Non Linear Cellular Automata in Improving the Quality of Clustering For Medical Image Processing

P.Kiran Sree[1], I Ramesh Babu[2], N.S.S.S.N Usha Devi[3]

Abstract

Clustering can be applied to medical image processing on the grounds of its potential improved effectiveness over conventional image search. Clustering is a mostly unsupervised procedure and the majority of the clustering algorithms depend on certain assumptions in order to define the subgroups present in a data set .A clustering quality measure is a function that, given a data set and its partition into clusters, returns a non-negative real number representing the quality of that clustering. Moreover, they may behave in a different way depending on the features of the data set and the input parameters values. Therefore, in most applications the resulting clustering scheme requires some sort of evaluation as regards its validity. The quality of clustering can be enhanced by using a Cellular Automata Classifier for medical image processing. In this paper we take the view that if cellular automata with clustering is applied to search results, then it has the potential to increase the retrieval effectiveness compared both to that of static clustering and of conventional image search. We conducted a number of experiments using five image collections and four hierarchic clustering methods. Our results show that the effectiveness of query-specific clustering with cellular automata is indeed higher, and suggest that there is scope for its application to medical image processing.

Keywords : Non Linear Cellular Automata, Medical Image Processing, Clustering.

1. Introduction

Locating interesting information is one of the most important tasks in Medical Image Processing (MIR). An MIR system accepts a query from a user and responds with a set of images. The system returns both relevant and non-relevant material and an image organization approach are applied to assist the user in finding the relevant information in the retrieved set. Generally a search engine presents the retrieved image set as a ranked list of image titles. The images in the list are ordered by the probability of being relevant to the user's request. The highest ranked image is considered to be the most likely relevant image; the next one is slightly less likely and so on. This organizational approach can be found in almost any existing search engine. It is assumed that the user will start at the top of the list and follow it down examining the images one at a time. A number of alternative image organization approaches have been developed over the recent years. These approaches are normally based on visualization and presentation of some relationships among the images, terms, or the user's query. One of such approaches is image clustering. Image clustering has been studied in the Medical Image Processing for several decