{\sqrt{\sum_{i=1}^{N} W_i^{\,2}}\sqrt{\sum_{i=1}^{N} \hat{W}_i^{\,2}}} \qquad (18)$$

Where, N denotes the number of pixels in the watermark, w and wˆ represent the original and extracted watermarks respectively. The procedure for implementing digital image watermarking using GA is shown below:

```
Initialize watermark amplification factor α between 0 and 1, initialize the population
Generate the first generation of GA individuals based on the parameters specified by performing the watermark embedding procedure. A different watermarked image is generated for each individual.
While max iterations have not reached do
        Evaluate the perceptual transparency of each watermarked image by computing the corresponding PSNR value
        Apply a common attack on the watermarked image.
        Perform the watermark extraction procedure on each attacked watermark image.
        Evaluate robustness by computing the correlation between the original and extracted watermarks
        Evaluate the fitness function for the PSNR and ρ values
        Select the individuals with the best fitness values.
        Generate new population by performing the crossover and mutation functions on the selected individuals.
End While
```

The parameters like robustness, transparency, fitness value, CPU time and elapsed time are determined using genetic algorithm when the watermark images are attacked by filtering, scaling, rotation, and JPEG compression.

### 3.2 Particle Swarm Optimization

Particle Swarm Optimization (PSO) is a scheme for optimizing functions based on the allegory of social behavior of flocks of birds and schools of fish. It was first designed to simulate birds seeking food which is defined as a cornfield vector [27]. An individual bird in the flock would find a path for food through social cooperation with other birds around it. No bird is a leader, if one bird changes its style of flying all the other follow the same. In digital watermarking, for each time step a particle has to move to a new position, by adjusting its velocity. The movement is adjusted in the direction of its best current velocity or according to the direction of the neighborhood best. Having worked out a new velocity, its position is simply its old position plus the new velocity. PSO algorithm takes each particle in the swarm representing a solution to the problem and it is defined with its position and velocity. In D-dimensional search space, the position of the $i^{th}$ particle can be represented by a D-dimensional vector, present []= ($present_{i1}$,…, $present_{id}$, …, $present_{iD}$). The velocity of the particle $v_{[]}$ can be represented by another *D*-dimensional vector $V_{[]}$= ($V_{i1}$,…, $V_{id}$, …, $V_{iD}$). The best position visited by the $i^{th}$ particle is denoted as $gbest_i$=($gbest_{i1}$,…,$gbest_{id}$,…,$gbest_{iD}$), and $P_g$ as the index of the particle visited the best position in the swarm, then gbest becomes the best solution found so far, and the velocity of the particle and its new position will be determined by the following equations.

$$V[\,] = Co * V[\,] + C1 * rand(\,) * (pbest[\,] - present[\,]) + C2 * ran \qquad (19)$$

$$present[\,] = present[\,] + v[\,] \qquad (20)$$

The PSO algorithm for digital image watermarking tries to determine optimal Scale Factors (SFs). The transparency and robustness can be achieved by optimizing these scaling factors. While applying PSO to digital image watermarking, each string (combinations of 1s and 0s) in the swarm represents a possible solution to the problem with a set of SFs [28]. The scale factors for the initial swarm solutions are generated in a random manner. If the pixel values of the watermarked image are out of the desired range, then they are, rescaled based on the host image pixel values.

There are various signal processing and image processing operations for which the robustness has to be evaluated. The robustness and transparency of the proposed watermarking algorithm is evaluated in this paper using the attacks that are commonly employed in literature such as filtering, Gaussian noise, rotation, scaling and JPEG compression. In PSO optimization procedure, the attacking scheme refers to removing the extracted image after adding noise.

The watermarks are computed from the attacked watermarked images using the extraction procedure. The two dimensional correlation values are calculated between the original and watermarked images (*corrI*= *corr*(*I*, *IW*)) and between the original watermark and the extracted watermarks (*corrW*= *corr*(*W*,*W**)). The correlation values are then feed backed to PSO to evaluate the appropriateness of the SFs. The appropriateness of a solution is calculated depending on both the transparency (*corrI*) and the robustness (*corrW*) under noise attack at each iteration of optimization process as mentioned above. The objective function to be minimized is defined as:

$$f_i = \left[ \frac{1}{\frac{1}{t}\sum_{i=1}^{t} \max( corr_{Wi}(W, W_i^*)} - corr_I(I, I_W) \right]^{-1} \quad (21)$$

where, $corr_I$ and $corr_W$ are related to transparency and robustness measure, respectively; *fi* and *t* are the objective value of the $i^{th}$ solution and the number of attacking methods, respectively. Since each watermark pixel is embedded into each corresponding sub-band, maximum value of correlations (max(*corrW*)) is considered in the calculations. The processes explained above are repeated until a predefined stopping criterion is satisfied, for example maximum iteration number. The pseudocode of the PSO for DWM is shown below:

```
Initialize swarm size, acceleration constant and inertia weight of the swarm
Generate the initial swarm randomly
While stopping condition is false do
        Present the watermarked images into the swarm population
        Apply attacks on the watermarked images and extract the watermarks
        Compute the similarity between the watermark and the extracted ones
        Evaluate the objective function
        Feedback the appropriateness value to the PSO to get new values
End While
```

The swarm size, acceleration constant and inertia weight of the swarm are initialized. The watermarked images are produced using the solutions in the swarm by means of embedding process. The Normalised Correlation (NC) values are computed between the host image and each watermarked images. The attacks are applied upon the watermarked images one by one and the watermarks are extracted from the attacked images using the extraction procedure. The objective function is evaluated and the procedure is repeated until a constant number of generations are reached. In our experiment, by using Particle Swarm Optimization the fitness function is determined by evaluating the quality of each solution, so that individuals with high quality will survive to the next swarming process and the parameters like fitness value, CPU time and elapsed time are determined from the watermarked image.

## 4 Result Analysis

The experimental analysis and simulation determine the efficiency and capability of the work. In this section, the simulation results for various modules implemented namely, image segmentation, feature extraction, orientation assignment, image normalization, watermark embedding and extraction, optimization techniques like Genetic Algorithm, Particle Swarm Optimization and Hybrid Particle Swarm Optimization are explained to quantify the benefits of digital watermarking in image processing.

### 4.1 Image Segmentation

To evaluate the performance of proposed scheme, set of 50 texture images were chosen among which 512x512 images were taken as the original image and 256x256 images as the watermark image. The purpose of image segmentation is to partition an image into meaningful regions with respect to a particular application. The segmentation is based on measurements taken from the image and might be grey level, colour, texture, depth or motion. Using the EM segmentation, the image is segmented into number of homogeneous regions and the parameters are updated for each segmented regions.

The original image *house.tiff* is segmented into five different homogeneous regions, and each segmented region is represented based on colors. The variance of the regions are high if the region contains more dark regions. This implies that the darker sections contain more high frequency components. Figures 4 and 5 show the original and the segmented images of *house.tiff* using EM segmentation.

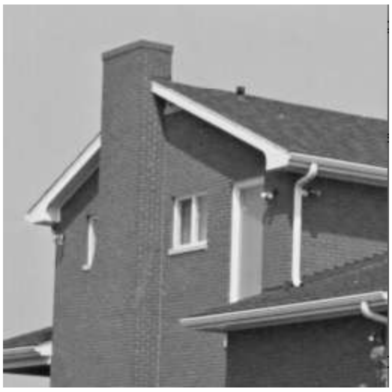


**Figure 4** Original Image

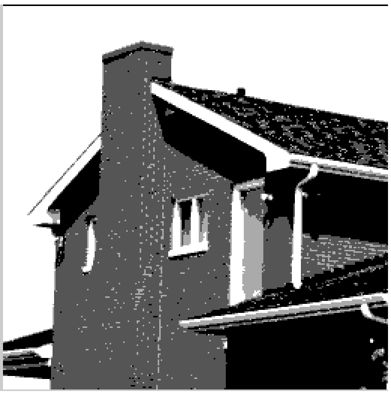


**Figure 5** Segmented Image

**Table 1** EM Image Segmentation

| Class | Mean | Variance |
|---|---|---|
| K=1 | 118.8056 | 0.003 |
| K=2 | 13.3041 | 0.0139 |
| | 144.8486 | 1.4098 |
| K=3 | 12.9443 | 0.0101 |
| | 123.1324 | 1.9331 |
| | 166.2306 | 0.1464 |
| K=4 | 12.8613 | 0.0094 |

| | | |
|---|---|---|
| | 108.3290 | 2.8983 |
| | 168.0908 | 0.1282 |
| | 132.8610 | 0.2781 |
| **K=5** | 12.7761 | 0.0089 |
| | 45.1680 | 0.4034 |
| | 127.3312 | 0.4092 |
| | 167.5977 | 0.1388 |
| | 188.2837 | 1.1109 |

Table 1 shows the value for mean and variance of the segmented image determined for each segment or class. The covariance can be set as the identity matrix and the mean is calculated by finding the average of different regions of the image. From the table, it is found that as the number of regions increase, the mean and variances of the segmented regions are also increased and this indicates the strength of the segmented region to embed the watermark image.

### 4.2 Feature Extraction

Difference of Gaussian (DoG) is a grayscale image enhancement algorithm to select the feature points from the segmented image. This algorithm computes the difference between one blurred version of the original grayscale image and another less blurred version of the original image. The difference of Gaussians is similar to a band-pass filter that discards all the coefficients but a handful of spatial frequencies are present in the original grayscale image. The parameters used in the DOG algorithm and their values are,

Smoothing parameter ($s_p$) – fixed
Radius (r) – 10
Sigma (σ) – [0-200]

**Table 2** Gaussian values for Original and Blurred image

| Sigma | G<br>(diff of original image) | G1<br>(diff of blurred image) | g<br>(Gaussian of original image) | g1<br>(Gaussian of blurred image) |
|---|---|---|---|---|
| 70 | 0.0057 | 0.0040 | 255 | 255 |
| 80 | 0.0050 | 0.5902 | 255 | 78 |
| 90 | 0.0044 | 0.0031 | 255 | 24 |
| 100 | 0.0040 | 0.0028 | 255 | 11 |
| 110 | 0.0036 | 0.0026 | 97 | 6 |

The scale space extreme for the original and blurred image to extract the feature points is given by taking difference of Gaussian values as shown in Table 2. The local extrema in the DoG are found and by removing those with strong edge responses, the final results are selected as the feature points. DoG algorithm is chosen in our experiment to find the extracted points since, in each segmented region, one feature point is selected and the circular region centered at the selected feature point with radius will be used for the watermark embedding and detection. After the reference feature points are selected, the rotation and scaling invariant properties are assigned to the circular regions centered at the selected feature points.

### 4.3 Orientation Assignment

The circular regions are chosen and are made rotation invariant using orientation assignment. In order to achieve orientation invariance, the coordinates of the descriptor and the gradient orientations are rotated relative to the keypoint orientation. By finding the image gradients as in Figure 9, the key points in the image are extracted and modified to reduce the illumination change.

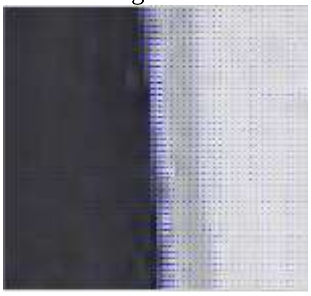


**Figure 6** Gradient of the image

### 4.4 Image Normalization

A technique for normalizing an image against geometric manipulation is implemented and the purpose is to obtain scaling, rotation invariance for the image during watermark embedding and extraction phases. Scaling normalization is employed to acquire the scaling invariance for the circular region. It transforms the image into its standard form by translating the origin of the image to its centroid(x,y). The normalized form of the *house.tiff* image is shown in Figure 7. In this figure, the normalization effect of the image is evaluated using local mean and standard deviation estimated by Gaussian kernel with sigma = 4.

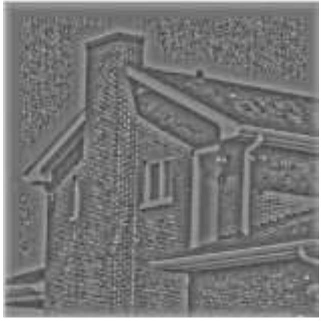


**Figure 7** Normalised image

### 4.5 Watermark Embedding and Extraction

For watermark embedding and extraction using DWT, original image ‘*house.tiff*’ with size 256x256 is taken as a cover image and the watermark image ‘*best.bmp*’ with size 60x24 is taken as the image to hide as shown in the Figure 8 and Figure 9 respectively. The image is decomposed into sub-image of different spatial domain and independent frequency distinct, and the resultant image will be of low pass-low pass (LL) images in the upper left corner, the low pass-high pass (LH) images on the diagonals and the high pass-high pass (HH) in the lower right corner. The process will continue to run the same wavelet transform on the low pass-low pass version of the image to get sub images.

The original image is first decomposed into several bands using the discrete wavelet transformation with the pyramidal structure. The sub-band LL represents the coarse-scale DWT coefficients while the sub-bands LH, HL and HH represent the fine-scale of DWT coefficients. If the information of low-frequency distinct is DWT transformed, the sub-level frequency distinct information will be obtained. The watermark is added to the largest coefficients in all bands of details which represent the high and middle frequencies of the image. The watermarked image by Discrete wavelet transform is shown in the Figure 10.

The parameters used for embedding the watermark into the cover image using DWT are,
Part – Decomposition
Data - input image
Mode - type of image (tiff,bmp.jpg,jpeg,png)
Level - 3

In the watermark extraction procedure both the received image and the original image are decomposed into the two levels. In the first level the image is filtered by lowpass, highpass filters and the lowpass region is further filtered by both (LP and HP) filters in second level to obtain the LL plane.

**Figure 10** Original Image

**Figure 11** Image to Hide

**Figure 12** Watermarked Image

### *4.6 Optimization in Watermarking*

This section illustrates the implementation and simulated results of the computational intelligence techniques applied for digital image watermarking.

*Genetic Algorithm*

The GA training procedure was executed with an initial population size of 120. The watermark bits of the '*best.bmp*' are inserted into the cover image '*house.tiff*' and this forms the initial population. The fitness function is evaluated based on the PSNR and the CPU time, and the elapsed time is determined. The crossover rate is 0.9 with single point crossover and the mutation rate is 0.02 with the flip bit type of mutation. The GA procedure is repeated until the maximum number of generations, 100 are reached. The watermark amplification factor is initially set as 0.5, as the generations progress, the value varies with respect to the sub-bands. The same GA procedure was tested on 6 different images chosen in a random manner from a set of 50 images with the same watermark image. The parameters used for digital watermarking using GA are shown below:

| | |
|---|---|
| Population size | : 120 (multiples of 4) |
| Crossover probability | : 0.9 |
| Mutation probability | : 0.02 |
| Type of crossover | : Single point crossover |
| Type of mutation | : Flip bit |
| Number of generations | : 100 |

**Table 3** Fitness Value and Best Point of the Extracted Watermark Image using GA

| Watermark Images | Fitness value | CPU time (sec) | Elapsed time (sec) |
|---|---|---|---|
| **House.tiff** | 1.0e+006 | 14.8760 | 17.5560 |
| **Rice.png** | 1.0e+007 | 13.0313 | 16.1316 |
| **Fingerprint.bmp** | 1.0e+007 | 19.4563 | 21.6565 |
| **Stone.jpg** | 1.06e+006 | 11.5713 | 14.5175 |
| **Goldhill.bmp** | 1.02e+007 | 15.4973 | 13.5783 |
| **Zoneplate.bmp** | 1.01e+007 | 14.1345 | 19.3285 |
| **Lena.tif** | 1.0e+006 | 12.1874 | 16.7832 |

The fitness value, CPU time and the elapsed time are computed for a set of images as shown in Table 3. The fitness value increases proportionately with the PSNR value and the watermark amplification factor α. The GA procedure was also tested with filtering attacks, JPEG compression, rotation and scaling attacks. The results are discussed in the following sections.

*Particle Swarm Optimization*

The initial particles are formed by embedding the chosen watermark image bits into the cover image, which comprises the initial population. The particle size in our simulation is chosen as 36, if the particle size is larger, more points can be searched in the search space to determine the global optimal solution, but this increases the number of iterations and hence the computational time. Based on several simulation results, we choose the cognitive factor $c_1$=1.8 and social coefficient factor $c_2$=2, and inertia weight *w*=0.3. The algorithm terminates after 200 iterations in accordance with various experimental observations. The parameters and their values of the PSO algorithm for watermarking optimization are configured as follows.

| | |
|---|---|
| Cognitive factor, $c_1$ | : 1.8 |
| Social coefficient $c_2$ | : 2 |
| Inertia weight, w | : 0.3 |
| Particle size | : 36 |
| Number of generations | : 200 |

**Table 4** Fitness value and best points of the extracted watermark image using PSO

| Images | Fitness value | CPU time (sec) | Elapsed time (sec) |
|---|---|---|---|
| **House.tiff** | 1.05e+006 | 12.543 | 14.454 |
| **Rice.png** | 1.21e+007 | 13.0313 | 15.4042 |
| **Fingerprint.bmp** | 1.0e+007 | 18.4563 | 21.3362 |
| **Stone.jpg** | 1.06e+005 | 14.5713 | 17.2371 |
| **Goldhill.bmp** | 1.04e+006 | 11.5213 | 15.1713 |
| **Zoneplate.bmp** | 1.02e+007 | 19.3463 | 21.3243 |
| **Lena.tif** | 1.03e+007 | 02.1138 | 03.1271 |

The fitness value and the computational time for the chosen set of images in PSO based digital watermarking are shown in Table 4.

### *4.7 Attacks*

To evaluate the performance of the optimization techniques, several experiments were conducted using the set of host images and a single watermark image. The common attacks employed to the watermarked image here are filtering, addition of Gaussian noise, rotation, scaling and JPEG compression. The two dimensional correlation values are calculated between the original and watermarked images (*Transparency measure*) and between the original watermark and the extracted watermarks (*Robustness measure*). The GA, and PSO procedures were repeated by applying these attacks and the robustness measure was evaluated.

*Filtering attacks*

The watermarked image under consideration *house.tiff* is subject to several types of filtering like average filtering, Gaussian filter, median filtering, and Wiener filtering which are regarded as attacks. The mask for the filter is usually a window which can take various sizes. Average filtering removes the high frequency components present in the image acting like a low pass filter. The average filter with a 5 x 5 mask was applied to the watermark image during the optimization process of GA, and PSO to evaluate the robustness measure. While using Gaussian filter attack, the mean was set to '0' and the variance set to '1', with a window size 3 x 3. The median filtering was applied four times on the watermarked image with a mask size of 3 x 3, and this seems to preserve the edges while recovering the watermark. Wiener or adaptive noise removal filtering is an adaptive process tailoring itself to the local image variance which is inversely proportional to the smoothing process. This works best in removing the Gaussian white noise and in our experiment the mask size was set to 3 x 3. Table 5 shows the attacked watermarked image with different types of filtering attacks. The correlation factor is evaluated based on the similarity between the original watermark and the attacked watermark for all the filtering techniques and tabulated. The results show that the PSO watermark optimization technique has the best similarity of extracted watermarks among the compared approach such as GA.

*Additive Noise*

A Gaussian noise was added to the watermarked image with zero mean and different variance σ, indicating the percentage of gray levels added into the image. The robustness measure is computed by varying σ between [0.001, 1] for GA, and PSO algorithms as indicated in Table 6.

**Table 5** Computed robustness for Filtering Attack

| Optimization Technique | Average Filtering | Gaussian Filtering | Median Filtering | Adaptive Noise Removal Filtering |
|---|---|---|---|---|
| Watermarked Image | | | | |
| Robustness measure using GA | 0.972 | 0.989 | 0.995 | 0.992 |
| Robustness measure using PSO | 0.921 | 0.979 | 0.994 | 0.984 |

**Table 6** Evaluated Results for Gaussian noise Attack

| Gaussian Noise | σ=0.001 | σ=0.01 | σ=0.1 | σ=0.5 | σ=1 |
|---|---|---|---|---|---|
| Watermarked Image | | | | | |
| Robustness measure using GA | 0.875 | 0.871 | 0.864 | 0.821 | 0.81 |
| Robustness measure using PSO | 0.887 | 0.882 | 0.879 | 0.862 | 0.842 |

**Table 7** JPEG Compression Attack and robustness computation

| JPEG Compression Quality Factor | 10% | 20% | 36% | 50% | 70% | 85% | 95% |
|---|---|---|---|---|---|---|---|
| Watermarked Image | | | | | | | |
| Robustness measure using GA | 0.823 | 0.845 | 0.878 | 0.889 | 0.894 | 0.910 | 0.936 |
| Robustness measure using PSO | 0.872 | 0.884 | 0.891 | 0.917 | 0.930 | 0.952 | 0.972 |

**Table 8** Robustness Results for Rotation Attack

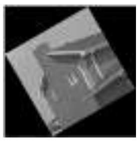
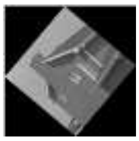
| Degree | 5º | 10º | 20º | 30º | 40º | 50º |
|---|---|---|---|---|---|---|
| Watermarked Image | | | |  | |  |
| Robustness measure using GA | 0.88 | 0.82 | 0.75 | 0.658 | 0.595 | 0.545 |
| Robustness measure using PSO | 0.93 | 0.912 | 0.9 | 0.856 | 0.823 | 0.78 |

**Table 9** Experimental Results for Scaling Attack

| Scale Factor | 0.5 | 0.75 | 0.9 | 1 | 1.2 | 1.5 | 2 |
|---|---|---|---|---|---|---|---|
| Watermarked Image | | | | | | | |
| Robustness measure using GA | 0.786 | 0.825 | 0.932 | 0.987 | 0.86 | 0.789 | 0.753 |
| Robustness measure using PSO | 0.833 | 0.947 | 0.984 | 0.996 | 0.953 | 0.811 | 0.712 |

*JPEG Compression*

JPEG is one of the most widely used lossy compression algorithms, and any watermarking technique should be resilient to some degree of JPEG compression attacks. For images that are published on the internet, robustness of watermarks against the JPEG compression standard is particularly important. In general, such lossy compression algorithms discard the redundant and perceptual insignificant information during the coding process, but watermark embedding schemes add invisible information to the image. JPEG compression calculates the visual components based on the relationship with the neighboring pixels in the image. The specific positions to embed the watermark are derived from these visual components which are proportion to the quality levels of the JPEG compression. In practice, it is difficult to choose the minimal quality factor (QF) of JPEG for compression. Low values of quality factor indicate high compression ratio and vice versa. The Quality Factors (QF) were set from 10% to 95% for simulation, and the results are shown in Table 7. The watermark was detected well even after the image was compressed using a quality factor of 10%. The correlation factor seems to be high for the quality factor of 95%, indicating that the similarity values prove best match between the original watermark and the extracted watermark. The robustness measure is 0.989 when the PSO algorithm is applied to optimize the watermarking process and this shows that the proposed hybrid watermarking technique is superior to GA whose robustness measure is 0.936.

*Rotation*

Geometric attacks usually make the watermark detector loose the synchronization information and one of the major attacks among this group is rotation. The watermarked image is rotated by an angle in the counterclockwise direction before extracting the watermark. The watermarked image is rotated by angles 5º, 10º, 20º, 30º, 40º and 50º to the right and then rotated back to their original position using bilinear interpolation. While rotating, the black pixels left after rotation in the corners have been included to maintain the image size and the shape. For larger angles, more black pixels are padded and hence the correlation factor or the robustness measure tends to decrease. For each degree of rotation, the correlation factor was measured between the original watermark and the attacked watermark to determine the degree of similarity between them. Table 8 shows the watermarked image rotated for a set of angles and the corresponding robustness measure while applying GA, and the PSO algorithm. It can be inferred that the watermarked image is robust against rotation attack when PSO algorithm is applied since the correlation factor is 0.93 denoting more similarity features between the watermark image and the attacked watermark image.

*Scaling*

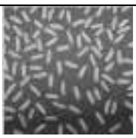
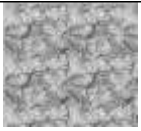
Scaling is generally considered more challenging than other attacks due to the fact that changing the image size or its orientation even by slight amount, could dramatically reduce the receiver ability to retrieve the watermark. The scaling factors are selected such that the robustness, invisibility and quality of the extracted watermark is maintained, usually higher in the low frequency band and lower in the high frequency band. In our experiment, the watermarked image was scaled by using different scale factors within the range [0.5, 2] as shown in Table 9. For a reasonable range of the scale factor, the applied optimization techniques are robust to the scaling concept.

The experiments were conducted on a set of images and the transparency measure is computed for the images house.tiff, rice.png, fingerprint.bmp, goldhill.bmp, stone.jpg and lena.tif and the results are tabulated. Table 13 shows the computed values of perceptual transparency when the optimization techniques are applied to the corresponding set of images. The invisibility and fidelity are comparatively high for the PSO technique (transparency=0.9994) for *house.tiff*, which is closer to one when compared against GA.

## 5. Conclusion

Digital image watermarking algorithms based on the discrete wavelet transform (DWT) have been widely recognized to be more prevalent than the existing steganographic techniques. In this paper, several pre-watermarking stages were included to enhance the quality and to maintain the invariance property of the image to be watermarked. The watermarked image was optimized using computational intelligence techniques like GA, and PSO and the performance was compared. Imperceptibility of images and robustness of abstracting digital watermark are an important criterion of judging digital watermarking algorithm. Inorder to prove this, several attacks were imposed on the watermarked image and optimized using the evolutionary algorithms. The PSO algorithm seemed to yield a better watermark, invisible to the human eye, when filtering attacks were applied. The fact that geometrical attacks like rotation and scaling are more stronger than filtering attacks is proved from the results obtained in Table 11 and 12. In JPEG compression attacks, the extracted watermarks are fully recognizable with PSO with quality factors between the ranges 10% and 95%. The experiments and results show that the PSO is not only robust to attacks, but also ensures the imperceptibility and fidelity of the watermark embedded image. The experiments can still further be expanded with additional attacks like stirmark, image enhancement, JPEG2000, translation and also a combination of attacks can be imposed to understand the performance of the optimization techniques in a better perspective.

**Table 13** Comparison of optimization techniques based on transparency measure

| Optimization Technique | Transparency Measure using GA | Transparency Measure using PSO |
|---|---|---|
| **House.tiff** | 0.9989 | 0.9994 |
|  **Rice.png** | 0.9988 | 0.9990 |
| **Fingerprint.bmp** | 0.9986 | 0.9990 |
| **Goldhill.bmp** | 0.9987 | 0.9989 |
|  **Stone.jpg** | 0.9990 | 0.9992 |
| **Lena.tif** | 0.9989 | 0.9993 |

## REFERENCES

[1] H, Vallabha V. "Multiresolution Watermark Based on Wavelet Transform for Digital images." Cranes Software International Ltd, 2003.

[2] Hernández., Fernando Pérez-González and Juan R. "A tutorial on digital watermarking." Proc. of the 33rd IEEE Annual Carnahan Conference on Security Technology, Spain, October 1999.

[3] Serkan EMEK, Melih PAZARCI. "A Cascade DWT-DCT Based Digital Watermarking Scheme" 13th European Signal Processing Conference, Turkey, 2005.

[4] Yi-leh Wu, Divyakant Agrawal,Amr El Abbadi. "A comparison of DFT and DWT based similarity search in time-series databases." Proc. Ninth international conference on Information and knowledge management, United States of America, 2000: pp. 488 - 495.

[5] X.H. Shia, Y.C. Lianga, b,H.P. Leeb, C. Lub and L.M. Wanga. "An improved GA and a novel PSO-GA-based hybrid algorithm." Information Processing, Elsevier, Volume 93, Issue 5, 2005: 255-261.

[6] W. T. Li, X. W. Shi, and L. Xu. "Improved GA And PSO Culled Hybrid Algorithm For Antenna Array Pattern Synthesis" Progress In Electromagnetics Research,, 2008: 461-476.

[7] J.-S. Pan, H.-C. Huang, L.C. Jain. Intelligent Watermarking Techniques (Innovative Intelligence). World Scientific Press , 2004.

[8] Dong Zheng, Sha Wang, and Jiying Zhao*," RST Invariant Image Watermarking Algorithm With Mathematical Modeling and Analysis of the Watermarking Processes",* IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 18, NO. 5, MAY 2009, 1055-1068

[9] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," Int. *J.* Comput. Vis., vol. 60, no. 2, pp. 91–110, 2004.

[10] C.-S. Shieh, H.-C. Huang, F.-H. Wang, and J.-S. Pan, "Genetic watermarking based on transform domain techniques," Pattern Recognition, vol. 37, no. 3, pp. 555-565, 2004.

[11] Abu-Errub, Ali Al-Haj and Aymen. "Performance Optimization of Discrete Wavelets Transform Based Image Watermarking Using Genetic Algorithms." Journal of Computer Science, 2008: 834-841.

[12] Zhicheng Wei, Hao Li, Jufeng Dai" Image Watermarking Based On Genetic Algorithm" School of Electronic Information Engineering, Tianjin University, Tianjin 300072, China, Department of Network Engineering, Hubei Normal University, Shijiazhuang 050016, China.

[13] Cong Jin: Affine invariant watermarking algorithm using feature matching. Digital Signal Processing 16(3): 247-254 (2006)

[14] G. Boato, V. Conotter and F. G. B. De Natale "GA-based Robustness Evaluation Method for Digital Image Watermarking" Proc. of IWDW 2007, Guangzhou, December 2007.

[15] C.-S. Shieh, H.-C. Huang, F.-H. Wang, and J.-S. Pan, "Genetic watermarking based on transform domain techniques," Pattern Recognition, vol. 37, no. 3, pp. 555-565, 2004.

[16] Zne-Jung Lee a,*, Shih-Wei Lin a, Shun-Feng Su b, Chun-Yen Lin bA hybrid watermarking technique applied to digital images, Applied Soft Computing Elsevier, vol 8, pp.798-808, 2008.

[17] Ziqiang Wang, Xia Sun and Dexian Zhang, “Novel Watermarking Scheme based on PSO Algorithm”, Bio-Inspired Computational Intelligence and Applications, Lecture Notes in Computer Science, 2007, Volume 4688, pp.307-314, 2007.

[18] Belongie, S. Carson, C. Greenspan, H. Malik, J. "Color- and texture-based image segmentation using EM and its application to content-based image retrieval." IEEE Proc. Computer Vision, 1998: 675-682.

[19] Batzoglou, Chuong B Do and Serafim. "What is the expectation maximization algorithm?" nature biotechnology volume 26 number 8, 2008: 897-899.

[20] Andrzejewski, David. Expectation Maximization. Machine Learning Notes, University of Wisconsin-Madison, 2010.

[21] Qiu Chen Kotani, K. Feifei Lee Ohmi, T. Scale-Invariant Feature Extraction by VQ-Based Local Image Descriptor. 1217-1222, Vienna : IEEE Conf. Proc. Computational Intelligence for Modelling Control & Automation, 2008.

[22] Alghoniemy, M. Tewfik, A.H. "Geometric distortion correction through image normalization." IEEE Int. Conf. Multimedia and Expo, 2000: 1291-1294.

[23] Wavelets and Subband coding. USA: Prentice Hall, 1995.

[24] Nataša Terzija, Markus Repges, Kerstin Luck, Walter Geisselhardt. "Digital Image Watermarking Using Discrete Wavelet Transform: Performance Comparison of Error Correction Codes"

[25] M. Ketcham, and S. Vongpradhip. "Intelligent Audio Watermarking using Genetic Algorithm in DWT Domain." World Academy of Science, Engineering and Technology, 2007: 336-341.

[26] Goldberg, David E. Genetic Algorithms in Search, Optimization, and Machine Learning. Addison Wesley, 1989.

[27] Angeline, P. "Evolutionary Optimization versus Particle Swarm Optimization: Philosophy and Performance Difference." The 7th Annual Conference on Evolutionary Programming, USA, 1998.

[28] Aslantas, V. Dogan, A.L. Ozturk, S. "DWT-SVD based image watermarking using Particle Swarm Optimizer." IEEE int. COnf. Multimedia and Expo, 2008: 241-244.

**Surekha P** is currently a research scholar with the Electrical and Electronics Engineering department at PSG College of Technology, Coimbatore, India. She received her B.E. Degree in Electrical and Electronics Engineering during 2001 from PARK College of Engineering and Technology, Coimbatore, Tamil Nadu, and Masters Degree in Control Systems during 2006 from PSG College of Technology, Coimbatore, Tamil Nadu.

She has published the following books: LabVIEW based Advanced Instrumentation Systems (Germany, Springer-Verlag, 2007), Evolutionary Intelligence (Germany, Springer-Verlag, 2008) and Computational Intelligence (Taylor and Francis, CRC Press, 2010). Her current research work includes Computational Intelligence Methodologies in various engineering applications. Her major interests include Robotics, Virtual Instrumentation, Mobile Communication, and Computational Intelligence.

**Dr. S Sumathi** is an Asst. Professor in the Electrical and Electronics Engineering department at PSG College of Technology, Coimbatore, India. She has completed B.E. Degree in Electronics and Communication Engineering at and a Masters Degree in Applied Electronics at Government College of Technology, Coimbatore, Tamil Nadu. The Author obtained her Ph.D. Degree in the area of Data Mining.

She has published a large number of articles in International conferences and journals. She has also published books on Neural Networks, Fuzzy Systems, Data Mining and its Applications, LabVIEW, Evolutionary Intelligence, and Computational Intelligence. She has reviewed papers in National/International Journals and Conferences. Her Research interests include Neural Networks, Fuzzy Systems and Genetic Algorithms, Pattern Recognition and Classification, Data Warehousing and Data Mining, Operating Systems and Parallel Computing.

# Steganography and Image Enhancement- A Proposed Method

Nitika Kapoor[1], Aashima[2] and Navneet Malik[3]

[1,2]Lovely Professional University, Phagwara,
[2]Rayat & Bahra Institute of Engineering & Bio-Technology, Sahauran
er.nitikakapoor@gmail.com, er.aashima@yahoo.co.in

***Abstract*:** In this paper we have proposed a method to combine the features of image enhancement and Steganography. Various still images will be used on which the tests will be implemented.

***Keywords*:** Image enhancement, Steganography

## 1. Introduction

Removing and reducing impulse noise is very active research area in image processing. Present day applications require various kinds of images and pictures as sources of information for interpretation and analysis. Whenever an image is converted from one form to another, some form of degradation occurs at the output. The output image has to undergo a process called image enhancement. An effective method for image enhancement was presented by Russo, which was controlled by tuning of one parameter.

Digital communication has become an essential part of infrastructure nowadays, a lot of applications are Internet-based and in some cases it is desired that the communication be made secret. Two techniques are available to achieve this goal: one is cryptography, where the sender uses an encryption key to scramble the message; this scrambled message is transmitted through the insecure public channel, and the reconstruction of the original, unencrypted message is possible only if the receiver has the appropriate decryption key. The second method is steganography, where the secret message is embedded in another message. Using this technology even the fact that a secret is being transmitted has to be secret.

Our method is to combine these two techniques.

## 2. Proposed Work

Steganography and Enhancement are the two broad categories in the field of image processing. We are tried to combine these two fields. The method is discussed here.

### 2.1 Noisy Image

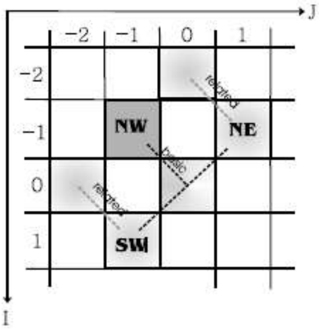
The method we are going to develop will be for the noisy image. We assume that the image contain the salt and pepper noise.

### Removal of Impulse noise

We start from a gray scale image in order to better explain how the new algorithm is constructed. Let the grayscale image be represented by a matrix F of size N1 ×N2, F ={F(i, j)€ {0, . . . ,255}, i = 1, 2, . . . ,N1, j = 1, 2, . . . ,N2}. Our construction starts with the introduction of the similarity function μ: [0 ;∞) →R. We will need the following assumptions for μ:

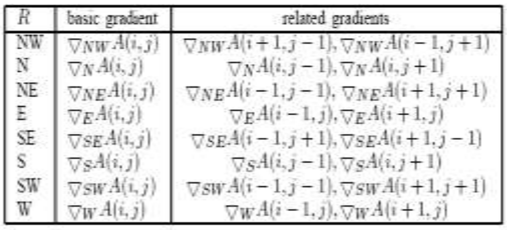
(1) μ is decreasing in [0;∞),
(2) μ is convex in [0;∞),
(3) μ(0) = 1, μ(∞) = 0.

In the construction of filter, the central pixel in the window W is replaced by that one, which maximizes the sum of similarities between all its neighbors. Basic assumption is that a new pixel must be taken from the window W.



For each pixel (i, j) of the image (that isn’t a border pixel) use a 3 X 3 neighborhood window. Each neighbor with respect to (i, j) corresponds to one direction {NW = North West, N = North, NE = North East, W = West, E = East, SW = South West, S = South, SE= South East}. Each such direction with respect to (i, j) can also be linked to a certain position.

INVOLVED GRADIENT VALUES TO CALCULATE THE FUZZY GRADIENT.



| $R$ | basic gradient | related gradients |
|---|---|---|
| NW | $\nabla_{NW}A(i,j)$ | $\nabla_{NW}A(i+1,j-1), \nabla_{NW}A(i-1,j+1)$ |
| N | $\nabla_{N}A(i,j)$ | $\nabla_{N}A(i,j-1), \nabla_{N}A(i,j+1)$ |
| NE | $\nabla_{NE}A(i,j)$ | $\nabla_{NE}A(i-1,j-1), \nabla_{NE}A(i+1,j+1)$ |
| E | $\nabla_{E}A(i,j)$ | $\nabla_{E}A(i-1,j), \nabla_{E}A(i+1,j)$ |
| SE | $\nabla_{SE}A(i,j)$ | $\nabla_{SE}A(i-1,j+1), \nabla_{SE}A(i+1,j-1)$ |
| S | $\nabla_{S}A(i,j)$ | $\nabla_{S}A(i,j-1), \nabla_{S}A(i,j+1)$ |
| SW | $\nabla_{SW}A(i,j)$ | $\nabla_{SW}A(i-1,j-1), \nabla_{SW}A(i+1,j+1)$ |
| W | $\nabla_{W}A(i,j)$ | $\nabla_{W}A(i-1,j), \nabla_{W}A(i+1,j)$ |

Each direction R corresponds to central position. Column 2 gives the basic gradient for each direction; column 3 gives the two related gradients. The fuzzy gradient value for direction R is calculated by following fuzzy rule:

IF $|\nabla_R\ A(i,j)|$ is large AND $|\nabla'_R\ A(i,j)|$ is small

OR

$|\nabla_R\ A(i,j)|$ is large AND $|\nabla''_R\ A(i,j)|$ is small

OR

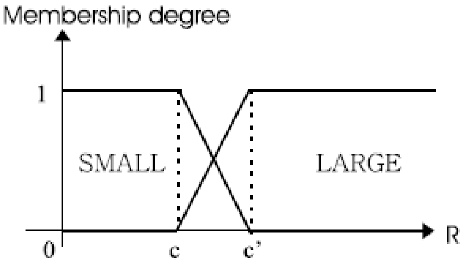
$\nabla_R\ A(i,j)$ is big positive AND $\left(\nabla'_R\ A(i,j)\right.$ AND $\left.\nabla''_R\ A(i,j)\right)$ are big negative

OR

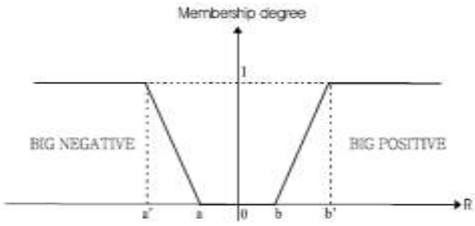
$\nabla_R\ A(i,j)$ is big negative AND $\left(\nabla'_R\ A(i,j)\right.$ AND $\left.\nabla''_R\ A(i,j)\right)$ are big positive

THEN $\nabla^F_R\ A(i,j)$ is large

The membership function used are LARGE (for the fuzzy set large), SMALL (for the fuzzy set small), BIG POSITIVE (for the fuzzy set big positive) and BIG NEGATIVE (for the fuzzy set big negative).



The above graph shows the membership function for fuzzy set SMALL and LARGE.



The above graph shows the membership function for fuzzy set BIG NEGATIVE and BIG POSITIVE.

The pixels of the image are arranged in these membership functions. The noisy pixels are then sort out and form the member of the function more or less impulse noise. The noisy pixel values are then changed according to the following formula:

$$F(i,j) = \frac{\sum_{h=-1}^{1}\sum_{l=-1}^{1}[1-\mu(A(i+h,j+l))]A(i+h,j+l)}{\sum_{h=-1}^{1}\sum_{l=-1}^{1}1-\mu(A(i+h,j+l))}$$

## Improving contrast of the image

For improving the contrast of the image following steps are done: setting the shape of membership function (regarding to the actual image) setting the value of fuzzifier Beta calculation of membership values modification of the membership values by linguistic hedge generation of new gray-levels.

Using the notation of fuzzy sets, we can write,

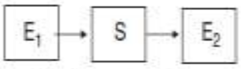
$$X = \begin{pmatrix} \mu_{11}/x_{11} & \mu_{12}/x_{12} & \cdots & \mu_{1M}/x_{1M} \\ \mu_{21}/x_{21} & \mu_{22}/x_{22} & \cdots & \mu_{2M}/x_{2M} \\ \vdots & \vdots & & \vdots \\ \mu_{N1}/x_{N1} & \mu_{N2}/x_{N2} & \cdots & \mu_{NM}/x_{NM} \end{pmatrix}$$

Where 0 ≤ μmn ≤ 1,m = 1, 2 . . . M,n = 1, 2 . . . N.

Contrast within an image is measure of difference between the gray-levels in an image. The greater the contrast, the greater is the distinction between gray-levels in the image. Images of high contrast have either all black or all white regions; there is very little similar gray-levels in the image, and very few black or white regions. High-contrast images can be thought of as crisp, and low contrast ones as completely fuzzy. Images with good gradation of grays between black and white are usually the best images for purposes of recognition by humans.

The object of contrast enhancement is to process a given image so that the result is more suitable than the original for a specific application in pattern recognition. As with all image-processing techniques we have to be especially careful that the processed image is not distinctly different from the original image, making the identification process worthless. The technique used here makes use of modifications to brightness membership value in stretching or contracting the contrast of an image.

Many contrast enhancement methods work as shown in the figure below, where the procedure involves primary enhancement of the image, denoted with an E1 in the figure, followed by a smoothing algorithm, denoted by an S, and a subsequent final enhancement, step E2.



Method of contrast enhancement

The function of the smoothing operation of this method is to blur (make fuzzier) the image and this increased blurriness then requires the use of final enhancement step E2. Generally smoothing algorithms distribute a portion of the intensity of one pixel in the image to adjacent pixels. This distribution is greatest for pixels nearest to the pixels being smoothed, and it decreases for pixels farther from the pixel being smoothed.

## 2.2 Steganography

Steganography embeds a secret message in a cover message, this process is usually parameterized by a stego-key, and the detection or reading of embedded information is possible only having this key.

We will implement steganography on noisy and low contrast images. We have opted two methods for the same.

**LSB**

LSB is the lowest bit in a series of numbers in binary. e.g. in the binary number: 10110001, the least significant bit is far right 1.

The LSB based Steganography is one of the steganography methods, used to embed the secret data in to the least significant bits of the pixel values in a cover image. e.g. 240 can be hidden in the first eight bytes of three pixels in a 24 bit image.

**DCT**

DCT coefficients are used for JPEG compression. It separates the image into parts of differing importance. It transforms a signal or image from the spatial domain to the frequency domain. It can separate the image into high, middle and low frequency components.

**METHOD**

Firstly, we will take the noisy and low contrast images. At the sender end we will embed an image in the said image.

Then, the noise will be removed from the image and the embedded image is extracted.

The stego image is extracted using two algorithms LSB and DCT.

## References

[1] József LENTI, STEGANOGRAPHIC METHODS, PERIODICA POLYTECHNICA SER. EL. ENG. VOL. 44, NO. 3–4, PP. 249–258 (2000)

[2] K.T.Talele, Dr.S.T.Gandhe, Dr.A.G.Keskar, *International Journal of Computer and Network Security,Vol. 2, No. 4, April 2010*

# On Decomposition of Bitopological (1,2)*-*A*- Continuity

[1]O. Ravi, [2]K. Mahaboob Hassain Sherieff and [3]M. Krishnamoorthy

1. Department of Mathematics, P.M.Thevar College, Usilampatti, Madurai – Dt., Tamil Nadu, India. E-mail: siingam@yahoo.com
2. Department of Mathematics, S.L.S.M.A.V.M.M.A.V College, Kallampatti, Madurai – Dt., Tamil Nadu, India. E-mail: rosesheri14@yahoo.com
3. Department of Mathematics, R V S Engineering College, Dindigul, Tamil Nadu, India.

***Abstract:*** *- The aim of this paper is to give decompositions of continuity, namely (1,2)*-A-continuity by providing the concepts of (1,2)*-semi-continuity, (1,2)*-C-continuity, (1,2)*-β-continuity and (1,2)*-LC-continuity.*

***2000 Mathematics Subject Classification. 54E55.***

***Keywords and Phrases :*** *(1,2)*-A-set, (1,2)*-C-set, (1,2)*-A-continuity, (1,2)*-C-continuity,*

*(1,2)*-β-open set, (1,2)*-semi-open set*

## 1. Introduction

To give a decomposition of continuity, Tong[13] introduced the notions of A-set and A-continuous mappings and proved that a map f : X→Y is continuous if and only if it is both α-continuous and A-continuous. Again, Tong[14] introduced the notions of B-sets and B-continuous mappings, and together with the notion of precontinuity he proved another decomposition of continuity i.e., A mapping f : X→Y is continuous if and only if it is precontinuous and B-continuous. Ganster and Reilly[5] established a decomposition of A-continuity i.e., a mapping f : X→Y is A-continuous if and only if it is semi-continuous and LC-continuous.

In this paper, we obtain decompositions of bitopological (1,2)*-A-continuous. In most of the occasions, our ideas are illustrated and substantiated by suitable examples.

## 2. Preliminaries

Throughout this paper, X and Y denote bitopological spaces (X, $\tau_1$, $\tau_2$) and (Y, $\sigma_1$, $\sigma_2$), respectively, on which no separation axioms are assumed.

**Definition 2.1**

Let S be a subset of X. Then S is called $\tau_{1,2}$ -open [10] if S = A ∪ B, where $A \in \tau_1$ and $B \in \tau_2$.

The complement of $\tau_{1,2}$-open set is called $\tau_{1,2}$ - closed.

**Definition 2.2**

Let A be a subset of X.

(i) The $\tau_{1,2}$-closure of A [10], denoted by $\tau_{1,2}$-cl(A), is defined by

$$\cap\{U : A \subseteq U \text{ and } U \text{ is } \tau_{1,2}\text{- closed}\};$$

(ii) The $\tau_{1,2}$-interior of A [10], denoted by $\tau_{1,2}$-int(A), is defined by

$$\cup\{U : U \subseteq A \text{ and } U \text{ is } \tau_{1,2}\text{-open}\}.$$

**Remark 2.3 [10]**

Notice that $\tau_{1,2}$-open sets need not necessarily form a topology.

Now we recall some definitions and results, which are used in this paper.

**Definition 2.4**

A subset A of X is said to be

(i) (1,2)*-semi-open [10] if $A \subseteq \tau_{1,2}$-cl($\tau_{1,2}$-int(A)),

(ii) (1,2)*- preopen [10] if $A \subseteq \tau_{1,2}$-int($\tau_{1,2}$-cl(A)),

(iii) (1,2)*-β-open [12] if $A \subseteq \tau_{1,2}$-cl($\tau_{1,2}$-int($\tau_{1,2}$-cl(A))),

(iv) (1,2)*-α-open [10] if $A \subseteq \tau_{1,2}$-int($\tau_{1,2}$-cl($\tau_{1,2}$-int(A))),

(v) regular (1,2)*-open [10] if A = $\tau_{1,2}$-int($\tau_{1,2}$-cl(A)).

The complements of the above- mentioned open sets are called their respective closed sets.

The family of all (1,2)*-semi-open (resp. (1,2)*-preopen, (1,2)*-α-open, (1,2)*- β-open, regular (1,2)*-open) sets of X will be denoted by (1,2)*-SO(X) (resp. (1,2)*-PO(X), (1,2)*-αO(X), (1,2)*-βO(X), (1,2)*-RO(X)).

The (1,2)*-preclosure, (1,2)*-pcl(A), of a subset A is the intersection of all (1,2)*-preclosed subsets of X that contain A.

**Example 2.5**

Let X = {a, b, c}, $\tau_1$ = {φ, X, {a}} and $\tau_2$ = {φ, X, {c}}. Then the sets in {φ, X, {a}, {c}, {a, c}} are $\tau_{1,2}$-open and the sets in {φ, X, {b}, {a, b}, {b, c}} are $\tau_{1,2}$- closed.

**Definition 2.6**

A subset S of X is said to be

(i) a (1,2)*-A- set [10] if S = G ∩ R, where G is $\tau_{1,2}$-open and R is regular (1,2)*-closed,

(ii) a (1,2)*-t-set [10] if $\tau_{1,2}$-int($\tau_{1,2}$-cl(S)) = $\tau_{1,2}$-int(S),

(iii) a (1,2)*-B-set [10] if S = G ∩ R, where G is $\tau_{1,2}$-open and R is a (1,2)*-t- set,

(iv) a locally (1,2)*-closed [9] if S = G ∩ R, where G is $\tau_{1,2}$-open and R is $\tau_{1,2}$-closed.

The family of all (1,2)*-A-sets (resp. locally (1,2)*-closed sets, (1,2)*-B-sets) of X will be denoted by (1,2)*-A(X) (resp. (1,2)*-LC(X), (1,2)*-B(X)).

The following Proposition is a direct consequence of the definition of (1,2)*-t-sets.

**Proposition 2.7**

A subset A of a space X is a (1,2)*-t-set if and only if it is (1,2)*-semi-closed.

**Proof**

Let A be a (1,2)*-semi-closed set. Then $\tau_{1,2}$-int($\tau_{1,2}$-cl(A)) ⊆ A.Therefore $\tau_{1,2}$-int($\tau_{1,2}$-cl(A)) ⊆ $\tau_{1,2}$-int(A). We know that $\tau_{1,2}$-int(A) ⊆ $\tau_{1,2}$-int($\tau_{1,2}$-cl(A)). Hence $\tau_{1,2}$-int(A) = $\tau_{1,2}$-int($\tau_{1,2}$-cl(A)). Then A is (1,2)*-t-set.

Conversely, let A be a (1,2)*-t-set. Then $\tau_{1,2}$-int($\tau_{1,2}$-cl(A)) = $\tau_{1,2}$-int(A). We have $\tau_{1,2}$-int($\tau_{1,2}$-cl(A)) ⊆ $\tau_{1,2}$-int(A) ⊆ A. Hence $\tau_{1,2}$-int($\tau_{1,2}$-cl(A)) ⊆ A. Therefore A is (1,2)*-semi-closed set.

From the definitions, we can see (1,2)*-LC(X) ⊆ (1,2)*-B(X).

**Example 2.8**

Let X = {a, b, c}, $\tau_1$ = {φ, X, {a}} and $\tau_2$ = {φ, X}. We have (1,2)*-LC(X) = {φ, X, {a}, {b, c}} and (1,2)*-B(X) = {φ, X, {a}, {b}, {c}, {b, c}}. Clearly {b} is (1,2)*-B-set but it is not locally (1,2)*-closed.

**Remark 2.9 [10]**

(i) A (1,2)*-A-set is a (1,2)*-B-set but not conversely.

(ii) Every regular (1,2)*-open set is $\tau_{1,2}$ - open but not conversely.

**Proposition 2.10**

Let A be an $\tau_{1,2}$-open subset of a space X. Then $\tau_{1,2}$-cl(A) is regular (1,2)*-closed.

**Proof**

Clearly, $\tau_{1,2}$-cl($\tau_{1,2}$-int($\tau_{1,2}$-cl(A))) $\subseteq$ $\tau_{1,2}$-cl(A). So we need only to show that $\tau_{1,2}$-cl(A) $\subseteq$ $\tau_{1,2}$-cl($\tau_{1,2}$-int($\tau_{1,2}$-cl(A))). Now, from A $\subseteq$ $\tau_{1,2}$-cl(A), we have A $\subseteq$ $\tau_{1,2}$-int($\tau_{1,2}$-cl(A)). Therefore, $\tau_{1,2}$-cl(A) $\subseteq$ $\tau_{1,2}$-cl($\tau_{1,2}$-int($\tau_{1,2}$-cl(A))).

**Proposition 2.11 [11]**

Let A be a subset of a space X. Then (1,2)*-pcl(A) = A $\cup$ $\tau_{1,2}$-cl($\tau_{1,2}$-int(A)).

**Remark 2.12 [9]**

A subset S of X is locally (1,2)*-closed if and only if S = U $\cap$ $\tau_{1,2}$-cl(S), where U is $\tau_{1,2}$ -open.

**Definition 2.13 [10, 12]**

Let f : (X, $\tau_1$, $\tau_2$) → (Y, $\sigma_1$, $\sigma_2$) be a map. Then f is said to be (1,2)*-semi-continuous if $f^{-1}$(G) $\in$ (1,2)*-SO(X) for each $\sigma_{1,2}$-open set G of Y.

*The (1,2)*-β-continuity and (1,2)*-A-continuity are analogously defined.* **Remark 2.14 [10]**(1,2)*-A-sets and (1,2)*-semi-open sets are independent.

Let X = {a, b, c}, $\tau_1$ = {φ, X, {a}, {b, c}} and $\tau_2$ = { φ, X, {b}, {a, c}}. Then the sets in {φ, X, {a}, {b}, {a, b}, {b, c}} are $\tau_{1,2}$-open and the sets in {φ, X, {a}, {b}, {c}, {a, c}, {b, c}} are $\tau_{1,2}$-closed. We have {c} is (1,2)*-A-set but not (1,2)*-semi-open.

Let X = {a, b, c}, $\tau_1$ = {φ, X, {a}} and $\tau_2$ = { φ, X}. Then the sets in {φ, X, {a}} are $\tau_{1,2}$-open and the sets in {φ, X, {b, c}} are $\tau_{1,2}$-closed. We have {a, b} is not (1,2)*-A-set but it is (1,2)*-semi-open.

## 3. PROPERTIES OF BITOPOLOGICAL (1,2)*-SETS

In this section, we provide three theorems concerning decompositions of bitopological (1,2)*-A-continuity. In the second theorem, a notion of (1,2)*-C-sets which is weaker than that of locally (1,2)*- closed sets is used.

**Definition 3.1**

A subset S of a space X is called (1,2)*-C-set if S = G $\cap$ R, where G is $\tau_{1,2}$-open and R is a (1,2)*-preclosed.

**Remark 3.2**

(i) The family of all (1,2)*-C-sets of X will be denoted by (1,2)*-C(X).

(ii) Every $\tau_{1,2}$-open set is (1,2)*-C-set.

(iii) Every (1,2)*-preclosed set is (1,2)*-C-set.

**Remark 3.3**

By definition 3.1, it is clear that (1,2)*-A(X) $\subseteq$ (1,2)*-LC(X) $\subseteq$ (1,2)*-C(X).

The following example shows that a (1,2)*-C-set need not be a locally (1,2)*- closed set and a locally (1,2)*-closed set need not be a (1,2)*-A-set.

**Example 3.4**

Let X = {a, b, c}, $\tau_1$ = {φ, X, {a, b}} and $\tau_2$ = {φ, X, {b, c}}. Then (1,2)*-A(X) = {φ, X, {a, b}, {b, c}}; (1,2)*-LC(X) = {φ, X, {a}, {c}, {a, b}, {b, c}} and (1,2)*-C(X) = P(X), where P(X) is the power set of X. Clearly, {b} is (1,2)*-C-set but it is not locally (1,2)*-closed. Moreover, {a} is locally (1,2)*-closed but it is not (1,2)*-A-set.

**Definition 3.5**

A bitopological space ( X, $\tau_1$, $\tau_2$ ) equipped with the family of all $\tau_{1,2}$-open sets will be called DRT-space if $\operatorname{int}_{\tau_1}$ (S) = $\operatorname{int}_{\tau_2}$ (S) for each $\tau_{1,2}$-closed subset S of X.

Let X = {a, b, c}, $\tau_1$ = {φ, X, {a}, {b, c}} and $\tau_2$ = { φ, X, {b}, {a, c}}. Then ( X, $\tau_1$, $\tau_2$ ) is not DRT-space since $\operatorname{int}_{\tau_1}$ ({a}) = {a} ≠ φ = $\operatorname{int}_{\tau_2}$ ({a}) for the $\tau_{1,2}$-closed subset {a} of X. However, in Example 3.4., ( X, $\tau_1$, $\tau_2$ ) is DRT-space.

**Theorem 3.6**

Let X be a DRT-space. Then an (1,2)*-A-set in X is (1,2)*-semi-open.

**Proof**

Let S = U ∩ C be an (1,2)*-A-set, where U is $\tau_{1,2}$-open and C = $\tau_{1,2}$-cl($\tau_{1,2}$-int(C)). Since S = U ∩ C, we have $\tau_{1,2}$-int(S) ⊃ U ∩ $\tau_{1,2}$-int(C). It is easily seen that $\tau_{1,2}$-int(S) ⊂ S ⊂ C, hence $\tau_{1,2}$-int(S) = $\tau_{1,2}$-int($\tau_{1,2}$-int(S)) ⊂ $\tau_{1,2}$-int(C). But $\tau_{1,2}$-int(S) ⊂ S ⊂ U, hence $\tau_{1,2}$-int(S) ⊂ U ∩ $\tau_{1,2}$-int(C). Therefore $\tau_{1,2}$-int(S) = U ∩ $\tau_{1,2}$-int(C). Now we prove S ⊂ $\tau_{1,2}$-cl($\tau_{1,2}$-int(S)). Let x ∈ S and V be an arbitrary $\tau_{1,2}$-open set containing x. Then U ∩ V is also an $\tau_{1,2}$-open set containing x. Since x ∈ C = $\tau_{1,2}$-cl($\tau_{1,2}$-int(C)), there is a point z ∈ $\tau_{1,2}$-int(C) such that z ≠ x and z ∈ U ∩ V. Hence z ∈ U ∩ $\tau_{1,2}$-int(C) = $\tau_{1,2}$-int(S). Therefore x ∈ $\tau_{1,2}$-cl($\tau_{1,2}$-int(S)) and S $\tau_{1,2}$-cl($\tau_{1,2}$-int(S)). From $\tau_{1,2}$-int(S) ⊂ S ⊂ $\tau_{1,2}$-cl($\tau_{1,2}$-int(S)) we know that S is (1,2)*-semi-open.

**Example 3.7**

Let X = {a, b, c}, $\tau_1$ = {φ, X, {a}} and $\tau_2$ = { φ, X}. Then the sets in {φ, X, {a}} are $\tau_{1,2}$-open and the sets in {φ, X, {b, c}} are $\tau_{1,2}$-closed. In this DRT-space, we have {a, b} is not (1,2)*-A-set but it is (1,2)*-semi-open.

**Theorem 3.8**

Let X be a DRT-space. Then (1,2)*-A(X) = (1,2)*-SO(X) ∩ (1,2)*-LC(X).

**Proof**

Let S ∈ (1,2)*-A(X). Then S = G ∩ R where G is $\tau_{1,2}$-open and R is regular (1,2)*-closed. Clearly S is locally (1,2)*- closed. Now $\tau_{1,2}$ -int(S) = G ∩ $\tau_{1,2}$-int(R), so that S = G ∩ $\tau_{1,2}$-cl($\tau_{1,2}$-int(R)) ⊆ $\tau_{1,2}$-cl(G ∩ $\tau_{1,2}$-int(R)) = $\tau_{1,2}$-cl($\tau_{1,2}$-int(S)) and hence S is (1,2)*-semi-open.

Conversely, let S be (1,2)*-semi-open and locally (1,2)*-closed, so that S ⊆ $\tau_{1,2}$-cl($\tau_{1,2}$-int(S)) and S = U ∩ $\tau_{1,2}$-cl(S), where U is $\tau_{1,2}$-open. Then $\tau_{1,2}$-cl(S) = $\tau_{1,2}$-cl($\tau_{1,2}$-int(S)) and so is regular (1,2)*-closed. Hence S is an (1,2)*-A-set.

**Proposition 3.9**

Let A be a subset of X. Then A is (1,2)*-preclosed if and only if $\tau_{1,2}$- cl($\tau_{1,2}$-int(A)) ⊆ A.

**Lemma 3.10**

Let ( X, $\tau_1$, $\tau_2$ ) be a DRT-space and G be a subset of X. Then G ∈ (1,2)*-C(X) if and only if G = R ∩ (1,2)*-pcl(G) for some $\tau_{1,2}$-open set R.

**Proof**

Suppose that $G = R \cap (1,2)^*\text{-pcl}(G)$ for some $\tau_{1,2}$-open set R. It is obvious that $G \in (1,2)^*\text{-}C(X)$, since $(1,2)^*\text{-pcl}(G)$ is (1,2)*- preclosed.

Conversely , let $G \in (1,2)^*\text{-}C(X)$. Then $G = R \cap A$ where R is $\tau_{1,2}$-open and A is (1,2)*-preclosed. From $G \subseteq A$, we have $(1,2)^*\text{-pcl}(G) \subseteq (1,2)^*\text{-pcl}(A) = A \cup \tau_{1,2}\text{-cl}(\tau_{1,2}\text{-int}(A))$. Since A is (1,2)*-preclosed, by Proposition 3.9, we have $(1,2)^*\text{-pcl}(G) \subseteq A$. Thus, $R \cap (1,2)^*\text{-pcl}(G) \subseteq R \cap A = G \subseteq R \cap (1,2)^*\text{-pcl}(G)$, which shows that $G = R \cap (1,2)^*\text{-pcl}(G)$ with R is $\tau_{1,2}$-open.

**Lemma 3.11**

Let ( X, $\tau_1$, $\tau_2$ ) be a DRT-space and G be a subset of X. Then $G = R \cap \tau_{1,2}\text{-cl}(\tau_{1,2}\text{-int}(G))$ for some $\tau_{1,2}$-open set R if and only if $G \in (1,2)^*\text{-}C(X) \cap (1,2)^*\text{- }SO(X)$.

**Proof**

Suppose that $G = R \cap \tau_{1,2}\text{-cl}(\tau_{1,2}\text{-int}(G))$ where R is $\tau_{1,2}$-open. Then $G \subseteq \tau_{1,2}\text{-cl}(\tau_{1,2}\text{-int}(G))$ which shows that $G \in (1,2)^*\text{-}SO(X)$. Moreover, $\tau_{1,2}\text{-cl}(\tau_{1,2}\text{-int}(G))$ is $\tau_{1,2}$ -closed and therefore (1,2)*-preclosed. So, $G \in (1,2)^*\text{-}C(X)$.

Conversely, let $G \in (1,2)^*\text{-}C(X) \cap (1,2)^*\text{-}SO(X)$. From $G \in (1,2)^*\text{-}C(X)$, we have from Lemma 3.10, that $G = R \cap (1,2)^*\text{-pcl}(G)$, where R is $\tau_{1,2}$-open. From $G \in (1,2)^*\text{-}SO(X)$, we have $G \subseteq \tau_{1,2}\text{-cl}(\tau_{1,2}\text{-int}(G))$. But $(1,2)^*\text{-pcl}(G) = G \cup \tau_{1,2}\text{-cl}(\tau_{1,2}\text{-int}(G))$ (see Proposition 2.11). Thus $G = R \cap \tau_{1,2}\text{-cl}(\tau_{1,2}\text{-int}(G))$ with R is $\tau_{1,2}$ -open.

**Theorem 3.12**

Let ( X, $\tau_1$, $\tau_2$ ) be a DRT-space. Then $(1,2)^*\text{-}A(X) = (1,2)^*\text{-}C(X) \cap (1,2)^*\text{-}SO(X)$.

**Proof**

It is clear that $(1,2)^*\text{-}A(X) \subseteq (1,2)^*\text{-}C(X) \cap (1,2)^*\text{-}SO(X)$.

Conversely, let $G \in (1,2)^*\text{-}C(X) \cap (1,2)^*\text{-}SO(X)$. Then by Lemma 3.11, $G = R \cap \tau_{1,2}\text{-cl}(\tau_{1,2}\text{-int}(G))$, where R is $\tau_{1,2}$ -open. Since $\tau_{1,2}\text{-int}(G)$ is $\tau_{1,2}$ -open, by Proposition 2.10, $\tau_{1,2}\text{-cl}(\tau_{1,2}\text{-int}(G))$ is regular (1,2)*-closed. Therefore $G \in (1,2)^*\text{-}A(X)$.

**Remark 3.13**

It is clear from the definition 2.4 that $(1,2)^*\text{-}SO(X) \subseteq (1,2)^*\text{-}\beta O(X)$. However, the converse is not true.

**Example 3.14**

Let X = {a, b, c}, $\tau_1 = \{\varphi, X, \{a\}\}$ and $\tau_2 = \{\varphi, X, \{b, c\}\}$. Then the sets in $\{\varphi, X, \{a\}, \{b, c\}\}$ are $\tau_{1,2}$-open and $\tau_{1,2}$-closed. Clearly, $\{b\} \in (1,2)^*\text{-}\beta O(X)$, but it is not (1,2)*- semi-open.

**Theorem 3.15**

Let ( X, $\tau_1$, $\tau_2$ ) be a DRT-space. Then $(1,2)^*\text{-}A(X) = (1,2)^*\text{-}\beta O(X) \cap (1,2)^*\text{-}LC(X)$.

**Proof**

If $G \in (1,2)^*\text{-}A(X)$, then, obviously, $G \in (1,2)^*\text{-}\beta O(X) \cap (1,2)^*\text{-}LC(X)$. Conversely, let $G \in (1,2)^*\text{-}\beta O(X) \cap (1,2)^*\text{-}LC(X)$. From $G \in (1,2)^*\text{-}\beta O(X)$, we have $G \subseteq \tau_{1,2}\text{-cl}(\tau_{1,2}\text{-int}(\tau_{1,2}\text{-cl}(G)))$. From $G \in (1,2)^*\text{-}LC(X)$, we have, by Remark 2.12, $G = U \cap \tau_{1,2}\text{-cl}(G)$, where U is $\tau_{1,2}$-open. So $G \subseteq U$, which implies $G \subseteq U \cap \tau_{1,2}\text{-cl}(\tau_{1,2}\text{-int}(\tau_{1,2}\text{-cl}(G))) \subseteq U \cap \tau_{1,2}\text{-cl}(G) = G$. Hence we have $G = U \cap \tau_{1,2}\text{-cl}(\tau_{1,2}\text{-int}(\tau_{1,2}\text{-cl}(G)))$. By Proposition 2.10, $\tau_{1,2}\text{-cl}(\tau_{1,2}\text{-int}(\tau_{1,2}\text{-cl}(G)))$ is regular (1,2)*-closed, since $\tau_{1,2}\text{-int}(\tau_{1,2}\text{-cl}(G))$ is $\tau_{1,2}$-open. Therefore, $G \in (1,2)^*\text{-}A(X)$.

## 4. DECOMPOSITIONS OF (1,2)*-A-CONTINUITY

**Definition 4.1**

A mapping f : X→Y is said to be (1,2)*-continuous [9] if $f^{-1}$ (V) is $\tau_{1,2}$-open in X for every $\sigma_{1,2}$ -open set V of Y.

**Definition 4.2**

A mapping f : X→Y is said to be (1,2)* -C-continuous if $f^{-1}$ (V) $\in$ (1,2)*-C(X) for every $\sigma_{1,2}$ -open set V of Y.

**Definition 4.3**

A mapping f : X→Y is said to be (1,2)*-LC-continuous [9] if $f^{-1}$(V) $\in$ (1,2)*-LC(X) for every $\sigma_{1,2}$ -open set V of Y.

**Theorem 4.4**

Let ( X, $\tau_1$, $\tau_2$ ) be a DRT-space. Then a mapping f : X→Y is (1,2)*-A-continuous if and only if it is (1,2)*-semi-continuous and (1,2)* -C-continuous.

**Proof**

It follows from Theorem 3.12.

**Theorem 4.5**

Let ( X, $\tau_1$, $\tau_2$ ) be a DRT-space. Then a mapping f : X→Y is (1,2)*-A-continuous if and only if it is (1,2)*-semi-continuous and (1,2)*-LC-continuous.

**Proof**

It follows from Theorem 3.8.

**Theorem 4.6**

Let ( X, $\tau_1$, $\tau_2$ ) be a DRT-space. Then a mapping f : X→Y is (1,2)*-A-continuous if and only if it is (1,2)*-β-continuous and (1,2)*-LC-continuous.

**Proof**

It follows from Theorem 3.15.

# REFERENCES

**[1]** D. Andrijevic, Semi-preopen sets, Mat. Vesnik, 38(1986), 24-32.

**[2]** N. Bourbaki, General Topology, Part I, Addison-Wesley (Reading, Mass, 1996).

**[3]** J. Dugundji, Topology, Allyn and Bacon (Boston, 1972).

**[4]** Y. Erguang and Y. Pengfei, On decomposition of A-continuity, Acta Math. Hungar., 110 (4) (2006), 309-313.

**[5]** M. Ganster and I. L. Reilly, A decomposition of continuity, Acta Math. Hungar., 56 (1990), 299-301.

**[6]** K. Kayathri, O. Ravi, M. L. Thivagar and M. Joseph Israel, Decompositions of (1,2)*-rg-continuous maps in bitopological spaces, Antarctica J. Math., 6 (1) (2009), 13-23.

**[7]** A. S. Mashhour, M. E. Abd El-Monsef and S. N. El-Deeb, On precontinuous and weak precontinuous mappings, Proc. Math. and Phys. Soc. Egypt, 53 (1982), 47-53.

**[8]** O. Njastad , On some classes of nearly open sets, Pacific J. Math, 15 (1965), 961-970.

**[9]** O. Ravi, M. Lellis Thivagar and E. Ekici, Decompositions of (1,2)*- continuity and complete (1,2)*-continuity in bitopological spaces, Analele Universităţii Din Oradea Fasicola, Matematica, Tom XV (2008), 29-37.

**[10]** O. Ravi, M. Lellis Thivagar and E. Ekici, On (1,2)*-sets and decompositions of bitopological (1,2)*-continuous mappings. Kochi J. Math., 3 (2008), 181-189.

**[11]** O. Ravi, M. Lellis Thivagar and E. Hatir, Decomposition of (1,2)*-continuity and (1,2)*-α-continuity, Miskolc Mathematical notes, 10(2) (2009), 163-171.

**[12]** O. Ravi, M. Lellis Thivagar and M. Joseph Israel, Some decompositions of bitopological (1,2)*-α-continuity. (Submitted)

**[13]** J. Tong, A decomposition of continuity, Acta Math. Hungar., 48 (1986), 11-15.

**[14]** J. Tong, On decomposition of continuity in topological spaces, Acta Math. Hungar., 54 (1989), 51-55.

# Designing Robust Hybrid Wireless Sensor Network: Dual Technology Aspect

Ajay Jangra[1], Rajesh Verma[2], Priyanka[3]

[1] CSE department, U.I.E.T. Kurukshetra University, Kurukshetra, INDIA
[2]CSE [3]ECE department, Kurukshetra Institute of Engineering & Technology, INDIA

er_jangra@yahoo.co.in, vermar.rajesh1974@gmail.com , priyanka.jangra@gmail.com

***Abstract*****:** Wireless Sensor Networks are infrastructure less network characterized by dense node deployment, low power, unreliable sensor node, frequent topology change, and severe power, computation and memory constraints because the nodes will often operate with finite battery resources. Because of their wide range of applications, both research community and industry focus on the deployment of wireless sensor networks. In this paper we proposed a framework for hybrid wireless sensor networks using two popular wireless standards deployment. This paper elaborates the sensor network architecture and different wireless technologies. The co-existence behavior is also analyzed with respect to cases of interference occurrence in hybrid model.

***Keywords:*** wireless sensor network, co-existence, interference, network security

## 1. Introduction

Sensor networks are composed of a large number of sensing nodes, which are equipped with limited computing, radio communication capabilities characterized by dense node deployment, unreliable sensor node, frequent topology change, and severe power, computation and memory constraints because the nodes will often operate with finite battery resources. A typical network configuration consists of sensors working unattended and transmitting their observation values to some processing or control center, the so-called sink node, which serves as a user interface. Due to the limited transmission range, sensors that are far away from the sink deliver their data through *multihop* communications, i.e., using intermediate nodes as relays. In this case a sensor may be both a data source and a data router. Most application scenarios for sensor networks involve battery-powered nodes with limited energy resources. Recharging or replacing the sensors battery may be inconvenient, or even impossible in harsh working environments. Thus, when a node exhausts its energy, it cannot help but ceases sensing and routing data, possibly degrading the coverage and connectivity level of the entire network. A widely employed energy-saving technique is to place nodes in sleep mode, corresponding to low-power consumption as well as to reduce operational capabilities. [5, 6]

WSN is an emerging technology that promises a wide range of potential applications in both civilian and military areas, and has therefore received tremendous attention from both academia and industry in recent years. Depending on the application the large, sudden, and correlated synchronized impulses of data sent to a small number of sinks or base station without significantly disrupting the performance (i.e., fidelity) of the sensing application. This high generation of data packets is usually uncontrolled and often leads to congestion. One of the major challenges wireless sensor networks face today is security. While the deployment of sensor nodes in an unattended environment makes the networks vulnerable to a variety of potential attacks, the inherent power and memory limitations of sensor nodes makes conventional security solutions unfeasible. The sensing technology combined with processing power and wireless communication makes it profitable for being exploited in great quantity in future. The wireless communication technology also acquires various types of security threats[4,6,7]

## 2. Architecture of WSN

As shown in figure architecture of WSN consists of various components (Sensor, Sensor Unit, Memory, Processing Unit Power supply, Communication Unit, Radio, Mobilizer, Power Generator and Location Finding System). Each component has its own relative task that helps nodes to communicate with each other in WSN. Sensors sense the information from the environment and transfer it to Sensor Unit for further processing. After taking sensed information transfer it to processing unit that apply some calculation on it and make that raw information into a compatible format of given network. Processing unit is directly connected with Memory and Power Supply, Memory is used to store the processed information so that it may be further used and Power supply provides essential power so that all components work well. After that all processed information transfer to the Communication Unit (media) through which the requested information could easily transfer to the client through radio signals or by some other means. Power Supply is also directly connected with Power Generator that generates required power for all the components.

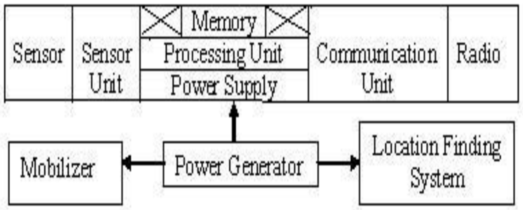


**Figure.1** WSN Architecture

Mobilizer is used to move sensor node when assigning any task to the node. A Location finding/detection system is used for finding the location with high accuracy [4, 5,8]

## 3. Applications of Wireless Sensor Network

Depending upon the requirement and characteristics of system, wide variety of application are there which require constant monitoring and detection of specific event. [5,8]

### 3.1 Military Applications

like military command & control communications, computing, intelligence, surveillance, reconnaissance & targeting systems, monitoring & Reconnaissance enemy forces, Targeting, Battle damage assessment and Military situation Awareness

### 3.2 Environmental Applications

like Habitat monitoring, Agriculture research-include sensing of pesticide, soil moisture, PH levels, Habitat Exploration of Animal, Forest Fire and Flood detection and Ocean Monitoring.

### 3.3 Structural health monitoring

like i) *Heavy industrial monitoring*: In warehouse monitoring to improve inventory control system, Manufacturing monitoring and Industrial automation and factory process control ii) *Health or Medical Applications* : monitor patient physiological data such as blood pressure or heart rate.

### 3.4 Home Application

Switch on/off, Monitoring product quality, Managing and Monitoring Inventory system, music, and lighting can be set automatically and Control of temperature and airflow adjustment.

## 4. Communication Technologies (WSN)

In this section we elaborate the various wireless technologies used for deploying a wireless sensor network. Characteristics and features of these technologies are explained below.

### 4.1 Infrared

Infrared is a wireless communication medium for aligned nodes. It has limited radius of communication and works on ISM band. It has limited processing capabilities as compare to other wireless technologies used in WSN. For a short distance and small data it is reliable to communicate and economic as well.[13]

### 4.2 Bluetooth (IEEE 802.15)

Bluetooth is wireless LAN technology design to connect devices of different functions such as telephones, notebooks, computers (desktop and laptop), cameras, printers, coffee makers, ad so on. Bluetooth follows two types of networks: piconet and scatternet. *Piconet:* A small network and can have up to eight nodes, among them one of which is called master; the rest are called slaves. In order to between master and slave can be one-to-one or one-to-many. The maximum numbers of slaves in a piconet can have seven. *Scatternet:* Two or more piconet combined together to form a scatternet. A slave node in one piconet can be the master in another piconet. This node can receive messages from both the piconets and act as master/slave at the same time and they can communicate by mans of multi-hopping [3,1,11,13]

### 4.3 Wireless Fidelity (Wi-Fi, IEEE 802.11)

Wi-Fi is the transmission of radio signals. In order to define data transmission and manages location independent network access using radio signals on the bases of that we can call it as a packet protocol. The structure of physical/link layer interface of Wi-Fi is similar to Ethernet. The layers above the physical and data link layers include TCP/IP. By the above introduction we can clearly see all programs and applications for TCP/IP that can run on an Ethernet can also be run on Wi-Fi interface. [1, 2, 10]

### 4.4 ZigBee (IEEE 802.15.4)

ZigBee technology (similar to Bluetooth) provides low data rate and low power connectivity for gadgets that follows low battery life as long as several months to several years. ZigBee has low cost and built to perform wireless networking protocol targeted towards automation and remote control application. The main features of ZigBee are developed for application with relaxed throughput requirements which cannot handle the power consumption of heavy protocol stack, very low power consumption, low data rate in an ad hoc self-organizing network among inexpensive fixed, low cost, network flexibility moving and portable devices. [12, 13]

### 4.5 WiMax (Worldwide Interoperability for Microwave Access, IEEE 802.16)

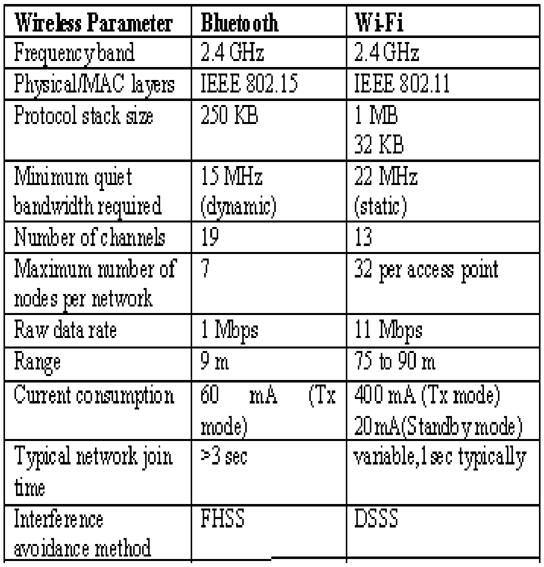
As we know in the modern era of broadband wireless access, WiMAX (IEEE 802.16) is an outstanding, well suitable, useful connection oriented protocol to which access fixed and mobile with low cost, high reliability, very high data rate and better efficiency. WiMAX standard defines the formal speciation for deployment of broadband wireless metropolitan area networks (wireless MANs) and with the help of WiMAX (802.16) we can access broadband anytime, on virtually any device and anywhere. While moving at a speed of approximate 125 kmph, in that speed we can also be able to access broadband. WiMAX (802.16) has data rate up to 70 mbps and can be able to work in both license free and licensed band and have high efficiency. WiMAX can have coverage area approximately is up to 50 km. [9, 10]

### 4.6 Mobile-Fi (IEEE 802.20):-

Mobile-Fi (IEEE 802.20) is the youngest IEEE standard. In order to access fully mobile broadband, it is the first standard designed to carry native IP traffic with licensed airwave below 3.5 GHz and provides symmetrical wireless rates over long distance (~15km). If we compare all the factors with other technologies for ad hoc network, it has lower power than WiMax but has high mobility and has latency of 10 ms. This features can pursue even with fast moving vehicles and we can also compare it with 3G has 500ms and for optimization of packets uses small antennas.[12,13]

## 5. Comparison of Wireless Standards

Now it is better to check the performance of Bluetooth and Wi-Fi and analyze them. We are comparing *(as shown in table 1)* the two widely used wireless technologies i.e. Bluetooth and Wi-Fi and check how they are different and compatible to each other. On the bases of some crucial wireless parameter as in table we analyze which one is reliable for what kind of network.



| Wireless Parameter | Bluetooth | Wi-Fi |
|---|---|---|
| Frequency band | 2.4 GHz | 2.4 GHz |
| Physical/MAC layers | IEEE 802.15 | IEEE 802.11 |
| Protocol stack size | 250 KB | 1 MB<br>32 KB |
| Minimum quiet bandwidth required | 15 MHz (dynamic) | 22 MHz (static) |
| Number of channels | 19 | 13 |
| Maximum number of nodes per network | 7 | 32 per access point |
| Raw data rate | 1 Mbps | 11 Mbps |
| Range | 9 m | 75 to 90 m |
| Current consumption | 60 mA (Tx mode) | 400 mA (Tx mode)<br>20mA(Standby mode) |
| Typical network join time | >3 sec | variable,1 sec typically |
| Interference avoidance method | FHSS | DSSS |

**Table 1** Comparison of Bluetooth and Wi-Fi Technologies

## 6. Co-existence scenario

This paper presents the coexistence scenario of Wi-Fi and blue-tooth wireless technology. Researchers claims that Wi-Fi and Bluetooth do not compete, because of high data rate and high power of the former. Wi-Fi used DSSS instead of FHSS and is a very high power, high cost scheme than Bluetooth with much greater range (45m indoor, 300m outdoors) let's take a look at the actual capabilities of these two technologies, as well as the corresponding requirements of the applications and real world considerations that affect the performance of the systems. [1, 2, 11, 13]

## 7. Hybrid wireless sensor network modal *(Co-existence aspect)*: The Proposed Model

In this paper we present a hybrid wireless sensor network modal which deploy both Bluetooth and Wi-Fi enable wireless devices. Wi-Fi nodes having communication range more then Bluetooth nodes so, in presented modal Wi-Fi node communicate each other with covering large geographic area and region between Wi-Fi node is covered by Bluetooth nodes. Network using Bluetooth nodes is characterized as low power requirement, low installation/maintain cost, small size, easy to install, secure, small range, multi-hop communication network. Its better to construct a low cost network rather then deploying high cost/power required Wi-Fi devices and Bluetooth nodes provides more information/data of a region *(i.e. more Bluetooth node in same area)*. Bluetooth forms scatter net to cover great area.

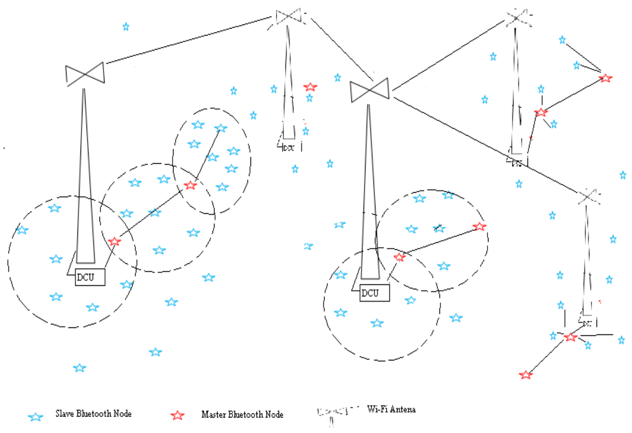


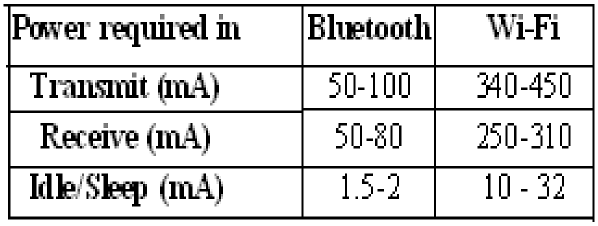
**Figure.2** hybrid wireless sensor network

As modal shown in fig.2 where BLUE nodes represents the slave Bluetooth node and RED nodes represents the master Bluetooth node. There is one master node in every piconet. Any slave node in a piconet can behave as a master node for other piconet this process leads a scatter net. There is a DCU (Data Compatible Unit) which works as an interfacing device between Wi-Fi and Bluetooth devices. DCU receives the data from scatter net , process that data in to useful information (removal of redundant data) and convert that data in the form as required by Wi-Fi devices and vice-versa. The DCU contain buffers to store un-processed and processed data, controls the flow of data, cancel the duplicate data (bandwidth utilization). Bluetooth devices separated by short distance which provides accurate and high degree of information. This proposed model could also be feasible and reliable for both stationary and mobile node.

Routing and relaying are the two approaches to provide multi-hop connectivity in wireless ad hoc networks. In relaying, the switching or forwarding decisions take place at the MAC layer, while routing schemes make these decisions at the network layer. Routing is a widely researched and practiced solution, whereas relaying has not been adequately explored in wireless ad hoc networks. The IEEE 802.11 MAC protocol is the standard for wireless LANs; it is widely used in test beds and simulations for wireless multi-hop ad hoc networks. However, this protocol was not designed for multi-hop networks. Although it can support some ad hoc network architecture, it is not intended to support the wireless mobile ad hoc network, in which multi-hop connectivity is one of the most prominent features.

### 7.1 Power management

Power consumption is a critical consideration as it directly affects device battery life specially in case of wireless sensor networks. This consideration is obviously most crucial for devices that spend most or all their operating hours on battery power. Bluetooth was designed to be a small-form factor, low-cost, low-power technology. The Bluetooth specification incorporates a number of power saving features in order to keep power use to a minimum. These features include a *standby mode* as well as four connected modes – (*parked, hold, sniff and active*). An adaptive transmission power feature further minimizes power use. Wi-Fi offers a power save mode in which STAs "*sleep*," then *reawaken* periodically to check for messages. Table 2 shows the power requirement for both technologies. [3, 10, 11]



| Power required in | Bluetooth | Wi-Fi |
|---|---|---|
| Transmit (mA) | 50-100 | 340-450 |
| Receive (mA) | 50-80 | 250-310 |
| Idle/Sleep (mA) | 1.5-2 | 10 - 32 |

**Table 2**. Typical Bluetooth and Wi-Fi power requirements comparison

Bluetooth devices required very less amount of power as compared to Wi-Fi. The implication is that Bluetooth will drain the battery less quickly than will Wi-Fi, making Bluetooth a more attractive option for users with smaller devices.

### 7.2 Security

Security threats to any network include the physical security of the network, unauthorized access, eavesdropping and attacks from within the network's authorized user community. Wireless sensor networks are more susceptible to threats due to the fact that signals from the network are more accessible to potential hackers. Nevertheless, both Bluetooth and Wi-Fi are highly resistant to security threats based on the security procedures implemented in the protocols as well as commonly used adjunct security procedures. Some of the basic security concepts addressed by the technologies include: [3, 4,]

***i) Authentication*** – verifying who is at the other end of a link between devices;

***ii) Authorization*** – the process of determining what a device or user is *allowed to do*;

***iii) Encryption*** – disguising information to make it *inaccessible to unwanted listeners*.

### 7.3 Interference

Wi-Fi WLAN installations and the anticipated growth in the use of Bluetooth-enabled devices ensure that the two technologies will find themselves sharing space because both use 2.4GH spectrum (ISM band). The question naturally arises as to how these two technologies will get along. Interference between Bluetooth and Wi-Fi will occur any time there is an overlap of both time and frequency between transmissions associated with each technology.

### 7.4 Cases of interference occurrence

**Case-I** Wi-Fi receiver senses a Bluetooth signal at the same time when a Wi-Fi signal is being sent to it.

**Case-II** Bluetooth receiver senses a Wi-Fi signal at the same time when a Bluetooth signal is being sent to it.

### 7.5 Discussion

Bluetooth is considered less susceptible to interference because of its frequency hopping capability. It has the ability to "hop away" from interfering signals and does so pseudo-randomly. Wi-Fi is considered more susceptible to interference because it inhabits a specific 22 MHz pass band and cannot "hop away" from interference as Bluetooth can. Its collision avoidance mechanism also results in retransmission following Bluetooth interference events, leading to successful transmission but reduced throughput. [2]

## 8. Interference reduction methods

To minimize any potential interference between Wi-Fi and Bluetooth systems can follow a few simple guidelines to help ensure optimal coexistence between the two technologies.

i) Ensure adequate spacing between Wi-Fi APs and Bluetooth APs to minimize the probability of interference b/w the two types of devices most likely to be transmitting.

ii) Do not deploy any devices that are simultaneously equipped with both Bluetooth and Wi-Fi.

iii) Increase the number of Wi-Fi APs deployed in order to yield a shorter average distance between wireless LAN STAs and APs.

## 9. Conclusion

Low cost, performance and security are remains the major objectives of a sensor network. The hybrid sensor network framework proposed in this paper deployed Bluetooth and Wi-Fi in a same network model. Bluetooth nodes perform low cost and short distance scatternet based communication, which can enhanced the sensor networks performance by deploying heterogeneous nodes for observing different parameters for a common network, co-existed with Wi-Fi. It is also claim that both Bluetooth and Wi-Fi technologies are complementary to each other but not competing.

## References

1. Shoemake, M., *Wi-Fi (802.11b) and Bluetooth: Coexistence Issues and Solutions for the 2.4 GHz ISM Ban*d, February 2001.
2. Mobilian Corporation, *Wi-Fi (802.11b) and Bluetooth: An Examination of Coexistence Approache*s, 2001.
3. Jakobbson, M. et al., *Security Weaknesses in Bluetoot*h, February 2001
4. Schenk, R. et al., *Wireless LAN Deployment and Security Basic*s, ExtremeTech.com, August 2001

5. .F. Akyildiz, W.Su, Y.Sankarasubramaniam, E. Cayirci, “Wireless Sensor Networks: A Survey”, IEEE commun. Mag., published by Elsevier Science B.V. in 2002.
6. Abdul-Halim Jallad and Tanya Vladimirova,”Data-centicity in wireless sensor network”, springer-Verlag London limited 2009.
7. Raymond Mulligan, Habib M. Ammari,” Coverage in Wireless Sensor Network: A Survey” , network protocol and algorithm, Vol 2, published in 2010
8. Chiara Buratti, Andrea conti, Davide Dardari, Roberto Verdone,” An Overview on Wireless Sensor Network Technology and Evolution”, ISSN 1428-8220, published on 31August 2009.
9. N. Gupta and G. Kaur, “WiMAX: Applications,” ser. The WiMAX Handbook, S. Ahson and M. Ilyas, Eds. CRC Press (Taylor and Francis Group), 2008, ch. 3: WiMAX Technology for Broadband Wireless Communication, pp. 35 – 54, ISBN 9781420045474.
10. “An Introduction to Wi-Fi” 019-0170 • 090409-B USA 2007-2008, Caroline Gabriel, “WiMax”, ARCchart ltd., London EC2A 1LN
11. Ajay Jangra, Sunita Beniwal, Anil Garg, “Co-existence behavior study of Bluetooth & Wi-Fi for 2.4 GHz ISM band”2006
12. Sinem Coleri ,Ergen, ZigBee IEEE 802.15.4” September 10, 2004
13. Ajay Jangra, Nitin Goel, Priyanka , Komal Kumar Bhatia, “IEEE WLANs Standards for Mobile Ad-hoc Networks (MANETs): Performance Analysis” global Journals of Computer Science and Technology (GJCST) Volume 10 Issue 14 Version 1.0 november 2010.

## Author Biographies

***Er. Ajay Jangra*** (11th March 1979) received his B.Tech. in 2001 from kurukshetra university, kurukshetra india, M.Tech. Computer Engineering (with honour) in 2004 from YMCA institute of engineering & technology *(now YMCA university of science & technology)*, faridabad india & M.B.A.(information technology) from guru jambheshwar university, hisar, india in 2008. Presently working as *Assistant Professor* in CSE department UIET, kurukshetra university, kurukshetra,india. He has published 21 research papers repute journals/conference. His area of interest is smart sensors, Ad-hoc & sensor networks, digital & data communication, Mobile computing, software engineering, testing etc.

**Dr. Rajesh Verma** (May 1974) received his Bachelor degree in 1994 from UCK, kurukshetra university, kurukshetra india, Masters in Computer Application in 1999 and Ph.D in Computer Science & Engg. In 2009 from Department of Computer Science and Applications from kurukshetra university, kurukshetra india. Presently he is *Professor and Head* in CSE department Kurukshetra Institute of Technology and Management, Kurukshetra, INDIA. He has published 30 research papers repute journals/conference. His area of interest is smart sensors, Ad-hoc & sensor networks, AI, Computer Architecture, Simulation, software engineering, testing etc.

***Ms. Priyanka*** (18th October1980) received his B.Tech. (Electronics & communication engineering) in 2002 from kurukshetra university, kurukshetra india, M.Tech. (electronics & communication engineering) with honour from N.I.T. kurukshetra ,india in 2006 & M.B.A.(information technology) from guru jambheshwar university, hisar, india in 2008.Presently working as *Sr. Assistant Professor* in ECE department, kurukshetra institute of technology & management, kurukshetra, india. She has published 10 research papers reputed international journals. Her area of interest is Ad-hoc & sensor networks, Mobile computing, analog & digital communication, antenna & wave propagation etc.

# Artificial Intelligent &PI in Load Frequency Control of Interconnected Power system

**Surya Prakash** [1*] **S K Sinha** [2]

[1]Department of Electrical & Electronics Engineering, Shepherd School of Engineering & Technology,
Sam Higginbottom Institute of Agriculture, Technology & Sciences- Deemed University, Allahabad, India
[2]Department of Electrical Engineering , Kamala Nehru Institute of Technology, Sultanpur-UP, India

*corresponding author
sprakashgiri0571@yahoo.com[1], sinhask98@engineer.com[2]

***Abstract***: This paper present the use of Artificial Intelligent and conventional PI controller to study the load frequency control of interconnected power system. In the proposed scheme, control methodology developed using Artificial Neural Network (ANN), Fuzzy Logic controller (FLC) and PI controller for interconnected hydro-thermal power system. The control strategies guarantees that the steady state error of frequencies and inadvertent interchange of tie-lines power are maintained in a given tolerance limitations. The performances of these controllers are simulated using MATLAB/SIMULINK package. A comparison of PI controller, Fuzzy controller and ANN controller based approaches shows the superiority of proposed ANN based approach over Fuzzy and conventional one for same conditions. The simulation results also tabulated as a comparative performance in view of settling time and peak over shoot.

***Keywords*** **:** Load Frequency Control(LFC), Fuzzy Logic Controller, ANN Controller, PI Controller, Area Control error(ACE), Tie-line, MATLAB / SIMULINK**.**

## 1. Introduction

Automatic Generation Control (AGC) or Load Frequency Control is a very important issue in power system operation and control for supplying sufficient and reliable electric power with good quality. An interconnected power system can be considered as being divided into control area, all generators are assumed to form a coherent group[1]. In the steady state operation of power system, the load demand is increased or decreased in the form of Kinetic Energy stored in generator prime mover set, which results the variation of speed and frequency accordingly. Therefore, the control of load frequency is essential to have safe operation of the power system[2]-[4]. Automatic generation control (AGC) is defined as, the regulation of power output of controllable generators within a prescribed area in response to change in system frequency, tie-line loading, or a relation of these to each other, so as to maintain the schedules system frequency and / or the established interchange with other areas within predetermined limits [5]. Therefore, a control strategy is needed that not only maintains constancy of frequency and desired tie-power flow but also achieves zero steady state error and inadvertent interchange. Among the various types of load frequency controllers, the most widely employed is the conventional proportional integral (PI) controller. The PI controller is very simple for implementation and gives better dynamic response, but their performances deteriorate when the complexity in the system increases due to disturbances like load variation boiler dynamics[6-7]. Therefore, there is need of a controller which can overcome this problem. The Artificial Intelligent controller like Fuzzy and Neural control approach is more suitable in this respect. Fuzzy system has been applied [8] to the load frequency control problems with rather promising results. The salient feature of these techniques is that they provide a model- free description of control systems and do not require model identification. The fuzzy controller offers better performance over the conventional controllers, especially, in complex and nonlinearities associated system. In [9] Fuzzy control was applied to the two region interconnected reheat thermal and hydro power system. However, it is demonstrated good dynamics only when selecting the specific number of membership function, so that the method had limitation. To over come this Artificial Neural Network (ANN) controller, which is an advance adaptive control configuration, is used because the controller provides faster control than the others. [10 ].

In this paper, the performance evaluation based on PI, Fuzzy controller and Artificial Neural controller for two area interconnected hydro-thermal power plant is proposed. To enhance the performance of PI, fuzzy and neural controller sliding surface is included. The sliding concept arises due to variable structure concept. The objective of VSC has been greatly extended from stabilization to other control functions. The most distinguished feature of VSC is its ability to result in very robust control systems, in many cases it results invariant control system.

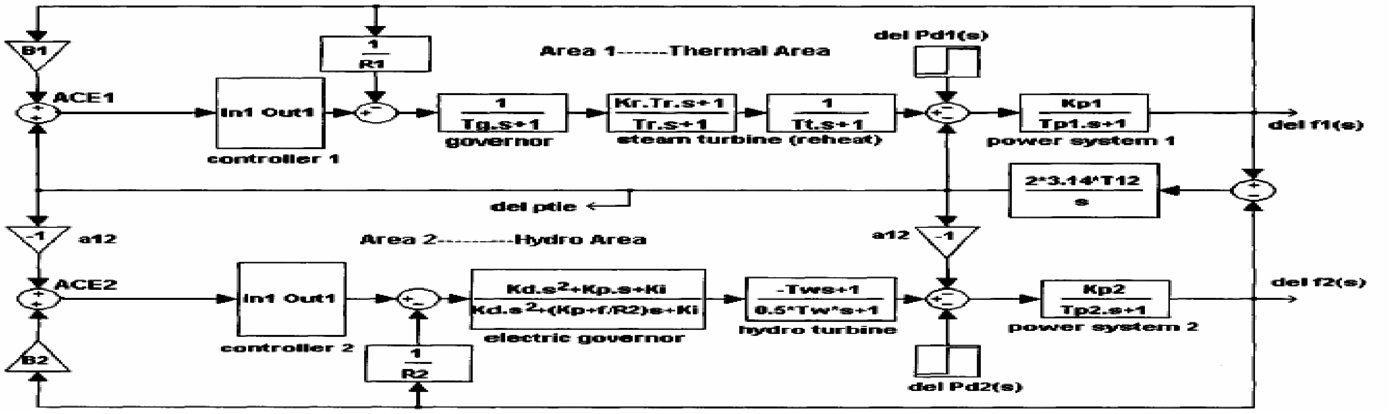


**Fig. 1**: Block diagram model of hydro-thermal reheat power system

The term 'invariant' means that the system is completely insensitive to parametric uncertainty and external disturbances [12-13].

## 2. The Investigated Power System

The detailed block diagram modeling of two area thermal-hydro power system for load frequency control investigated, is shown in figure 1. An extended power system can be divided into a number of load frequency control areas interconnected by means of tie-lines. Without loss of generality one can consider a two- area case connected by single tie- line [11].

The control objectives are as follows:

(i) Each control area as for as possible should supply its own load demand and power transfer through tie line should be on mutual agreement.

(ii) Both control areas should controllable to the frequency control.

In an isolated control area case the incremental power $(\Delta P_G - \Delta P_D)$ was accounted for by the rate of increase of stored kinetic energy and increase in area load caused by increase in frequency. Since a tie line transports power in or out of an area, this fact must be accounted for in the incremental power balance equation of each area.

### 2.1 Modeling of Tie-Line

The power transfer equation through tie- line is [11],

$$P_{12} = \frac{|V_1||V_2|}{x}\sin(\delta_1 - \delta_2) \qquad (1)$$

considering area 1 has surplus power and transfers to area 2

$P_{12}$ = Power transferred from area 1 to 2 through tie line.

$$P_{12} = \frac{|V_1|.|V_2|}{X_{12}}.\mathrm{Sin}(\delta 1 - \delta 2) \qquad (2)$$

Where

$\delta_1$ and $\delta_2$ = Power angles of end voltages $V_1$ and $V_2$ of equivalent machine of the two areas respectively.

$X_{12}$ = reactance of tie line.

The order of the subscripts indicates that the tie line power is define positive in direction 1 to 2.

For small deviation in the angles and the tie line power changes with the amount i.e. small deviation in $\delta_1$ and $\delta_2$ changes by $\Delta\delta 1$ and $\Delta\delta 2$,

Power $P_{12}$ changes to $P_{12} + \Delta P_{12}$

Therefore,

Power transferred from Area 1 to Area 2 as given in [11] is

$$\Delta P_{12}(s) = \frac{2\pi T^o}{s}\left(\Delta f_1(s) - \Delta f_2(s)\right) \qquad (3)$$

$T^0$ = Torque produced

The above equation can be represented as in Fig. 2

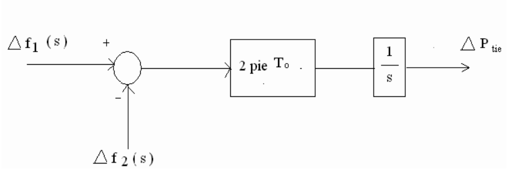


**Fig. 2** : Block Diagram Representation of a Tie -Line

Tie-line bias control is used to eliminate steady state error in frequency in tie-line power flow. This states that the each control area must contribute their share to frequency control in addition for taking care of their own net interchange.

Let ACE1 = area control error of area 1

ACE2 = Area control error of area 2

In this control, ACE1 and ACE2 are made linear combination of frequency and tie line power error[11].

ACE1 = $\Delta P_{12+} \mathrm{b1}\Delta \mathrm{f1}$ (4)

ACE2 = $\Delta P_{21+} \mathrm{b2}\Delta \mathrm{f2}$ (5)

where the constant b1 & b2 are called area frequency bias of area 1 and area 2 respectively.

Now $\Delta PR1$ and $\Delta PR2$ are mode integral of ACE1 and ACE2 respectively.

$$\Delta PR1 = -\mathrm{Ki}_1\int_0^t(\Delta \mathrm{P}_{12} + b_1\Delta f_1)dt \qquad (6)$$

$$\Delta PR2 = -\mathrm{Ki}_2\int_0^t(\Delta \mathrm{P}_{21} + b_2\Delta f_2)dt \qquad (7)$$

Taking Laplace transform of the above equation

We get

$$\Delta PR_{1(s)} = -\frac{\mathrm{Ki}_1}{\mathrm{s}}\left[\Delta P_{12}(s) + \mathrm{b}_1\Delta \mathrm{f}_1(\mathrm{s})\right] \qquad (8)$$

$$\Delta PR2_{(s)} = -\frac{\mathrm{Ki}_2}{\mathrm{s}}\left[\Delta P_{21}(s) + \mathrm{b}_2\Delta \mathrm{f}_2(\mathrm{s})\right] \qquad (9)$$

The step changes $\Delta P_{D1}$ and $\Delta P_{D2}$ are applied simultaneously in control area 1 and 2 respectively. When steady state conditions are reached, the output signals of all integrating blocks will be constant and their input signal must become zero. i.e.

$\Delta P_{12} + b_1 \Delta f_1 = 0$ (input of integrating block - $\frac{Ki_1}{s}$) (10)

$\Delta P_{21} + b_2 \Delta f_2 = 0$ ( input of integrating block - $\frac{Ki_2}{s}$) ( 11)

$\Delta f_1 - \Delta f_2 = 0$ (input of integrating block - $\frac{2\pi T_{12}}{s}$) ( 12)

$$\Delta P_{12} = \Delta P_{tie,1} \quad \text{and} \quad \Delta P_{21} = \Delta P_{tie,2}$$

Therefore $\frac{\Delta P_{tie,1}}{\Delta P_{tie,2}} = -\frac{T_{12}}{T_{21}} = -\frac{1}{a_2} = \text{constant}$ (13)

Hence $\Delta P_{tie,1} = \Delta P_{tie,2} = 0$

$$\Delta P R_1 = \Delta P R_2 ,$$

And $\Delta f_1 = \Delta f_2 = 0$

Thus, under steady condition change in the tie- line power and frequency of each area is zero. This has been achieved by integration of ACEs in the feedback loops of each area [11]. Control methodology used (PI, FLC & ANN) is mentioned in next preceding sections.

## 3. Conventional Integral Control

When an integral controller is added to each area of the uncontrolled plant in forward path the steady state error in the frequency becomes zero. The task of load frequency controller is to generate a control signal u that maintains system frequency and tie-line interchange power at predetermined values [11]. The block diagram of PI controller is shown in Fig. 3. Where

$$u_i = -K_i\int_0^{\tau} (ACE_i)dt = -K_i\int_0^{\tau} (\Delta P_{tie,i} + b_i\Delta f_i)dt \qquad (14)$$

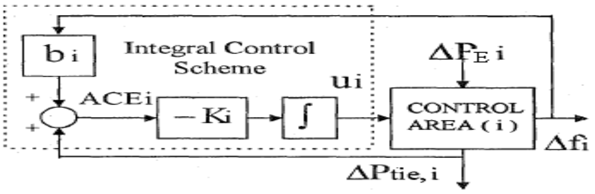


**Fig. 3** : Conventional PI Controller

Taking the derivative of above equation

$$\dot{u}_i = -K_i(ACE_i) = -K_i(\Delta P_{tie,i} + b_i\Delta f_i) \qquad (15)$$

## 4. Fuzzy Logic Control

Fuzzy logic is a thinking process or problem-solving control methodology incorporated in control system engineering, to control systems when inputs are either imprecise or the mathematical models are not present at all. Fuzzy logic can process a reasonable number of inputs but the system complexity increases with the increase in the number of inputs and outputs, therefore distributed processors would probably be easier to implement.

Fuzzification is process of making a crisp quantity into the fuzzy [14]. They carry considerable uncertainty. If the form of uncertainty happens to arise because of imprecision, ambiguity, or vagueness, then the variable is probably fuzzy and can be represented by a membership function. Defuzzification is the conversion of a fuzzy quantity to a crisp quantity, just as fuzzification is the conversion of a precise quantity to a fuzzy quantity [14]. The out put of a fuzzy process can be the logical union of two or more fuzzy membership functions defined on the universe of discourse of the output variables. There are many methods of defuzzification, out of which smallest of maximum method is applied in making fuzzy inference system [14].

SOM ( SMALLEST OF MAXIMUM) METHOD: This is also called first (or last of maximum) and this method uses the overall output or union of all individual output fuzzy sets $C_k$ to determine the smallest value of the domain with maximum z membership degree in $C_k$ [14]. The equation for $z^*$ are as follows:

$$z^* = \inf_{z\in Z}\{z \in Z \,|\, \mu C_k(z) = hgt(C_k)\} \qquad (16)$$

The Fuzzy logic control consists of three main stages, namely the fuzzification interface, the inference rules engine and the defuzzification interface [15]. For Load Frequency Control the process operator is assumed to respond to respond to variables error (*e*) and change of error (*ce*).

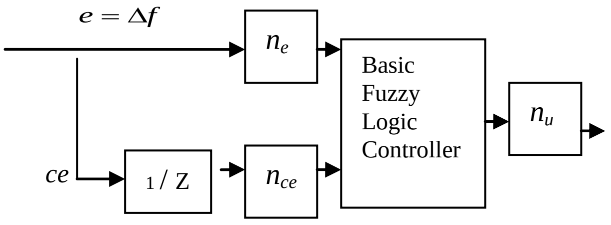


**Fig. 4** Block diagram of a Fuzzy Logic controller

The variable error is equal to the real power system frequency deviation ( $\Delta f$ ). The frequency deviation $\Delta f$ , is the difference between the nominal or scheduled power system frequency ($f_N$) and the real power system frequency (*f*). Taking the scaling gains into account, the global function of the FLC output signal can be written as.

$$\Delta P_C = F[n_c\, e(k), n_{ce}\, ce(k)] \qquad (17)$$

Where $n_e$ and $n_{ce}$ are the error and the change of error scaling gains, respectively, and F is a fuzzy nonlinear function. FLC is dependant to its inputs scaling gains [15]. The block diagram of FLC is shown in Fig 4,. $n_u$ is output control gain[6]. A label set corresponding to linguistic variables of the input control signals, *e(k) and* ce(k), with a sampling time of 0.01 sec is as follows:

L(*e*, *ce*) = { NB, NM, ZE, PM, PB},

Where, NB = Negative Big, NM = Negative Medium,
ZE = Zero, PM = Positive Medium, PB = Positive Big

Fuzzy logic controller has been used in hydro-thermal interconnected areas. Attempt has been made to examine with five number of triangular membership function (MFs) which provides better dynamic response with the range on input( error in frequency deviation and change in frequency deviation) i.e universe of discourse is -0.25 to 0.25. The numbers of rules are 25.

The membership functions (MFs) for the input variables are shown in Fig.5.

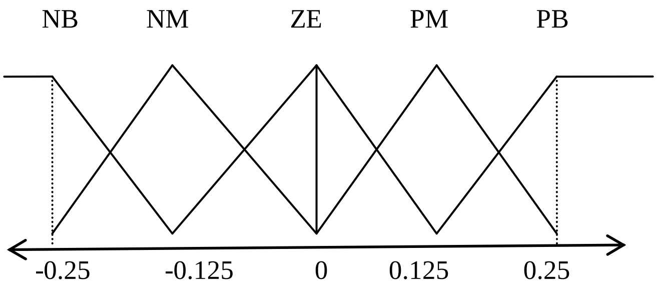


**Fig.5** : Membership Function for the control input variables

**Table. 1**: Fuzzy inference rule for Fuzzy Logic Controller

| Input | *e(k)* | | | | | |
|---|---|---|---|---|---|---|
| *ce(k)* | | NB | NM | ZE | PM | PB |
| | NB | NB | NB | NM | NM | ZE |
| | NM | NB | NB | NM | ZE | ZE |
| | ZE | NM | NM | ZE | PM | PM |
| | PM | ZE | PM | PM | PB | PB |
| | PB | ZE | ZE | PM | PB | PB |

## 5. Artificial Neural Network (ANN) Controller

ANN is information processing system, in this system the element called as neurons process the information. The signals are transmitted by means of connecting links. The links process an associated weight, which is multiplied along with the incoming signal (net input) for any typical neural net. The output signal is obtained by applying activations to the net input. The field of neural networks covers a very broad area. Neural network architecture-the multilayer perceptron as unknown function are shown in Fig 6, which is to be approximated. Parameters of the network are adjusted so that it produces the same response as the unknown function, if the same input is applied to both systems. The unknown function could also represent the inverse of a system being controlled, in this case the neural network can be used to implement the controller [17].

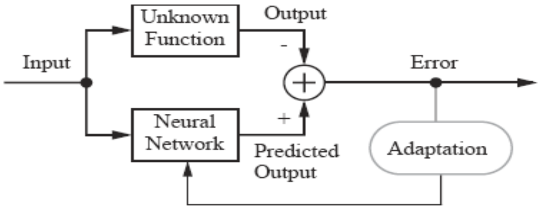


**Fig. 6**: Neural Network as Function Approximator

A neuron has more than one input. A neuron with inputs is shown in Fig. 7. The individual inputs $p_1, p_2, \ldots p_R$ are each weighted by corresponding elements $w_{1,1}, w_{1,2}, \ldots w_{1,R}$ of the weight matrix *W*. The neuron has a bias *b*, which is summed with the weighted inputs to form the net input n.

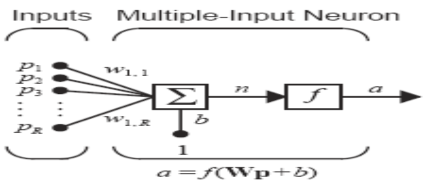


**Fig. 7** : Multi input neuron model

$$n = w_{1,1}p_1 + w_{1,2}p_2 + \cdots + w_{1,R}p_R + b\,. \tag{20}$$

This expression can be written in matrix form:

$$n = \mathbf{Wp} + b\,, \tag{21}$$

Where the matrix W for the single neuron case has only one row. Now the neuron output can be written as

$$a = f(\mathbf{Wp} + b)\,. \tag{22}$$

One of the most commonly used functions is the *log-sigmoid transfer function,* which is shown in Figure 8.

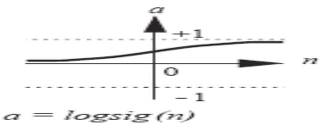


**Fig. 8**: Log-Sigmoid Transfer Function

This transfer function takes the input (which may have any value between plus and minus infinity) and squashes the output into the range 0 to 1, according to the expression:

$$a = \frac{1}{1 + e^{-n}}\,. \tag{23}$$

The log-sigmoid transfer function is commonly used in multilayer networks that are trained using the back propagation algorithm, in part because this function is differentiable.

### 5.1 NARMA-L2 Control

The ANN controller architecture employed here is a Non linear Auto Regressive Model reference Adoptive Controller. This controller requires the least computation of the three architectures. This controller is simply a rearrangement of the neural network plant model, which is trained offline, in batch form. It consists of reference, plant out put and control signal. The controller is adaptively trained to force the plant output to track a reference model output. The model network is used to predict the effect of controller changes on plant output, which allows the updating of controller parameters. In the study, the frequency deviations, tie-line power deviation and load perturbation of the area are chosen as the neural network controller inputs.

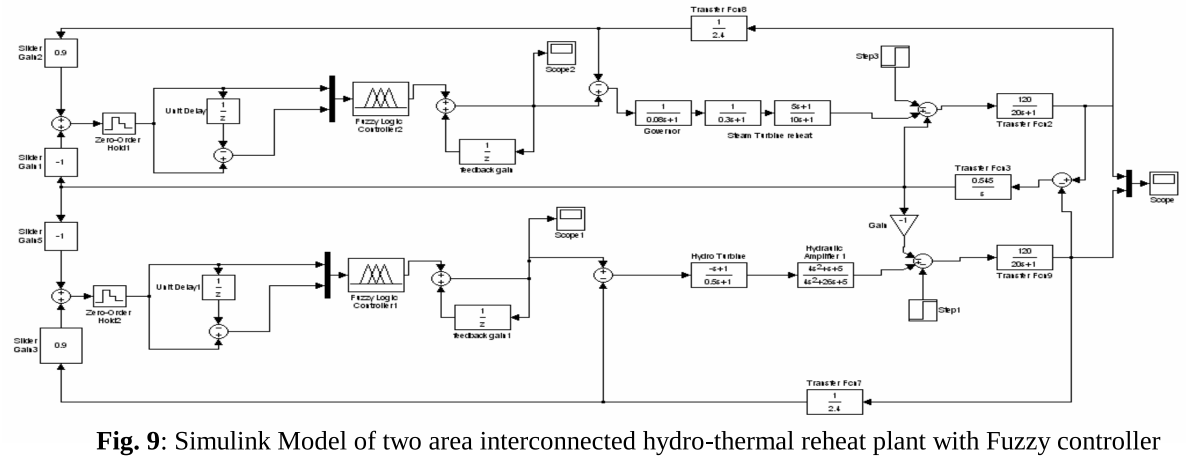


**Fig. 9**: Simulink Model of two area interconnected hydro-thermal reheat plant with Fuzzy controller

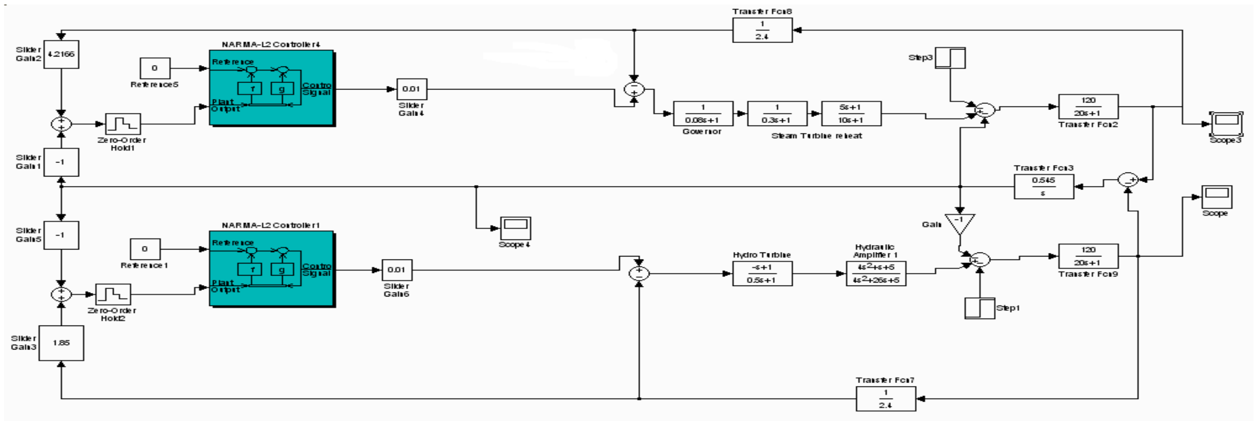


**Fig. 10**: Simulink Model of two area interconnected Hydro-Thermal Reheat plant with Neural Network controller

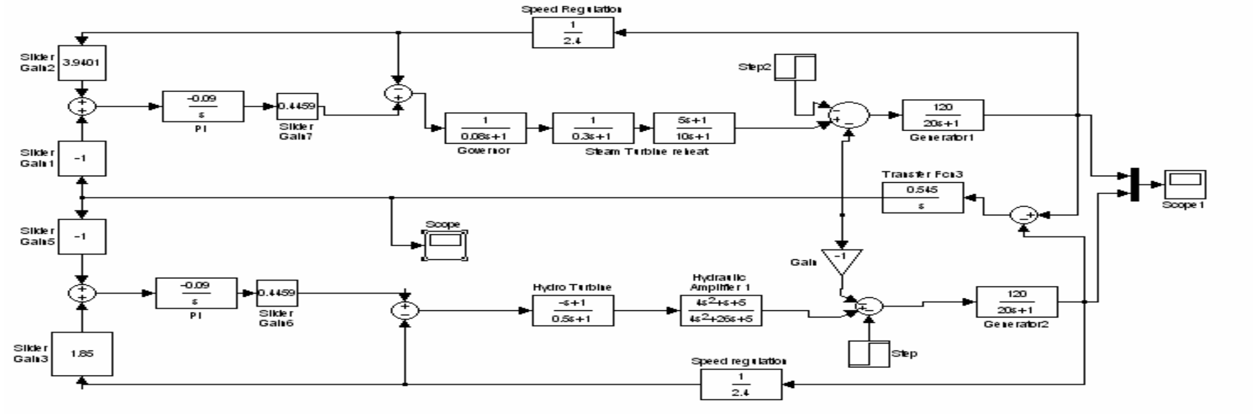


**Fig. 11**: Simulink Model of two area interconnected Hydro-Thermal Reheat plant with PI controller

The outputs of the neural network are the control signals, which are applied to the governors in the area. The data required for the ANN controller training is obtained from the designing the Reference Model Neural Network and applying to the power system with step response load disturbance. After a series of trial and error and modifications, the ANN architecture provides the best performance. It is a three-layer perceptron with five inputs, 13 neurons in the hidden layer, and one output in the ANN controller. Also, in the ANN Plant model, it is a three-layer perceptron with four inputs, 10 neurons in the hidden layer, and one output. The activation function of the networks neurons is trainlm function.300 training sample has been taken to train 300 no of epochs. The proposed network has been trained by using the learning performance. Learning algorithms causes the adjustment of the weights so that the controlled system gives the desired response.

## 6. Simulation and Results

In this presented work, the hydro-thermal interconnected power system have been developed with PI, fuzzy logic and ANN controllers to illustrate the performance of load frequency control using MATLAB/SIMULINK package. The parameters used for simulation are given in appendix [12]. Three types of simulink models are developed with fuzzy , ANN and PI controller based as shown in Fig. 9 -11 respectively to obtain better dynamic behavior. Refer to Simulink models, frequency deviation plots for thermal and hydro both cases are drawn combined and separately for 1% step load change in system frequency and tie-line power as shown in Fig. 12-20 respectively.

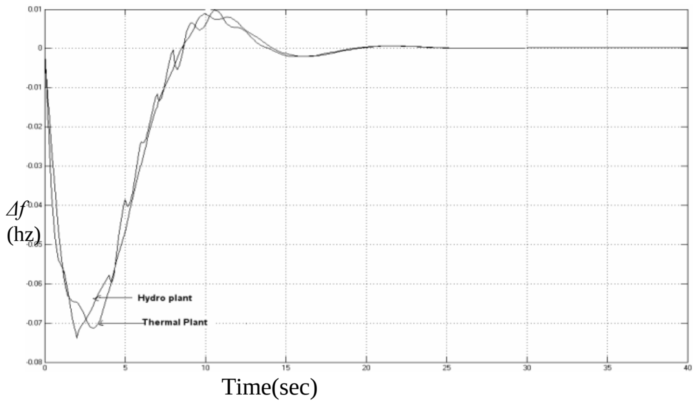


**Fig.12** Response of Hydo-thermal plant with fuzzy controller

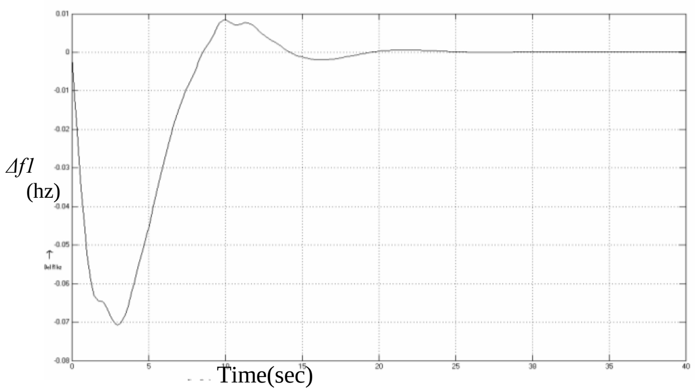


**Fig. 13** Change in frequency (thermal plant) –Fuzzy controller $(\Delta f_1)$

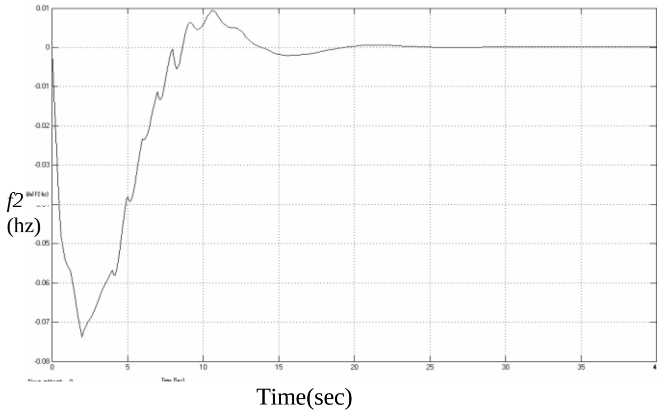


**Fig. 14** Change in frequency (Hydro plant)- Fuzzy controller $(\Delta f_2)$

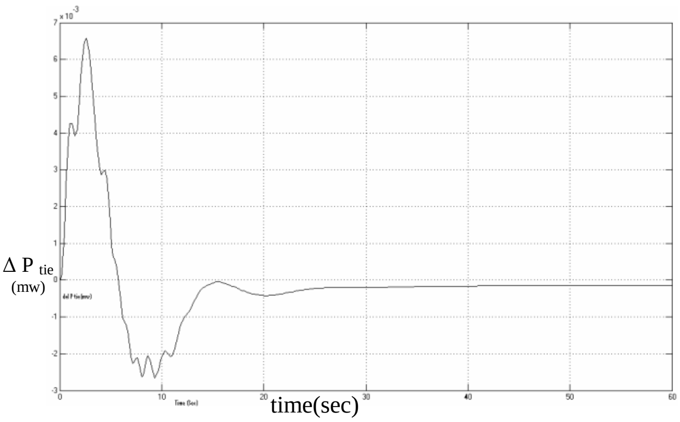


**Fig. 15** Change in Tie-line power (hydo-theraml plant) with fuzzy control $(\Delta P_{tie})$

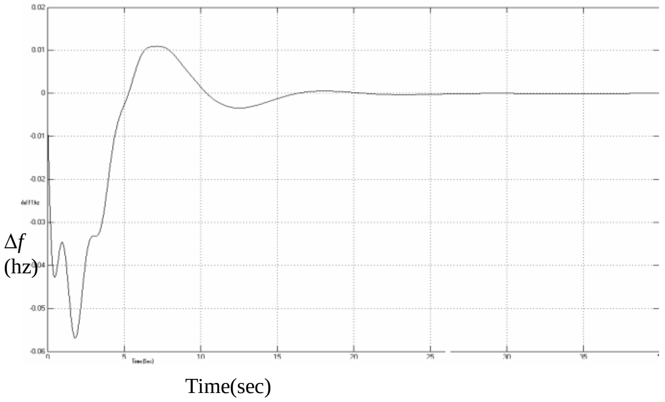


**Fig. 16** Change in frequency (Hydro plant) with ANN controller $(\Delta f_1)$

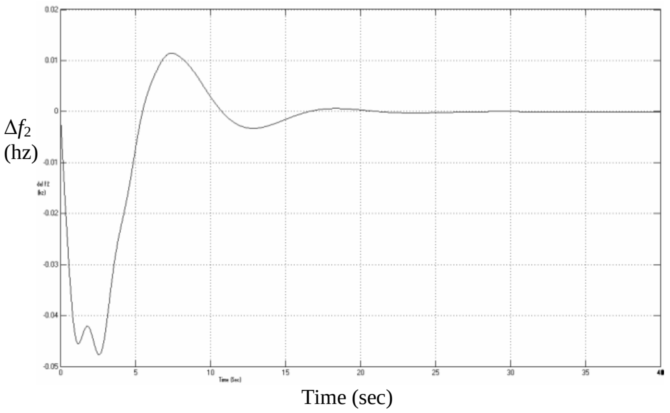


**Fig.17** Change in frequency (Thermal plant) with ANN controller $(\Delta f_2)$

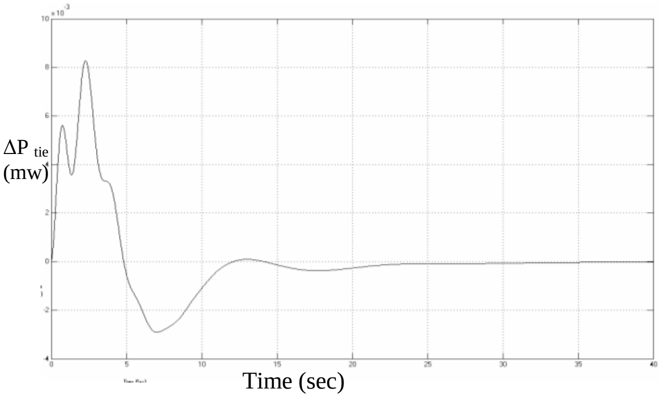


**Fig. 18** Change in Tie-line power (hydo-theraml plant) with ANN controller $(\Delta P_{tie})$

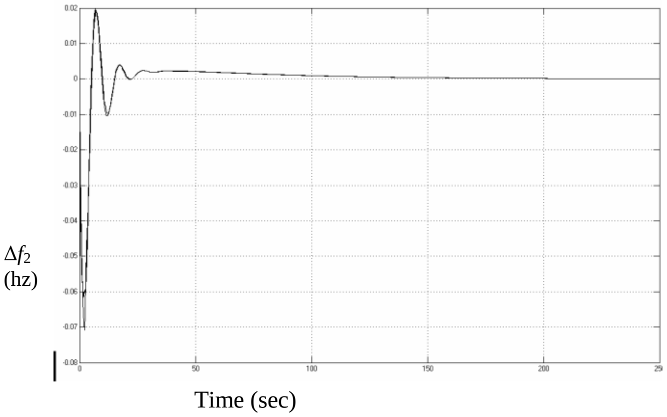


**Fig. 19**: Change in frequency (hydro-thermal plant) with PI Controller (Δ*f*)

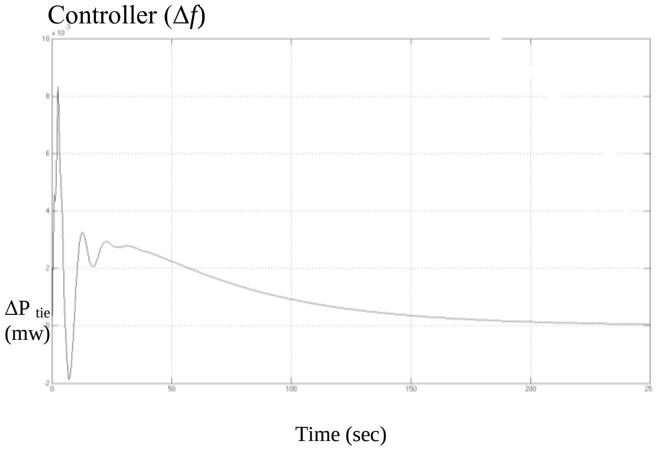


**Fig. 20** Change in Tie-line power (hydo-theraml plant) with PI controller (ΔP $_{tie}$)

With 1% step load change in both Hydro-Thermal Reheat areas with fuzzy, ANN and PI controllers, the steady state error is minimized to zero. Settling time and maximum peak overshoot in transient condition for both change in system frequency and change in tie-line power are given in table 2 & 3 respectively.

## 7. Conclusion

With 1% load variation in power system the following results are obtained. The Intelligent control approach (Fuzzy Controller and ANN controller) with inclusion of slider gain provides better dynamic performance and reduces the oscillation of the frequency deviation and the tie line power flow in each area in hydro-thermal combination.

**Table: 2**. Comparative study of settling time

| Controllers | Δ$f_1$ Area 1 (sec) | Δ$f_2$ Area 2 (sec) | ΔP $_{tie}$ (sec) |
|---|---|---|---|
| PI | 105 | 105 | 202 |
| Fuzzy | 23 | 23 | 27 |
| ANN | 17 | 17 | 23 |

**Table: 3**. Comparative study of peak overshoot

| Controllers | Δ$f_1$ Area 1 in pu | Δ$f_2$ Area 2 in pu | ΔP $_{tie}$ in pu |
|---|---|---|---|
| PI | 0.072 | 0.072 | 0.0081 |
| Fuzzy | 0.071 | 0.075 | 0.0035 |
| ANN | 0.045 | 0.055 | 0.003 |

From the above table it is clear that responses obtained, reveals that ANN controller with sliding gain provides better settling performance than Fuzzy & PI. Therefore, the intelligent control approach using ANN & fuzzy concept is more accurate and faster than the conventional PI control scheme even for complex dynamical system.

## 8. References

[1] George Gross and Jeong Woo Lee, "Analysis of Load Frequency Control Performance Assessment Criteria", IEEE transaction on Power System Vol. 16 No, 3 Aug 2001

[2] D. P. Kothari , Nagrath " Modern Power System Analysis"; Tata Mc Gro Hill, Third Edition.

[3] Kundur P, " Power System Stability and Control", Mc Graw hill New York, 1994.

[4] CL Wadhawa " Electric Power System" New Age International Pub. Edition 2007.

[5] Elgerd O. I. , " Elctric Energy System Theory; An Introduction " Mc Gro Hill.

[6] Jawad Talaq and Fadel Al- Basri , " Adaptive Fuzzy gain scheduling for Load Frequency Control", IEEE Transaction on Power System, Vol. 14. No. 1. Feb 1999.

[7] P. Aravindan and M.Y. Sanavullah, "Fuzzy Logic Based Automatic Load Frequency Control of Two Area Power System With GRC" International Journal of Computational Intelligence Research, Volume 5, Number 1 (2009), pp. 37–44

[8] J. Nanda, J.S Kakkarum, "Automatic Generation Control with Fuzzy logic controllers considering generation constraints", in Proceeding og 6th Int Conf on Advances in Power System Control Operation and managements" Hong Kong , Nov, 2003.

[9] A. Magla , J Nanda, " Automatic Generation Control of an Interconnected hydro- Thermal System Using Conventional Integral and Fuzzy logic Control", in Proc. IEEE Electric Utility Deregulation, Restructuring and Power Technologies, Apr 2004.

[10] A. Demiroren, H.L. Zeynelgil, N.S. SengorThe pplication of ANN Technique toLoad-frequency ControlFor Three-rea Power System" Paper acceptcd for presentation at PPT *001*,2001 IEEE Porto Power Tech ConferenceIOth - 13'h Septcrnber, Porto, Portugal.

[11] Surya Prakash , SK Sinha, "Impact of slider gain on Load Frequency Control using Fuzzy Logic Controller" ARPN journal of Engineering and Applied Science, Vol 4, No 7, Sep 2009.

[12] John Y, Hung, " Variable Structure Control: A Survey", IEEE Transaction on Industrial Electronics, Vol. 40, No.1 Feb 1993.

[13] Ashok Kumar, O.P. Malik., G.S. Hope. “Variable-structure-system control applied to AGC of an interconnected power System” I.E.E.E. *1EE PROCEEDINGS, Vol. 132, Pt. C, No. 1, JANUARY 1985*

[14] Timothy. J. Ross ; ‘Fuzzy logic with engineering application’, Mc Gro Hill, International Edition 1995

[15] Q. P. Ha, “ A Fuzzy sliding mode controller for Power System Load Frequency Control”,1998 Second International Conference of Knowledge based Intelligent Electronic System, 21-23 Apr 1998, Adelaide, Australia,

[16] M. Masiala,, M. Ghnbi and A. Kaddouri “An Adaptive Fuzzy Controller Gain Scheduling For Power System Load-Frequency Control” 2004 lEEE International Conference on Industrial Technology (ICIT).

[17] S Hykin “ Neural Network ‘ Mac Miller NY 1994.

[18] J N Mines’ MATLAB Supliment to Fuzzy & Neural approach in Engineering’ John Wiley NY 1997.

## Appendix

### Parameters

Parameters are as follows:

*f* = 50 Hz, R1 =R2= 2.4 Hz/ per unit MW, Tg = 0.08 sec, Tp=20 sec

P tie, max = 200 MW

Tr = 10 sec kr = 0.5,

H1 =H2 = 5 sec Pr1 = Pr2 =2000MW

Tt = 0.3 sec Kp1=Kp2 = 120 Hz.p.u/MW

Kd =4.0 ki = 5.0 Tw = 1.0 sec

D1 =D2= 8.33 * $10^{-3}$ p.u MW/Hz.

*F* : Nominal system frequency

$P_{ri}$ : Area rated power, $H_i$: Inertia constant

### Nomenclature

$K_j$ : Integral gain

$K_d$, $K_p$, $K_i$ : Electric governor derivative, proportional and integral gains, respectively.

$T_w$ : Water starting time, *ACE* : Area control error

*P*: Power, *E* :Generated voltage

*V*: Terminal voltage, $\delta$ : Angle of the Voltage V

$T_t$ : Steam turbine time constant

$R_i$ : Governor speed regulation parameter

$B_i$: Frequency bias constant

$T_{pi}$ : 2Hi / f * Di, $K_{pi}$: 1/ Di

$T_{12}$ : Synchronizing coefficient,

$T_g$: Steam governor time constant

$K_r$ : Reheat constant, $T_r$ : Reheat time constant

$\Delta\delta$ : Change in angle, $\Delta P$ :Change in power

$\Delta f$ : Change in supply frequency

$$Di = \frac{\Delta P_{Di}}{\Delta fi}$$

*R*: Speed regulation of the governor

$K_H$ : Gain of speed governor

$T_H$ :Time constant of speed governor

$K_1$, $K_2$, $K_3$, $K_4$, $K_5$ : Constants

$K_p$: 1/B= Power system gain

$T_p$: 2H / B $f_0$ = Power system time constant

$\Delta P_{Di}$ : Incremental load change

$\Delta Pg_i$ : Incremental generation change

## Author Biographies

**Surya Prakash** Allahabad, 01.05.1971, Received his Bachelor of Engineering degree from The Institution of Engineers(India) in 2003, He obtained his M.Tech. in Electrical Engg. .(Power System) from KNIT, Sultanpur.UP-India in 2009. Presently he is Pursuing Ph. D in Electrical Engg. Load Frequency Control and working as Assistant Professor in SSET, SHIATS(Formerly Allahabad Agriculture Institute, Allahabad- India). His field of interest is Intelligent Control & Power System operation and Control.

e-mail: sprakashgiri0571@yahoo.com *MB 09956722055*
sprakashgiri0571@gmail.com

**Dr. S. K. Sinha** belongs to Varanasi and his date of birth is 17th Apr 1962. He received the B.Sc. Engg degree in Electrical from R.I.T. Jamshedpur, Jharkhand, India, in 1984, the M. Tech. degree in Electrical Engineering from Institute of Technology, B.H.U, Varanasi, India in 1987, and the Ph.D. degree from IIT, Roorkee, India, in 1997. Currently he is working as Professor & Head, Department of Electrical Engineering, Kamla Nehru Institute of Technology, Sultanpur, UP, India . His fields of interest includes estimation, fuzzy control, robotics, and AI applications.

e-mail: *sinhask98@engineer.com* *MB*

# In Search of a Suitable Indian Language for Huffman Data Compression Algorithm

Satyendra Nath Mandal[1], Md. Iqbal Quraishi[2], Kuntal Bhowmick[3] and J. Pal Chaudhuri[4]

[1]Lecturer, Dept. of IT, Kalyani Government Engineering College, West Bengal, India.
[2]Lecturer, Dept. of IT, Kalyani Government Engineering College, West Bengal, India.
[3]4th Year, Dept. of CSE, Kalyani Government Engineering College, West Bengal, India.
[4]Asst. Professor, Dept. of IT, Kalyani Government Engineering College, West Bengal, India.

Kalyani Govt. Government Engineering College, Kalyani, Nadia, 741235, West Bengal, India
{satyen_kgec@rediffmail.com, iqbalqu@gmail.com, kuntal.kgec.cse@gmail.com jnpc193@yahoo.com}

*Abstract*: Huffman data compression algorithm is used many data compression application. In this paper, this algorithm has been used on data files of same size made by different languages. The same efforts have been made on different size files. The languages have been taking from the different part of India. A comparison has been made based on compression ratio, compression time and decompression time. Finally, one language has been selected based on performance.

***Keywords***: Huffman Data Compression Algorithm, Lossless data Compression, Compression Ratio, Compression Time and Decompression Time.

## 1. Introduction

In last decade has witnessed tremendous growth in the Information Technology innovations and applications. Information Technology has become a vital component for the success of business because most of the organizations require fast information dissemination, information processing, storage and retrieval of data. The growth in this area occurred at such a fast rate due to the fact that Information Technology [1][ 2] opened new vistas in almost all day-to-day problems related with common man. Information Technology has revolutionized our life and has made a significant impact on all dimensions of our day-to-day life. In banking sector, use of credit, debit card, ATM, Tele-banking, Net banking[6]; in transportation, reservation of air tickets[5], railway tickets, buying & selling items on internet, electronic market, inquiry of department, bank transaction on net, entertainment, education, communication, hotel reservation[3], tourism have become reality. Internet is one of the mediums, which being used to access the pool of information.

Proper transformation of data is main theme in this era. Sometimes information becomes so large that it becomes problematic to transmit or storing information in their proper format. So, the concepts of data compression arise. That means, transformation of information in certain format which will take much small space comparing to original data. Shannon-Fanon algorithm [7], Huffman algorithm [4] & Arithmetic Coding [8] are some process of data compression. The works include "Optimal Huffman Tree-Height Reduction for Instruction Level Parallelism", Dept. of Computer Sciences, the University of Texas at Austin reports on a work of exploiting instruction level parallelism(ILP) is a key component of high performance for modern processor. For this purpose, Huffman Algorithm was taken to (1) tree height reduction rewriting expression trees of commutative and associative operations to make the height of the tree reduced (2) software fan-out generating software to fan out tree even when expression store intermediates of the instruction. [8]

Another work on the concurrent update and generation of the dynamic Huffman Code on is made for the dynamic Huffman Encoding. The concurrent procedure performs the tree update and code generation processes in parallel and therefore reduces over 45% number of steps required by the Knuth's work [13].Work on the Fast Adaptive Huffman Encoding Algorithms state that Huffman code suffers from two problems: the prior knowledge of the probability distribution of the data source to be encoded is necessary, and the encoded data propagate errors. The first problem can be solved by an adaptive coding, while the second problem can be party solved by segmenting data into segments. But the adaptive Huffman code performs badly when segmenting data into relatively small segments because of its relatively slow adaptability [7]

The paper on "Optimal Multiple Bit-Huffman Decoding" proposes a new optimal multi-bit Huffman decoding method that combines the barrel shifter and look-ahead approaches. Specifically, the work is on the development of approval partition of the state diagram corresponding Huffman diagram [14].The development of an efficient compression scheme to process the Unicode format data represents a very difficult task. Our work mainly concentrates on this.

In this paper, the design of an efficient scheme of compression using the Huffman Coding to process multiple languages that supports Unicode formatting. Huffman algorithm coding on data compression with variable length bit coding has been used on files of same size made by 10 different languages, Arabic, Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Oriya, Tamil and Telugu. Same things have been made on files of different size from different languages. A comparison has been made on compression ratio, time taken for compression and time taken for decompression for files of same size on different languages. Finally, one language has been chosen based on their performance. This type of work has not been made so far. This is the reason for making this paper.

The paper is divided into following parts. The first part of this paper has been described the overall data compression. The Huffman data compression algorithm with example has been described in next section i.e. article number 3. The article number 4 has been given the algorithm for use of the Huffman Data Compression algorithm in different language. Finally, the results, conclusion and references have been described in article number 5, 6 and 7.

## 2. Overview of Data Compression

Data-compression techniques can be divided into two major families; **lossy** and **lossless**.

### 2.1 Lossy Data Compression Technique

Lossy data compression concedes a certain loss of accuracy in exchange for greatly increased compression. Lossy compression proves effective when applied to graphics images and digitized voice. Most lossy compression techniques can be adjusted to different quality levels, gaining higher accuracy in exchange for less

### 2.2 Lossless Data Compression

Lossless compression consists of those techniques guaranteed to generate an exact duplicate of the input data stream after a compress/expand cycle. This is the type of compression used when storing database records, spreadsheets, or word processing files.

### 2.3 Compression Ratio

Compression Ratio (CR) is defined as,

$$CR = \frac{(\text{Size of Original Data - Size of Compressed Data})*100}{\text{Size of Original Data}}$$

## 3. Huffman Algorithm

The Huffman Algorithm for Text Compression is an improvement over Shannon-Fano algorithm.

### 3.1 Huffman Algorithm Overview

Although similar in approach, the Huffman algorithm differs from Shannon-Fano algorithm in the construction of Binary Tree. The Huffman algorithm generates variable length code in such way that high frequency symbols are represented with a minimum number of bits and low frequency symbols are represented by relatively higher number of bits. The decoding of Huffman codes is done by using Huffman decodes Tree. Major difference in Shannon-Fano and Huffman algorithm, is that in SF algorithm ,the tree is built on the Top-Down approach while Huffman Tree is built using the Bottom-Up approach.

### 3.2 Construction of Huffman Tree and Generating code

1. Pick up two symbols (a, b) from the last two symbols in the sorted list of symbols.
2. Create two free nodes of the binary tree and assign A and B to these nodes.
3. Create a parent node for both nodes and assign it the frequency equal to the sum of frequencies of the child nodes.
4. Delete these two nodes from the list.
5. Parent node is added to the list of free nodes.
6. Repeat steps 1 to 5 until list of symbols becomes empty. This will generate the Huffman Tree.
7. Assign the bits similar to Shannon-Fano Tree i.e. left child is assigned '0' and right child is assigned '1'. Traverse from the root node of Huffman Tree to the leaf containing a particular symbol. This traversal will generate a code for that symbol.

Let us understand Huffman Algorithm with an example.

Table 1. Frequency table

| Symbol | Frequency |
|---|---|
| 'e' | 60 |
| 'a' | 20 |
| 'b' | 15 |
| 'd' | 5 |

The Huffman Tree is constructed in the following way and code is generated.

I. Find the sorted list of symbols in decreasing order of frequency. The list will be (e, a, b, d).
II. Pick up two Symbols with minimum frequency. These symbols are 'b' and'd'. Assign them two free nodes as shown figure1.

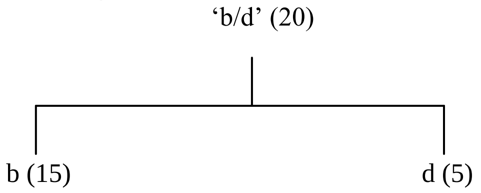


Figure 1. First Human Tree

This sub tree is called 'b/d'. Combined frequency of 'b/d' is 20.

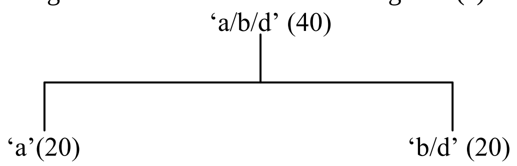
III. Update the sorted list as shown table 2.

Table 2. New Frequency Table

| Symbol | Frequency |
|---|---|
| 'e' | 60 |
| 'a' | 20 |
| 'b/d' | 20 |

Pick up two symbols with the minimum frequency. These symbols are ('a','b/d').

Assign them free node as under in figure 2(a) and 2(b).



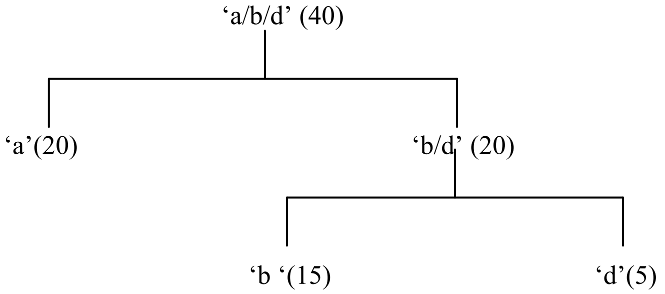


Figure 2(a) & 2(b): Next Huffman Tree

III. Update sorted list as shown in table 3.

Table 3. Modified frequency table

| Symbol | Frequency |
|---|---|
| ‘e’ | 60 |
| ‘a/b/d’ | 40 |

Pick up two symbols with the minimum frequency. These symbols are (’e’,’a/b/d’). Assign them free node as under is shown in figure 3.

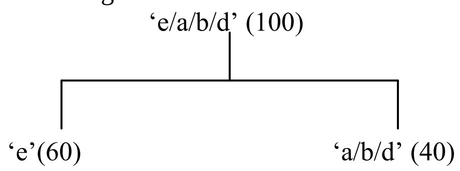


Figure 3. Third Huffman Tree

Finally, Huffman tree has been constructed based previous tree is shown figure 4. The two new symbols are nothing “b” and “d”.

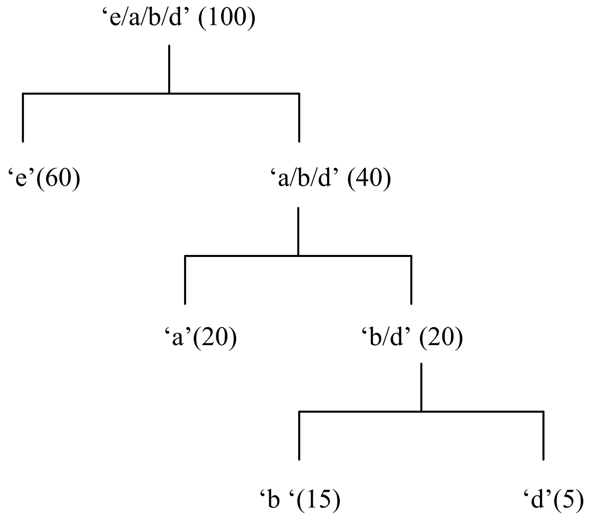
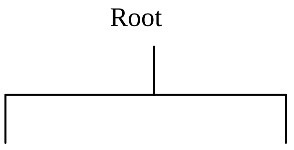


Figure 4. Final Huffman Tree

Update sorted list now it will contain only ‘e/a/b/d’ and no individual symbol is left. Therefore, the process halts and final Huffman Tree has been constructed. Now, assign the bit ‘0’ and ‘1’ to left and right subtree to obtain Huffman Code is shown in figure 5.



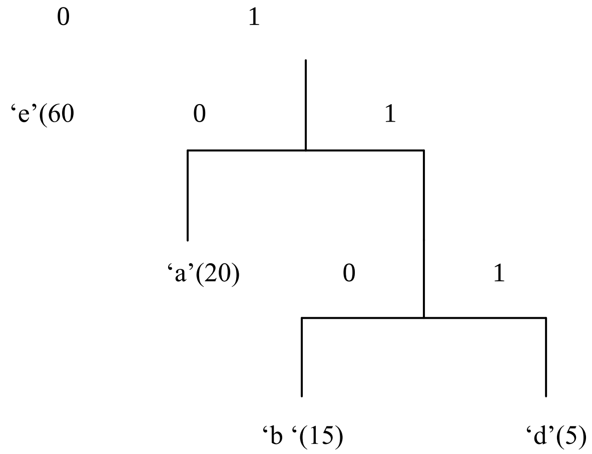


Figure 5. Huffman Tree with assigned code

From the above Huffman tree the code for all symbols are obtained. These codes are shown below in the following table 4.

Table 4. Symbol, Frequency, Huffman code and its size

| Symbol | Frequency | Huffman Code | Size of Huffman Code |
|---|---|---|---|
| E | 60 | 0 | 1 |
| A | 20 | 10 | 2 |
| B | 15 | 110 | 3 |
| D | 5 | 111 | 3 |

## 4. Method of selecting the language

This experimental process of selecting the best language has been done by taking ten standard Indian languages is shown figure 6. These files are served as input to the compressing algorithm.

The series of steps that we follow are given below.

**Step 1:** Initially the Unicode characters are written in Microsoft Word and then copy pasted in Notepad. The data is then saved as text file with the encoding option set to “Unicode” format.

**Step 2:** Ten files are prepared in the same way by copying and pasting the characters after being writing them in Microsoft word. The sizes of the files are increased gradually from 1Kb (1024 Bytes) to 10 Kb (10240 Bytes).

**Step 3:** In this way, 100 files are prepared, ten files of incrementing size for each of the ten languages.

**Step 4:** The Huffman data compression algorithm generates compressed files. The algorithm also generates a log file gives us information about the Original File Size, Compressed File Size, Time Required to Compression.

**Step 5:** The compressed files are then decompressed by the same algorithm. The decompressed file is compared with the original file to check it losslessness. The decompressed time also gets stored in the log file generated during compressing.

**Step 6:** A comparative study is made on the compression rate, Compressing and decompressing time from which the best language in selected.

The Figure 6 shows the above process diagrammatically.

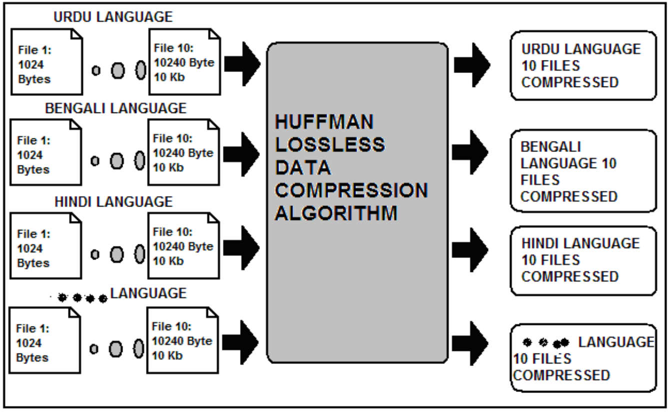

URDU LANGUAGE
File 1: 1024 Bytes
File 10: 10240 Byte 10 Kb
BENGALI LANGUAGE
File 1: 1024 Bytes
File 10: 10240 Byte 10 Kb
HINDI LANGUAGE
File 1: 1024 Bytes
File 10: 10240 Byte 10 Kb
LANGUAGE
File 1: 1024 Bytes
File 10: 10240 Byte 10 Kb
HUFFMAN LOSSLESS DATA COMPRESSION ALGORITHM
URDU LANGUAGE 10 FILES COMPRESSED
BENGALI LANGUAGE 10 FILES COMPRESSED
HINDI LANGUAGE 10 FILES COMPRESSED
LANGUAGE 10 FILES COMPRESSED

Figure 6: Diagrammatic Representation of the Language selection process

## 5. Experimental Results

For selecting the suitable language, an effort has been made a comparative study of ten languages already mentioned on compression ratio, compression time and decompression time. The selection of language based on the minimum compression ratio, minimum compression and decompression time are depicted in table 2,3 and 4. The results are also furnished in figure 7, 8 and 9.

## 6. Conclusion and Future Work

The result describes in tables 5, 6 & 7 shown that most of the times, the compression ratio, compression and decompression time of Arabic language is comparatively better than other nine languages. So, from this result, it can be concluded that Arabic language is suitable for Huffman data compression algorithms. Same study will be made in future on other languages in India.

## 7. References

[1] Blelloch, E., 2002, ***Introduction to Data Compression***. Computer science Department, Carnegie Mellon University.

[2] Cormark, V. and s. Horspool,, 1987, Data Compression using Dynamic Huffman Coding and Modelling Compute. J., 30: 541-550

[3] Vo Ngoc and M. Alistair , 2006, ***Improved word-aligned Binary compression for text indexing,*** IEEE Trans , Knowledge & Data Engineering, 18:857-861

[4] Kaufman, K. and T. Shumuel, 2005, Semi-Lossless text Compression, Intl. J Foundations of Computer Science.,16:1167-1178

[5] Capocelli, M., R. Giancarlo and J. Taneja, 1986 Bounds on the redundancy of Huffman Codes IEEE Transmission Information Theory, 32:854-857

[6] Gawthrop, J. and W. Liuping, 2005. Data Compression for Estimation of the Physical parameters of estimation of the physical parameters of stable and unstable linear Systems. Automation, 41: 1313-1321

[7] Kesheng, W., J. Otoo and S. Arie, 2006.optimizing Bitmap Image Compression Techniques with specification Index.ACM Databases Systems, 31:1-38

[8] D.A Huffman, "A method for the construction of minimum- redundancy codes,"Proc, . IRE, Vol. 40, pp 1098-1101, Sept. 1952.

[9] R.G Gallager, "Variations on a theme by Huffman," IEEE Trans. Inform Theory ., Vol It-24, pp,668-674, Nov, 1978.

[10] H. Yokoo, "An Improvement of Dynamic Huffman Coding with a simple repetation finder", IEEE Trans., Commun., Vol 39,pp. 8-10,Jan 1991.

[11] B. Landwehr and P. Marwedel, "***A new optimization Technique for improve ment resurce exploitation and critical path minization***," in Symposium on System Synthesis, (Antwerp, Belgium),pp. 65-72,September 1997.

[12] R. Sethi and J. d. Ullman "Using High Performance with an optimization compiler " in Proceedings architectures of high speed Processor pp :185-195,

[13] Sameh Ghwanmeh,Riyad Al-Shalabi "Efficient Data Compression Scheme using Dynamic Huffman Code Applied on Arabic Language" Journal of Computer Science 2(12) 885-888,2006 ISSN 1549-3636.

[14] Mark Nelson and Jean Loup Gailly "The Data Compression Book" 2$^{nd}$ Edition BPB Publisher,Indian Edition 1996.

Table 5: Different file size for different Languages and Compression ratio

| **Tamil** | **Telugu** | **Assamese** | **Oriya** | **Malayalam** | **Gujarati** | **Bengali** | **Hindi** | **Kannada** | **Arabic** |
|---|---|---|---|---|---|---|---|---|---|
| 65.917969 | 59.765625 | 63.671875 | 63.867188 | **59.082031** | 59.179688 | 64.550781 | 63.281250 | 65.234375 | 65.234375 |
| 60.156250 | 67.578125 | 65.917969 | 59.765625 | **59.130859** | 62.695312 | 63.330078 | 62.353516 | 65.329675 | 64.306641 |
| 62.076823 | 59.147135 | 60.774740 | 66.210938 | 63.346354 | 59.505208 | 61.100260 | 59.147135 | 64.290365 | **57.584635** |
| 63.696289 | 60.742188 | 62.353516 | 63.916016 | 64.331055 | 63.916016 | 64.599609 | 66.235352 | 61.718750 | **59.155273** |
| 59.765625 | 63.691406 | 66.328125 | **57.597656** | 62.968750 | 62.382812 | 62.714844 | 63.691406 | 58.183594 | 61.738281 |
| 60.791016 | 66.324870 | 64.664714 | **59.163411** | 61.735026 | 64.843750 | 65.543620 | 59.505208 | 58.544922 | 61.100260 |
| 65.652902 | **57.338170** | 62.039621 | 67.619978 | 62.067522 | 63.909040 | 57.589286 | 64.313616 | 66.629464 | 62.974330 |
| 64.318848 | 61.437988 | 65.270996 | 62.695312 | 59.509277 | 68.298340 | 61.743164 | 58.789062 | 62.365723 | **57.592773** |
| **58.279080** | 62.369792 | 62.369792 | 61.436632 | 67.610677 | 60.134549 | 65.332031 | 64.854601 | 64.854601 | 65.277778 |
| 61.748047 | 65.957031 | 66.630859 | 64.677734 | 65.332031 | 65.546875 | **56.992188** | 65.957031 | 66.923828 | 60.156250 |

Table 6: Different file size for different Languages and Compression Time

| Tamil | Telugu | Assamese | Oriya | Malayalam | Gujarati | Bengali | Hindi | Kannada | Arabic |
|---|---|---|---|---|---|---|---|---|---|
| 0.059 | 0.0548 | 0.0578 | 0.0592 | 0.0678 | **0.0549** | **0.0549** | 0.0568 | 0.0598 | 0.0568 |
| 0.062 | **0.0447** | 0.0525 | 0.0589 | 0.0598 | 0.0587 | 0.0574 | 0.0589 | 0.0549 | 0.0645 |
| 0.078 | 0.0688 | 0.0714 | 0.0635 | 0.0646 | **0.0574** | 0.0789 | 0.0587 | 0.0587 | 0.0654 |
| 0.082 | 0.0625 | 0.0845 | 0.0789 | 0.0712 | 0.0789 | 0.1345 | 0.1102 | **0.0574** | 0.0742 |
| **0.0102** | 0.0845 | 0.0845 | 0.0848 | 0.0845 | 0.0846 | 0.1289 | 0.0846 | 0.0789 | 0.0845 |
| 0.0589 | 0.0915 | 0.0987 | 0.0978 | 0.0973 | 0.0942 | 0.1156 | 0.0942 | 0.0897 | **0.0847** |
| 0.1648 | 0.1041 | 0.1258 | 0.1096 | 0.1198 | 0.1156 | 0.1345 | 0.1156 | 0.1548 | **0.0987** |
| 0.1847 | 0.1185 | 0.1389 | 0.1289 | 0.1385 | 0.1274 | 0.1274 | 0.1274 | 0.1045 | **0.1147** |
| 0.2197 | 0.1385 | 0.1149 | 0.1347 | 0.1398 | 0.1345 | 0.1897 | 0.0587 | 0.1356 | **0.1137** |
| 0.2198 | 0.1572 | 0.1489 | 0.1597 | 0.1596 | 0.1548 | 0.1945 | 0.1548 | 0.1945 | **0.1398** |

Table 7: Different file size for different Languages and Decompression Time

| Tamil | Telugu | Assamese | Oriya | Malayalam | Gujarati | Bengali | Hindi | Kannada | Arabic |
|---|---|---|---|---|---|---|---|---|---|
| **0.042** | 0.0578 | 0.0547 | 0.0589 | 0.0547 | 0.0548 | 0.0548 | 0.0601 | 0.0573 | 0.0587 |
| **0.045** | 0.0741 | 0.0647 | 0.0612 | 0.0593 | 0.0596 | 0.0658 | 0.652 | 0.0548 | 0.0789 |
| 0.069 | 0.0712 | 0.0798 | 0.0698 | 0.0679 | 0.0658 | 0.0758 | 0.0649 | **0.0596** | 0.0765 |
| 0.072 | 0.0756 | 0.0971 | 0.0756 | 0.0798 | 0.0858 | 0.1456 | **0.0654** | 0.0658 | 0.0897 |
| 0.081 | 0.0813 | 0.0925 | 0.0849 | 0.0973 | **0.0695** | 0.1045 | 0.0698 | 0.078 | 0.0965 |
| 0.0145 | 0.0105 | 0.1045 | 0.0845 | 0.1047 | 0.0985 | 0.1025 | 0.0985 | 0.0968 | 0.0879 |
| 0.1458 | 0.01156 | 0.1354 | 0.1145 | 0.1245 | 0.1025 | 0.1456 | 0.1025 | 0.1647 | **0.1012** |
| 0.1305 | 0.1347 | 0.1478 | 0.1374 | 0.1246 | 0.1256 | 0.1256 | 0.1256 | 0.1156 | **0.1145** |
| 0.15698 | 0.1498 | 0.1378 | 0.1454 | 0.1496 | 0.1456 | 0.1875 | 0.1596 | 0.1687 | **0.1193** |
| 0.15674 | 0.1497 | 0.1689 | 0.1647 | 0.1698 | 0.1647 | 0.2014 | 0.1647 | 0.2014 | **0.1354** |

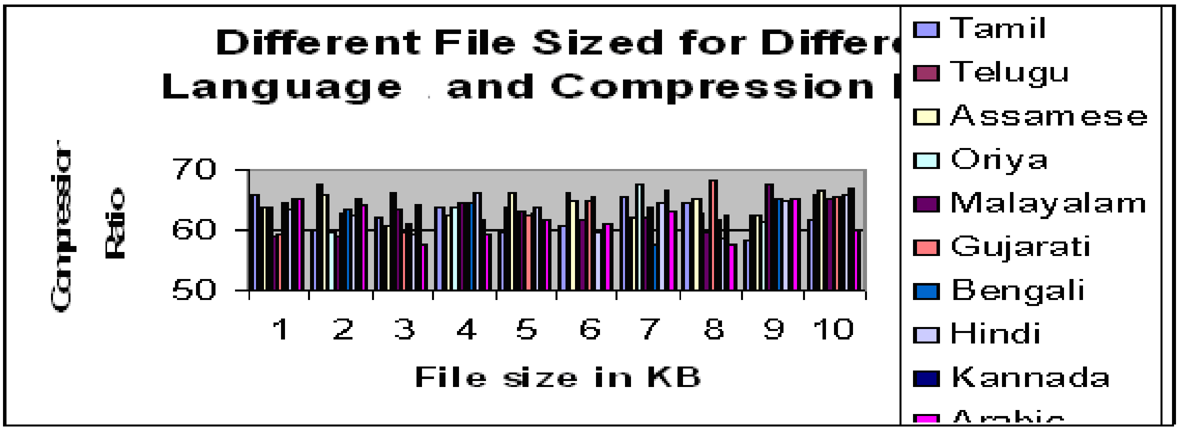


Figure 7. Graph for Different File sized for Different Languages and Compression Ratio

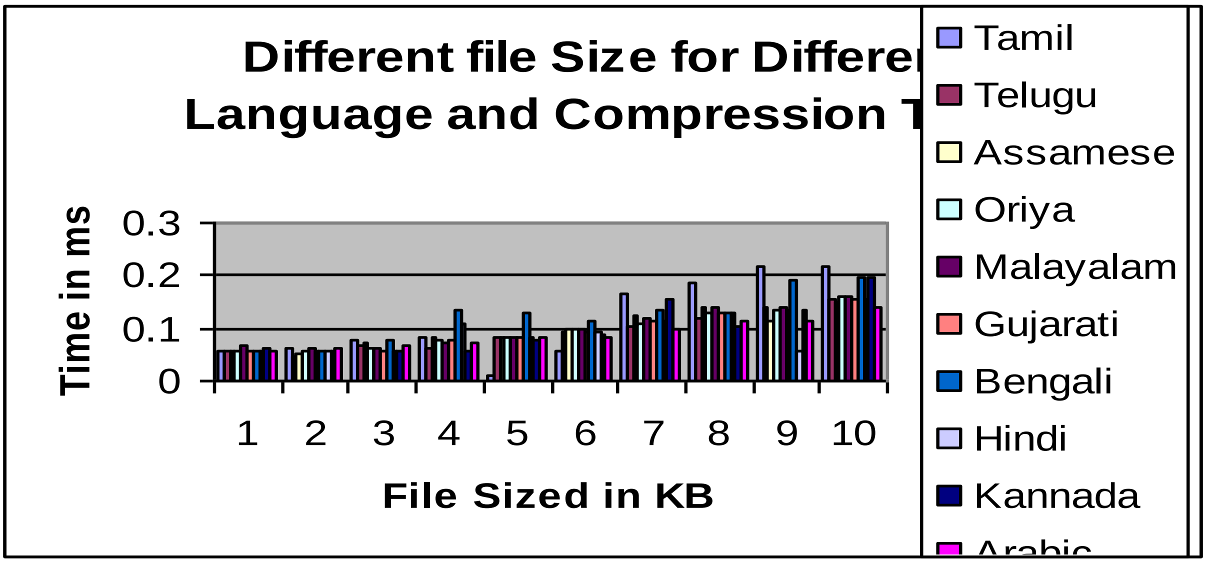


Figure 8: Graph for Different File sized for Different Languages and Compression Time

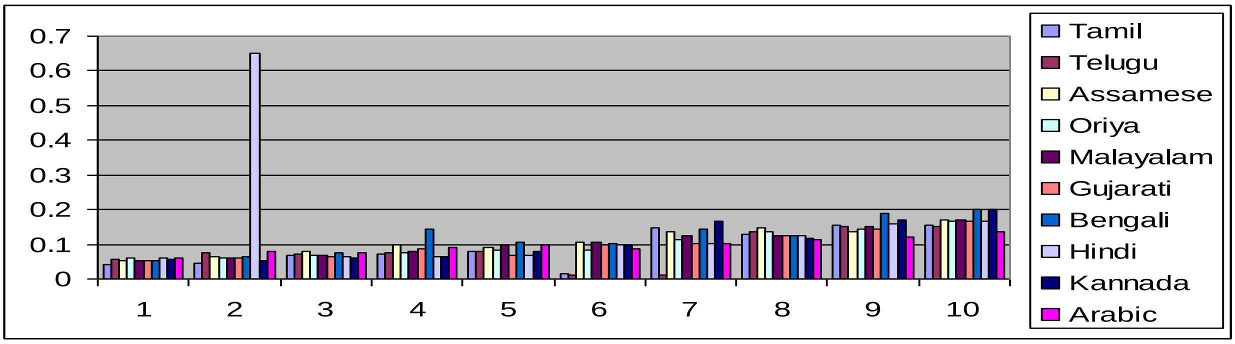


Figure 9: Graph for Different File sized for Different Languages and Decompression Time

# Handwritten English Character Recognition Using New Quadrant Splitting Feature Extraction Technique

Dayashankar Singh[1], Maitreyee Dutta[2], Sarvpal H. Singh[3], P. K. Singh[4]

[1]M.M.M.Engineering College, Gorakhpur,
[2]NITTTR, Chandigarh,
[3]M.M.M.Engineering College, Gorakhpur,
[4]M.M.M.Engineering College, Gorakhpur,

{dss_mec, d_maitreyee, singh_sarvpal}@yahoo.co.in, topksingh@gmail.com

***Abstract-*** In this paper, a new feature extraction technique based on divide and conquer approach has been developed. Each character divided in four quadrants is processed to search each quadrant in some predefined patterns. The values assigned to resultant patterns are combined and fed to the neural network as input for training the network. The technique named Quadrant Splitting Feature Extraction (QSFE) is implemented using Back-propagation Neural Network with one hidden layer and one output layer. An analysis has been carried out on Experimental results to show that the newly developed feature extraction technique QSFE requires less training time and provides high recognition accuracy of about 96%.

***Keywords:*** Neural Network, Feature Extraction Technique, Quadrant Splitting, Recognition Accuracy, Backpropagation neural network

## 1. Introduction

Neural Networks has been extensively used in the area of pattern recognition. Character recognition is one such aspect of pattern recognition that is being addressed by researches heavily. Many researches done in several languages such as Chinese, Japanese, Arabic, Farsi, etc. have been reported but still efficiency improvements in work related to hand-written English words and their conversions into Hindi using NN is an open problem. In many Indian offices such as Passport, Banks, Railway and sales tax etc., both Hindi and English languages are used [19]. This leaves a vast scope for English to Hindi translation.

Handwritten character recognition is a challenging problem in pattern recognition field. The difficulty is mainly because of large variations of individual's writing styles. Therefore, robust feature extraction to improve the performance of handwritten English character and word recognition has become very important to improve the performance of handwritten character recognition. In continuation to these efforts, a technique is being presented here that recognizes handwritten English characters and words. The work has also been enhanced to convert the words written in English into Hindi with high recognition accuracy and reduced training time. This technique is being referred to **Quadrant Splitting Feature Extraction (QSFE)** and is based on Back-propagation Neural Network with one hidden layer and one output layer.

Character recognition in several languages such as Chinese, Japanese etc has been an interesting issue and a good amount of work has been reported in this field. In India, about 80 percent people speak or know Hindi and a major part of this population also understand English. Therefore, considering the age of globalization and Information Technology, it is greatly felt that auto recognition of English words and their conversion into Hindi words are the needs of hour so that million of Hindi aware people around the world could get benefit with this revolutionized age of Information Technology. The methodology suggested in this paper has been proved to be the one recognizing English Character and words with high accuracy in less training and classification time.

The application of neural network to recognize characters is not new and the size of the input to such a net has remained a main; as the size of the input increases, the training time of network also increases. The QSFE minimizes the number of inputs to the neural network by dividing a character into four parts and then identifying the features of the character in each of the four parts. A complete character to the neural network is not needed to be given for recognition purpose; rather the values of the features of each quadrant are taken as inputs to the network. This requirement makes the feature extraction a key issue.

The organization of paper is as follow: Section-II describes the Extraction of features. Section-III provides detailed description of newly developed feature extraction technique named as Quadrant Splitting Feature Extraction (QSFE) technique. Section-IV offers experimental results and their analysis while Section-V covers conclusion and future scope.

## 2. Extractions of Features

Feature can be defined as a measurement taken on the input pattern to be classified. Features play an important role in handwritten character recognition in order to influence the recognition performance. When the input data to an algorithm is is too large to be processed and it is suspected to be notoriously redundant (much data, but not much information) then the input data will be transformed into a reduced representation set of features (also named features vector). Transforming the input data into the set of features is called features extraction.

The main objective of feature extraction is to divide the pattern by means of minimum number of features, which are effective in discriminating the different pattern classes.

Typically, it is being looked for the features that will provide a definite characteristic of that input type. The classifier is then supplied with a list of measured features, so that it maps these input features onto a classification state. On giving the input features, the classifier must decide which type of class or category they match most closely. Classification is rarely performed using a single measurement or feature taken from the input pattern. Usually, several measurements are required to be able to adequately distinguish between inputs that belong to different classes.

In this paper, various feature extraction techniques have been studied e.g. Conventional Feature Extraction, Boundary Tracing and Gradient Feature Extraction etc. A new feature extraction technique has been developed and it has been named as Quadrant Splitting Feature Extraction. The newly developed feature extraction technique (QSFE) has been implemented for English character recognition that provides high recognition accuracy and reduced training time.

## 3. Quadrant Splitting Feature Extraction (QSFE)

In this new approach, a 32x 32, black & white image in binary format is taken as input .The pixels which are covering the shape of the character are taken as the values 1 and rest of the pixels have the values 0. Now, the image is split into four quadrants each with size 16 x 16. Each character has its part in each of the four quadrants. Shape numbers are assigned to each part of the character which is lying in different quadrants.

Each quadrant has been separately scanned and matched the pixel pattern to different shapes. If all the four quadrant shapes are matching to a particular character, then the features of that character are extracted. In this way, 1024 input values are reduced to 4 values corresponding to the character part lying in each of the four quadrants.

For word recognition, the word is scanned column wise. The appearance of first empty column i.e. column having no character pixel, indicates the end of very first character. The number of empty columns between two consecutive characters will be the empty space between them. This procedure continues till the last column (up to column number 96).Characters scanned in this way are stored in separate arrays and they are made of the size 32 X 32.Thus by recognizing individual characters, the word made by their combination is recognized. Once the English word has been recognized, the process of conversion from English to Hindi starts. For word recognition, pattern matching technique has been performed. In this way, English words have been converted into Hindi words. Experiment has been carried out on these word conversions and it shows good result.

### 3.1. Features for English Characters

The features for characters of English Alphabet are given in Figure 1.

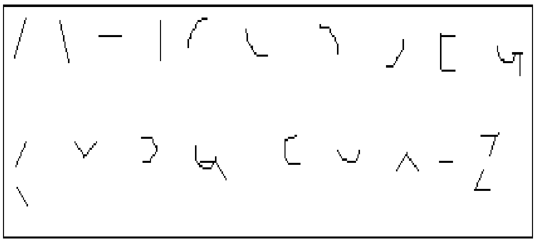


Figure1. Features of English Characters

### 3.2 Shape Representation

The various shape numbers which have been assigned for their corresponding features in various quadrants are given below.

**3.2.1 First quadrant:** Features and their corresponding values of shape numbers of first quadrant are given in figure 2.

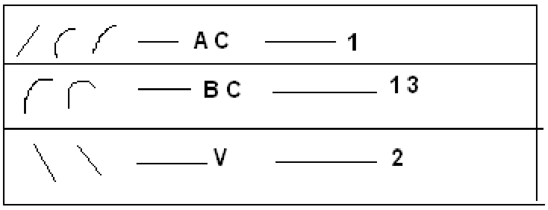


Figure 2. First Quadrant Features and their shape numbers

**3.2.2 Second Quadrant***:* Features and their corresponding values of shape numbers of second quadrant are given in figure 3.

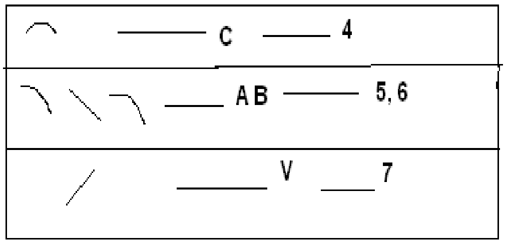


Figure 3. Second Quadrant Features and their shape numbers

**3.2.3Third Quadrant:** Features and their corresponding values of shape numbers of third quadrant are given in figure 4.

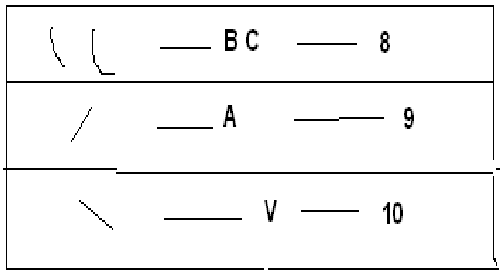


Figure 4. Third Quadrant Features and their shape numbers

**3.2.4 Fourth Quadrant:** Features and their corresponding values of shape numbers of fourth quadrant are given in figure 5.

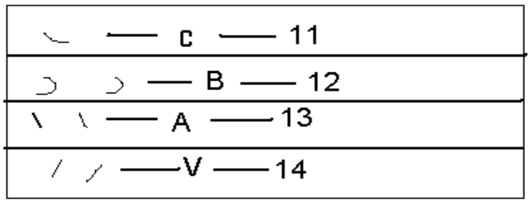


Figure 5. Fourth Quadrant Features and their shape numbers

### 3.4 Procedure

The procedure of implementing **QSFE** technique can be illustrated as follows:

- Capture the image of handwritten English character in 32 x32 pixels.
- Perform binarization on captured image into 32 x 32 pixels.
- Apply thinning process on binarized image.
- Split the image into four quadrants of 16x16 each.
- Scan each quadrant to match the pixel pattern of different shapes.
- All the four quadrant shapes are matching to a particular character and then the features of that character are extracted.
- In this way, 1024 input values have been reduced into 4 corresponding values to the character part lying in each of the four quadrants.
- Four corresponding values of a character are given as input to the Back propagation neural network for training as well as for simulation.

### 3.3 Flow Chart

The flowchart of automatic English character recognition system is given in figure 6.

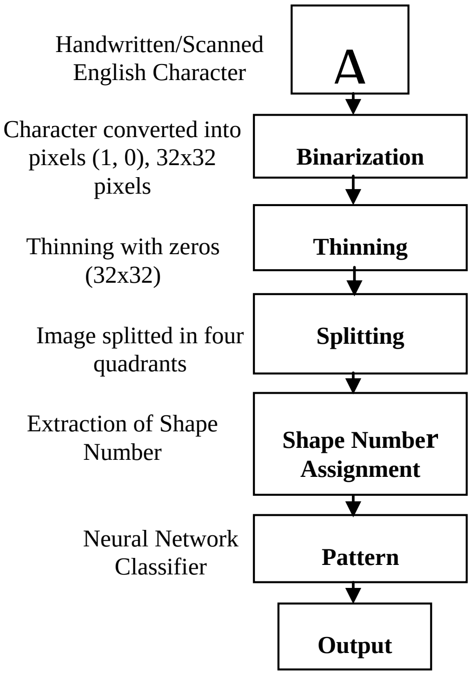

Handwritten/Scanned English Character
A
Character converted into pixels (1, 0), 32x32 pixels
Binarization
Thinning with zeros (32x32)
Thinning
Image splitted in four quadrants
Splitting
Extraction of Shape Number
Shape Number Assignment
Neural Network Classifier
Pattern
Output

Figure 6. Flowchart

***Binarization:*** A process that converts whole image into pixel values 0 and 1.

***Thinning:*** Through this operation skeleton of the image is obtained.

***Splitting:*** Skeletonized image is splitted into four quadrants.

***Shape Number Assignment:*** Different shape numbers have been assigned to the parts of the character which are lying in different quadrants. Pattern Recognition Pixel pattern is matched to find a particular shape number in each quadrant

## 4. Neural Network Classifiers

The neural network classification techniques such as multilayer perceptron (MLP) trained by Error backpropagation (EBP) algorithm has been used in this work. The feed-forward backpropagation network does not have feedback connections, but errors are backpropagated during training. Adjusting the two set of weights between the pair of layers and recalculating the output is an iterative process that is carried on until the error falls below a tolerance level. Learning rate parameters scale the adjustments to weights [13, 14]. The layered diagram of neural network is shown in Figure7.

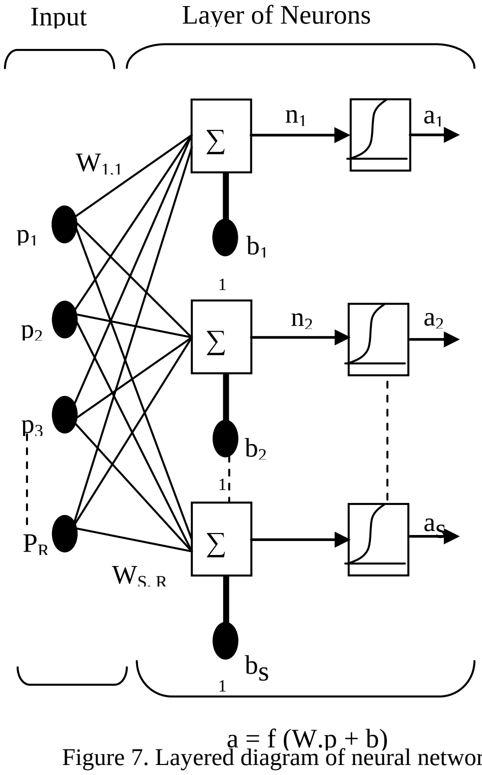


Figure 7. Layered diagram of neural network

In the layered diagram of neural network, $p_1, p_2 \ldots. p_R$ are inputs to the neural networks, $W_{11} \ldots W_{S,R}$ are the weights, $b_1, b_2 \ldots. b_S$ are biases given to the network initially 1 and $a_1$, $a_2 \ldots a_S$ are outputs of the neural network representing the number of inputs in the layer and S represents the number of neurons in the layer. At the hidden layer, tan sigmoid and at output layer, pure linear functions have been taken.

### 4.1 Backpropagation Training Algorithms

The simplest implementation of backpropagation learning updates the network weights and biases in the direction in which the performance function decreases most rapidly - the negative of the gradient. One iteration of this algorithm can be written as

$$W_{k+1} = W_k - \alpha_k . g_k$$

Where $W_k$ is a vector of current weights, $g_k$ is the current gradient, and $\alpha_k$ is the learning rate.

If gradient is greater than the threshold value then it performs next iteration. The batch steepest descent training function is traingd. The weights and biases are updated in the direction of the negative gradient of the performance function. There is only one training function associated with a given network.There are various training parameters associated with traingd: epochs, show, goal, time, min_grad, and lr. The learning rate lr is multiplied with the negative of the gradient to determine the changes to the weights and biases. The larger the learning rate, the bigger the step. If the learning rate is made too large, the algorithm becomes unstable. If the learning rate is set too small, the algorithm takes a long time to converge.The training status is displayed for every iteration of the algorithm. The other parameters determine when the training stops. The training stops if the number of iterations exceeds epochs, if the performance function drops below goal and if the magnitude of the gradient is less than mingrad, etc. Learning rate has been taken as 0.2 which is the optimum value.

### 4.2 Creating a Network (newff)

The first step in training a feedforward network is to create the network object. The function newff creates a feedforward network. It requires four inputs and returns the network object. The first input is an R by 2 Rx2 matrix of minimum and maximum values for each of the R elements of the input vector. The second input is an array containing the sizes of each layer. The third input is a cell array containing the names of the transfer functions to be used in each layer. The final input contains the name of the training function to be used.The following command creates a two-layer network. There is one input vector with two elements. The values for the first element of the input vector range between -1 and 2, the values of the second element of the input vector range between 0 and 5. There are three neurons in the first layer and one neuron in the second (output) layer. The transfer function in the first layer is tan-sigmoid, and the output layer transfer function is linear. The training function is traingd.

```
Net=newff([-12;0 5],[3,1],{'tansig','purelin'},'traingd');
```

This command creates the network object and also initializes the weights and biases of the network; therefore the network is ready for training. There are times when you may want to reinitialize the weights, or to perform a custom initialization. The next section explains the details of the initialization process.

### 4.3 Initializing Weights (init)

Before training a feedforward network, the weights and biases must be initialized. The newff command will automatically initialize the weights, but their reinitialization may be done with command init. This function takes a network object as input and returns a network object with all weights and biases initialized. Here is how a network is initialized (or reinitialized):

```
net = init(net);
```

### 4.4 Simulation (Sim)

The function sim simulates a network. sim takes the network input p, and the network object net, and returns the network outputs a. sim is called to calculate the outputs for a concurrent set of three input vectors. This is the batch mode form of simulation, in which all of the input vectors are place in one matrix. This is much more efficient than presenting the vectors one at a time.

```
p = [1 3 2; 2 4 1];
a=sim(net,p)
a = -0.1011   -0.2308   0.4955
```

### 4.5 Training

Once the network weights and biases have been initialized, the network is ready for training. The network can be trained for function approximation (nonlinear regression), pattern association, or pattern classification. The training process requires a set of examples of proper network behavior - network inputs p and target outputs t. During training the weights and biases of the network are iteratively adjusted to minimize the network performance function net.performFcn. The default performance function for feedforward networks is mean square error mse - the average squared error between the network outputs a and the target outputs t.

#### 4.5.1Batch Gradient Descent (traingd)

The batch steepest descent training function is traingd. The weights and biases are updated in the direction of the negative gradient of the performance function. If training a network using batch steepest descent is desired, the network trainFcn should be sat to traingd, and the function train is then called. There is only one training function associated with a given network.

Out of seven training parameters associated with traingd: epochs, show, goal, time, min_grad, and lr, the learning rate lr is multiplied with the negative of the gradient to determine the changes to the weights and biases. The larger the learning rate, the bigger the step. If the learning rate is made too large, the algorithm becomes unstable. If the learning rate is set too small, the algorithm takes a long time to converge. The training status is displayed for every show iteration of the algorithm.The other parameters determine when the training stops. The training stops if the number of iterations exceeds epochs, if the performance function drops below goal, if the magnitude of the gradient is less than mingrad, or if the training time is longer than time seconds. max_fail, which is associated with the early stopping technique is discussed in the section on improving generalization.

The following code creates a training set of inputs p and targets t. For batch training, all of the input vectors are placed in one matrix.

```
p = [-1 -1 2 2;0 5 0 5];
t = [-1 -1 1 1];
```

Next the feedforward network is created. Here e the function minmax is used to determine the range of the inputs to be used in creating the network.

```
net=newff(minmax(p),[3,1],{'tansig','purelin'},'traingd');
```

At this point, some of the default training parameters might be modified as follows:

```
net.trainParam.show = 50;
net.trainParam.lr = 0.05;
net.trainParam.epochs = 300;
net.trainParam.goal = 1e-5;
```

to use the default training parameters, the above commands are not necessary.
Now, the network is ready to be trained.

```
[net,tr]=train(net,p,t);

TRAINGD, Epoch 0/300, MSE 1.59423/1e-05, Gradient

2.76799/1e-10
TRAINGD, Epoch 50/300, MSE 0.00236382/1e-05,

Gradient
0.0495292/1e-10
TRAINGD, Epoch 100/300, MSE 0.000435947/1e-05,

Gradient
0.0161202/1e-10
TRAINGD, Epoch 150/300, MSE 8.68462e-05/1e-05,

Gradient
0.00769588/1e-10
TRAINGD, Epoch 200/300, MSE 1.45042e-05/1e-05,

Gradient
0.00325667/1e-10
TRAINGD, Epoch 211/300, MSE 9.64816e-06/1e-05,

Gradient
0.00266775/1e-10
TRAINGD, Performance goal met.
```

## 5. Experimental Results

### Quadrant Feature Extraction

The matlab v7 has been taken for the implementation of newly developed **QSFE** technique. Experiment has been carried out on 120 samples of training set and 150 samples of test set. Result of this experiment is given below in Table 1.

Table: 6.1 Result of Handwritten English character recognition using QSFE Technique

| Input to MLPN | No of Iterations | Training Time (sec) | Classification Time (ms) | Performance on Training set (%) | Performance on Test Set (%) |
|---|---|---|---|---|---|
| 4 x 1 direction input | 50 | 6.038 | 7.454 | 100 | 96 |

This technique is giving very good accuracy around 96% and it requires less time to train the network because of the discrete integer calculation being performed and due to more information content available for the network to learn and recognize handwritten English characters easily and quickly. There are small variations in direction values due to which network recognizes the character with high accuracy. Experiment shows that this technique requires less number of iterations for training the network.It also reduces training time. Experiment also shows that only 50 iterations are sufficient for training the network fully.

## 6. Conclusion & Future Scope

The work present in this paper is an effort towards recognition of hand-written English characters and words. This can be extended to any type of character provided it is given sufficient training time and sufficient inputs. The accuracy of this completed work depends on the level of training that has been given. This work demonstrates the application of MLP networks to the hand-written English character problem. The skeletonized and normalized binary pixels of English characters as well as features of these characters were used as the inputs of the MLP. As future extension to this research work, the recognition accuracy of the network may further be improved by using more training samples and by using more efficient feature extraction techniques. Finally the work cited in this paper is a tiny step towards the completion of a large goal which can bring new possibilities in the field of text recognition.

## REFERENCES

[1] Rajawelu, M.T. Husilvi, and M.V.Shirvakar, "A neural network approaches to character recognition." IEEE Trans. on Neural Networks, vol. 2, pp. 307-393, 1989.

[2] Parhami and M. Taragghi, "Automatic recognition of printed Farsi text,"IEEE Pattern Recognition, Vol. no. 8,
pp. 787-1308, 1990.

[3] C. Tappert, C.J. Suen and T. Wakahara,"The state of the art in outline handwriting recognition," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. PAMI-12, no.8, pp.707-808, 1990.

[4]Cheng-Lin Liu, "Normalization-Cooperated Gradient Feature Extraction for Handwritten Character Recognition" IEEE Transactions on Pattern Analysis and

Machine Intelligence, Volume 29, Issue 8, Aug. 2007 Page(s) 1465-1469.

[5] D.S. Yeung, "A neural network recognition system for handwritten Chinese character using structure approach," Proceeding of the World Congress on Computational Intelligence, vo1.7, pp. 4353-4358, Orlando, USA, June 1994.

[6] D.Y. Lee, "Handwritten digit recognition using K nearest-neighbor, radial basis function and backpropagation neural networks,"IEEE Neural computation, vol. 3, Page(s) 440- 449.

[7] E. Cohen, 1.1. Hull and S.N. Shrikari, "Control structure for interpreting handwritten addresses," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 16, no. 10, pp. 1049-1055, Oct. 1994.

[8] H. Almualim and S. Yamaguchi, "A method for recognition of Arabic cursive handwriting," IEEE Trans. on Pattern and Machine Intelligence, vol. PAMI-9, no 5, pp.715-722, Sept. 1987.

[9] Hailong Liu, Xiaoqing Ding, “Handwritten character recognition using gradient feature and quadratic classifier with multiple discrimination schemes” Eighth IEEE International Conference on Document Analysis and Recognition, Vol. 1, Page(s): 19-23, August 29-1 Sept., 2005.

[10] I.S.I. Abuhaiba and S.A. Mahmoud, "Recognition of handwritten cursive Arabic Characters," IEEE Transaction on PA&MI vol.16, no 6, pp. 664-672, June 1994.

[11] J. Hertz, A. Krogh and R. Palmer, “Introduction to the theory of neural computation," Addison-Wesley Publishing Company, USA, 1991.

[12 K. Yamada and H. Kami, "Handwritten numeral recognition by multilayered neural network with improved learning algorithm,” IJCNN Washington DC, vol. 2, pp. 259-266, 1989.

[13] Neural Computing Theory and Practices by Philip D. Wasserman.

[14] Neural Networks, Fuzzy Logic, and Genetic Algorithms by S. Rajasekaran and G.A. Vijaylakshmi Pai, PHI publication, India.

[15] P. Morasso, "Neural models of cursive script handwriting," IJCNN, WA, vol. 2, pp. 539-542, June 1989.

[16] S.J. Smith and M.O. Baurgoin, "Handwritten character classification using nearest neighbor in large database," IEEE Trans. on Pattern and Machine Intelligence, vol. 16, pp. 915-919, Oct. 1994.

[17] Starzyk,J.A.Ansari, “ Feedforward neural network for handwritten character recognition”, IEEE International Symposium on Circuits and Systems(ISCAS), Volume 6, Page(s) 2884-2887, 1992.

[18] Sutha.J, Ramraj.N, “Neural Network Based Offline Tamil Handwritten Character Recognition System”, IEEE International Conference on Computational Intelligence and Multimedia Application, 2007 Volume 2, 13-15, Dec.2007, Page(s): 446-450, 2007.

[19] Verma B.K, “Handwritten Hindi Character Recognition Using Multilayer Perceptron and Radial Basis Function Neural Network”, IEEE International Conference on Neural Network, vol.4, pp. 2111-2115, 1995.

[20] W.K. Verma, "New training methods for multilayer perceptrons," Ph.D Dissertation, Warsaw Univ. of Technology, Warsaw, March 1995.

[21 Weipeng Zhang; Yuan Yan Tang; Yun Xue. “Handwritten Character Recognition Using Combined Gradient and Wavelet Feature” IEEE International Conference on Computational Intelligence and Security, Volume 1, Page(s) 662-667, Nov. 2006.

# A Binary Tree Based Approach to Discover Multiple Types of Resources in Grid Computing

Leyli Mohammad khanli[1], Ali Kazemi Niari [2] and Saeed Kargar[2]
[1]Assistance Professor, Cs Dept. University of Tabriz, Tabriz, Iran, [2] MS student, Islamic Azad University-Tabriz Branch, Tabriz, Iran,
{l-khanli@tabrizu.ac.ir, a.kazemi.n@gmail.com, saeed.kargar@gmail.com}

***Abstract*:** Today grid technology considered as a solution to solve complex problems. Grid included a large number of heterogeneous resources. So, a resource discovery mechanism should be able to discover these heterogeneous and dynamic resources in such environment. In many of the previous methods, there are not possible for discover multiple types of resources in a framework simultaneously. In this paper, we propose a binary tree based to discover multiple types of resources for grid environment. In this method, for discover multiple types of resources, we send a unique request. We compare our method with other methods using simulation. The experimental results show that our method for resource discovery has better performance.

***Keywords*:** Grid, Discover multiple types of resources, Tree structure.

## 1. Introduction

Grid presented as a way to solve special problems [1]. This environment included a large number of heterogeneous resources. Resource discovery; i.e. finding users request is the most important challenges in the grid [2].

A user request may be a combination of multiple types of resources. This challenge should be able to discover the resources in a distributed environment.

A recent decentralizes method [3], uses tree structure for resource discovery in grid environment. Also, current method has presented a resource discovery method with one resource and different qualities. Therefore, discover more resources with different qualities needed to send separate and large number requests. In our method, we use tree structure, like [4]. In our proposed method, all tree nodes use the same table for identifying the resources available in it. The table contains all information related to the resources and their attributes. Our mechanism considers a combined package in every node in the grid and every node fills in its combined package using the table.

In our method users can be found to several resources simultaneously. So, in instead of the previous methods, it would be possible for the user to access to multiple types of resources only by one request.

This paper is organized as follows: Section 2 includes related works. In section 3, we explain our proposed method. Section 4 involved experimental results and section 5 would be the conclusions.

## 2. Related work

Grid provided infrastructure for sharing resources with different types in environment. Resource discovery approaches is of special importance in grid environment. Many of the methods proposed for resource discovery in grid.

One of the methods presented for resource discovery problem was the so called matchmaking one to solve the Condor problems [5]. Some researchers also used method matchmaking as a new method in their works [6]-[10]. In the current method, entities advertise their requirements and characters to the environment and a matchmaking service has to find a match between advertisement and entities.

Some resource discovery methods use resource brokers to match the resources between resource consumers and resource providers to find the best resource. Resource brokers use some factors in decision making resource availability, software/hardware capabilities, network bandwidth and resource price [11],[12].

In [3], Ruay-Shiung Chang et al. proposed a resource discovery tree for grid which using bitmap. Any request by the user, should be changed into bitmap form. If a request reaches to one of the tree nodes, it will be compared with its local resource bitmap in which the data related to the local resources kept in (by AND operation). If the requested resource not found, (provided that current node is a leaf node), then the request will be sent to the parent node. Otherwise, the requests are compared with index bitmap in which the data related to children resources kept in. If the resource not found in children, too, the request will be sent to the parent node and until the requested resource be found, current process continues.

In our previous work [4], we proposed another method in resource discovery which uses a weighted tree in grid environment. In the current method, users request directed to the target node through unique paths i.e. the requests are delivered to nodes in a determined format and the nodes send the request to their chosen children using the reserved information; but if no resource exists at the node and its children, it will be delivered to their parent node.

Our algorithm has some differences with previous ones:

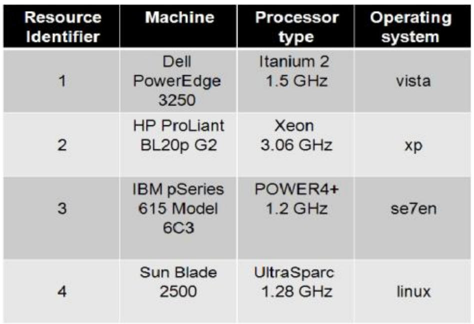
In our method, every node has maximum of 2 children which makes it easier for the parent nodes to keep and manage the children.

For all nodes, we use the same table. It means that information available in tree would be accessible to all nodes only once, regarding the available resources in grid environment. Table information will not change and the nodes fill their related combined package using the table and the users only manipulate the small packages.

## 3. Binary tree based approach to discover multiple types of resources

### 3.1 Request resources in different types

There are many heterogeneous resources in the grid which are geographically distributed. These resources have different types. The user may simultaneously need multiple types of resources. For example, machine (Dell PowerEdge 3250), processor type (Itanium 2, 1.5 GHz), operating system (vista), and so on. So, there should be a method in which the user may request resources in different types. Resources and their types should be known for the nodes. Managing the information of these heterogeneous resources, we propose a *resource-table* and some packages which discussed. First, we insert a list of resources with their types in the *resource-table*. Current table would be supplied to all nodes in the grid as a catalog. Figure 1 shows an example of *resource-table*. Any node has a "*Local Combined Package*" (LCP) which can extract the related code to its local resources and insert in the LCP.

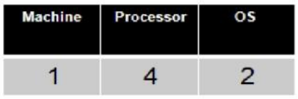


| Resource Identifier | Machine | Processor type | Operating system |
|---|---|---|---|
| 1 | Dell PowerEdge 3250 | Itanium 2 1.5 GHz | vista |
| 2 | HP ProLiant BL20p G2 | Xeon 3.06 GHz | xp |
| 3 | IBM pSeries 615 Model 6C3 | POWER4+ 1.2 GHz | se7en |
| 4 | Sun Blade 2500 | UltraSparc 1.28 GHz | linux |

**Figure 1.** An example of resource-table

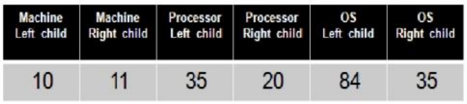
For example, in Figure 2, we show a LCP with the following resources: machine (Dell PowerEdge 3250), processor type (UltraSparc 1.28 GHz), operating system (XP). The node places the related code for their local resources on the LCP, using resource- table of Figure 1. Besides the *Local Combined Package*, the node which has children needs another combined package for children which called "*Children Combined Package*" (CCP) to determine resources available in children (Figure 3). So, the CCP contains information of collected resources from children.

When a node receive a request on resource discovery, it will be compared with LCP and then with CCP. The information of packages will be used to find the requested resources.



| Machine | Processor | OS |
|---|---|---|
| 1 | 4 | 2 |

**Figure 2.** An example of Local combined package.



| Machine Left child | Machine Right child | Processor Left child | Processor Right child | OS Left child | OS Right child |
|---|---|---|---|---|---|
| 10 | 11 | 35 | 20 | 84 | 35 |

**Figure 3**. An example of children combined package.

### 3.2 Initialization of our tree nodes

In this subsection, initialization means collecting data of children nodes by parent nodes during building the tree. Any node can place the received data of children in the CCP and send complete information to the parent node.

Figure 4 shows a typical grid environment with seven nodes which are connected with a tree structure. The tree is a binary one; i.e. any node can contain maximum of two children. The advantage of the tree would be easy management of children.

Here, we discuss how data collected for Machine resource. First, information of leaf nodes (D, E) which would be 2 and 3 are sent to node B and for nodes F and G (i.e. 1, 3) it will be sent to node C. Node B which received number 2 from his left child (D), insert it in left position of Machine (CCP of node B), but number 3 which is received from right child (E) are insert in right position of Machine. Node C does the same process for its children nodes (F, G); i.e. for CCP of node C, left position=1 and right position=3.

Nodes B and C should send their information to the parent node (A). We explain this process by more details. Left position of CCP=2, right position of CCP=3, LCP=2 which are in node B, are added to each other (2+3+2→7) and then send to the parent node (A). The same process performs for node C until number 7 insert in right position of Machine for node A (1+3+3→7).

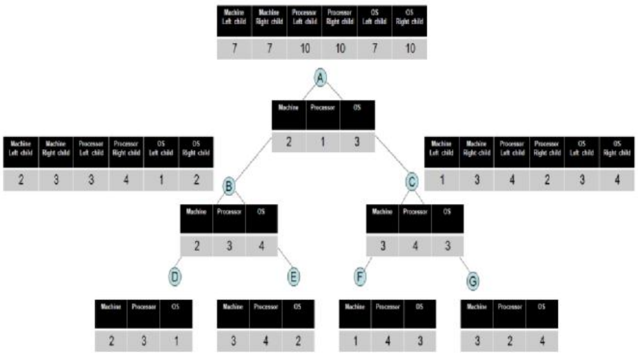
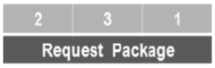

Machine Left child | Machine Right child | Processor Left child | Processor Right child | OS Left child | OS Right child
7 | 7 | 10 | 10 | 7 | 10
A
Machine | Processor | OS
2 | 1 | 3
Machine Left child | Machine Right child | Processor Left child | Processor Right child | OS Left child | OS Right child
2 | 3 | 3 | 4 | 1 | 2
B
Machine | Processor | OS
2 | 3 | 4
C
Machine Left child | Machine Right child | Processor Left child | Processor Right child | OS Left child | OS Right child
1 | 3 | 4 | 2 | 3 | 4
Machine | Processor | OS
3 | 4 | 3
D
E
F
G
Machine | Processor | OS
2 | 3 | 1
Machine | Processor | OS
3 | 4 | 2
Machine | Processor | OS
1 | 4 | 3
Machine | Processor | OS
3 | 2 | 4

**Figure 4.** An example of typical grid environment.

### 3.2 Resource discovery in typical grid environment

In initialization, information related to the resources sent from children node to parent one. Therefore, all nodes obtain complete information of resources available in their children and descendants. Finally, the final form of tree is created. Now, the requested resources of users can discover in the current tree structure. There would be a frame which the users can record their requested resources and deliver it to a node. Our goal is that the user would be able to simultaneously request multiple types of resources. In Figure 5, the user requests the following resources: machine (HP ProLiant BL20p G2) processor type (POWER4+ 1.2 GHz), operating system (vista) as a request package. After forming the request package, the user should deliver it to one of the tree node. Figure 6 shows an example for resource discovery on a binary tree.



| 2 | 3 | 1 |
|---|---|---|
| Request Package | | |

**Figure 5**. An example of request package

We suppose that in Figure 6, the request delivered to node G. Regarding the current algorithm, the request first compared with LCP of node G, but no match is found, so, the request delivered to parent node (C). Node C also compares it with its LCP and again no match is found. The request is compared with left positions of CCP (because the request comes from right child) and because no information of requested resources is found, then requested package forward to the parent node. Node A compared this request package with own LCP, then compared it with left positions of own CCP.

Here, a probability match is found for the requested resource in left child of node A (node B). So, the request package is forward to node B. Operation of node B would be as follows:

Node B first compared request package with its LCP and because this LCP not equals with request package then compared it with CCP. After this operation, node B found information about requested resources in its left child i.e. node D and right child i.e. node E and send request package to these nodes. The requested package that sent by node B which reaches node E, when compared with LCP of node E, can't discover the user's requested resources in this node, but the requested resources discovered in node D. Finally, send a success message to the node G and discovered resources would be reserve for user.

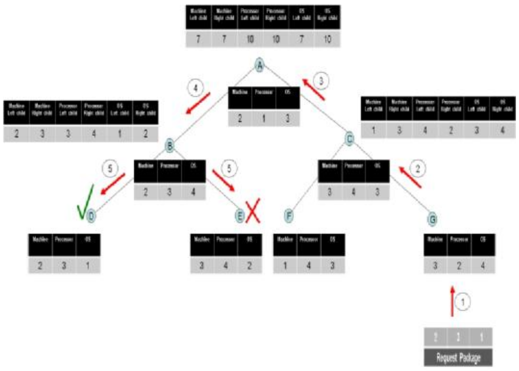


**Figure 6**. An example for resource discovery.

## 4 Experimental results

### 4.1 Setup

We performed the experiment at MATLAB environment and the resources and requests randomly distributed between nodes. Our experiment performed on a binary tree and compared with other approaches.

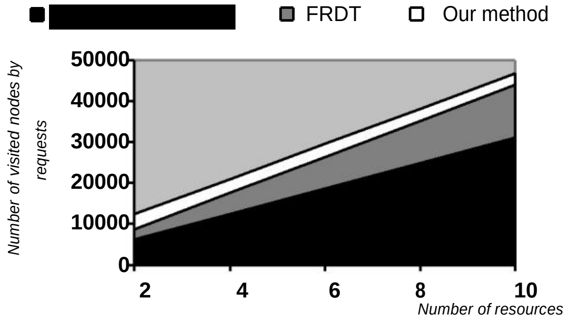
It has to be noticed that the methods which have been compared with ours, are the ones proposed for the discovery of just one resource based on a tree structure in the grid environment. Because, we didn't meet applied methods able to simultaneous discovery of multiple types of resources, so some methods recently presented for the discovery of one resource with different attributes on a tree will be compared with our method. In this process, we supposed that the current methods send the user's request separately and then discover these resources for the user.

For the other methods, we suppose that after the resources discovered and reserved for the user and before using them, the location of the discovered resources have to be compared and if all belong to one node, then they would be usable by the user. But in order to reduce the complexities available in simulation, we suppose for all simulations in other methods that all of the separated requests discover the resources from one node, but it is rarely possibility to occur.

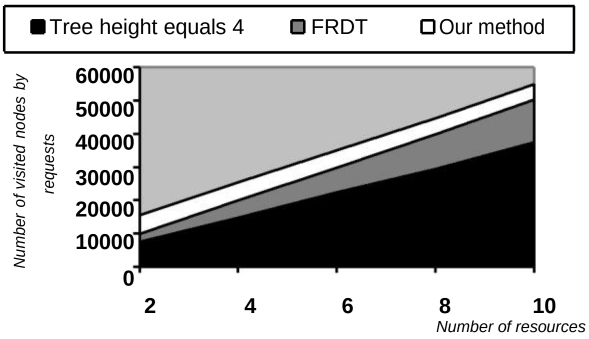
## 4.2 Simulation results

We conducted the experiments with different number of requested resources; i.e. one of the experiments the user requests only one resource and in some other experiments, user simultaneously requests two resources and so on. In other methods, for example, if user demands 3 resources, user should send 3 separate requests, but, in our method, just one request will be sent.

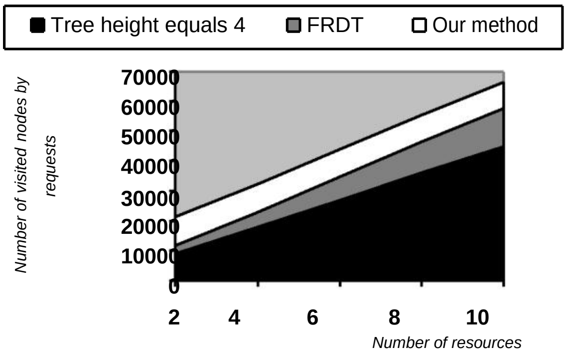
In the first experimental tests, the number of nodes which the requests send in tree method and FRDT with height 4, compared with our method. We supposed 300 users that everyone requested different number of resources (Figure 7 (a, b, c, d)).



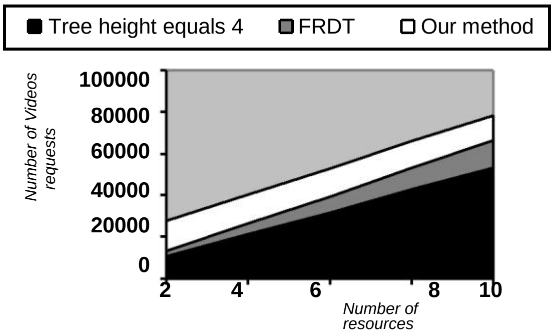
(a) # nodes= 85



(b) # nodes=156



(c) # nodes=259



(d) #nodes=400

**Figure 7.** The number of visited nodes for 300 requests.

In the second experimental tests, the average number of nodes in which the requests are sent, shows for methods flooding-based algorithm, MMO [13],[14] and our method. The results presented in Figure 8. In this test 300 users requesting one resource.

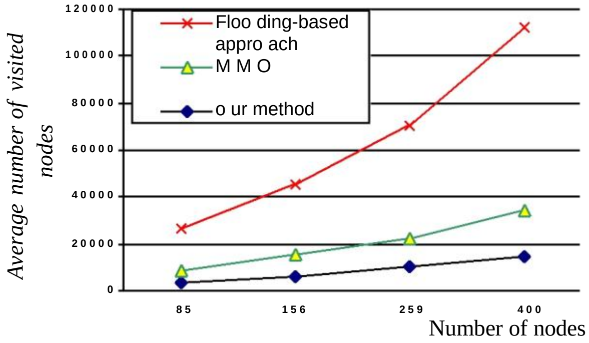


**Figure 8**. Average number of nodes that equests are forwarded using different approach.

In the last experimental tests, number of visited nodes during resource discovery and updating are show in tree and FRDT methods and also in our method. In the first environment, we supposed 100 users that everyone requested five resources (Figure 9) and in the second environment we supposed 100 users that everyone requested ten resources, each time (Figure 10). As observed in Figure 9 and Figure 10, if the number of user increases and every user request more resources, each time, our method would be more efficient comparing other methods.

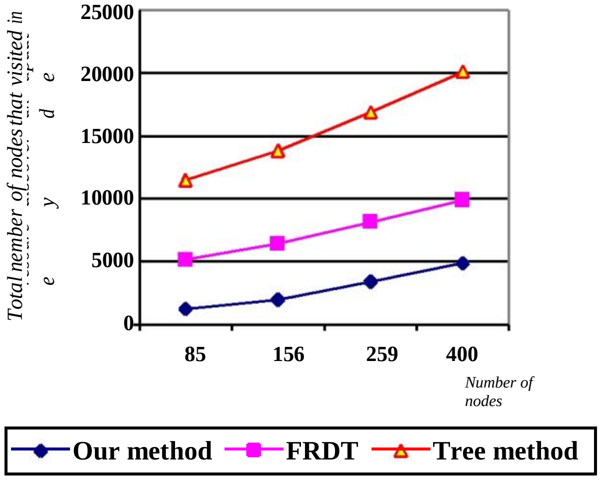


**Figure 9.** Total number of visited nodes by requests for 100 users each user requests five resources**.**

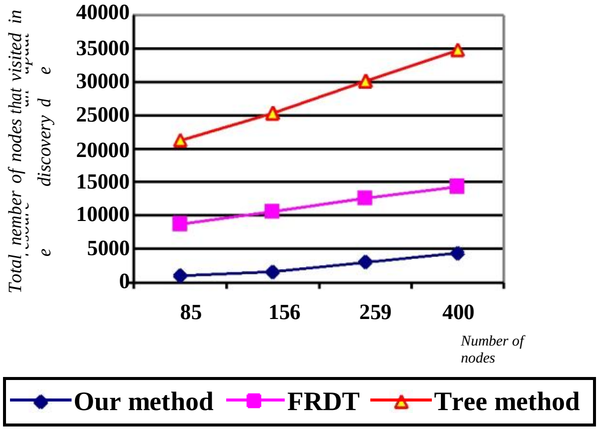


**Figure 10.** Total number of visited nodes by requests for 100 users each user requests ten resources**.**

## 5 Conclusions and Future Work

In this paper, we proposed a resource discovery mechanism which is on the basis of binary tree. So, every node has maximum of 2 children which makes it easier for the parent nodes to keep and manage the children. In our method, the users can send multiple requests in the form of a unique request.

In the future, if we can find a method that by using it, we discovered several resources on a free tree mode (not merely binary), we can improve the resource discovery mechanism significantly.

## References

[1] Ian Foster, Carl Kesselman, The Grid 2: Blueprint for a New Computing Infrastructure, Morgan Kaufmann Publishers Inc., San Francisco, CA, 2003.

[2] YiLi, G., FangPeng, D., Wei, L., ZhiWei, X.:VEGA Infrastructure for Resource Discovery in Grids. J. Comput. Sci. & Technol, pp.413-422 (2003)

[3] Shiung Chang, R., Shuo Hu, M.: A resource discovery tree using bitmap for grids. Future Generation Computer Systems (2009)

[4] L.M. Khanli, S. Kargar, FRDT: Footprint Resource Discovery Tree for grids, Future Gener. Comput. Syst. 27 (2011) 148–156.

[5] R. Raman, M. Livny, M. Solomon, Matchmaking: distributed resource management for high throughput computing, hpdc, in: Seventh IEEE International Symposium on High Performance Distributed Computing (HPDC-7’98), 1998, p. 140.

[6] K.I. Karaoglanoglou, H.D. Karatza, Resource Discovery in a dynamical grid based on Re-routing Tables, Simulation Modelling Practice and Theory 16 (2008) 704–720.

[7] Rajesh. Raman, Matchmaking Frameworks for Distributed Resource Management, University of Wisconsin-Maddison, 2001.

[8] Ye Zhu, Junzhou Luo, Teng Ma, Dividing Grid Service Discovery into 2-stage matchmaking, ISPA 2004, LNCS, vol. 3358, 2004, pp. 372–381.

[9] S .Tangpongprasit, T .Katagiri, H .Honda, T .Yuba, A time-to-live based reservation algorithm on fully decentralized resource discovery in grid computing, Parallel Computing 31 )6( )2005(529-543.

[10] Muthucumaru Maheswaran, Klaus Krauter, A parameter-based approach to Resource Discovery in Grid Computing Systems, GRID, 2000.

[11] Bradley, A., Curran, K., Parr, G., 2006. Discovering resource in computational GRID environments. The Journal of Supercomputing, 35, 27–49.

[12] D. Lacks, T. Kocak, Developing reusable simulation core code for networking: The grid resource discovery example, The Journal of Systems and Software 82 (2009) 89-100.

[13] M. Marzolla, M. Mordacchini, S. Orlando, Resource discovery in a dynamic environment, in: Proceedings of the 16th International Workshop on Database and Expert Systems Applications, DEXA'05, September 3_7, 2005, pp. 356_360.

[14] M. Marzolla, M. Mordacchini, S. Orlando, Peer-to-peer systems for discovering resources in a dynamic grid, Parallel Computing 33 (4_5) (2007) 339_358.

www.ijcset.excelingtech.co.uk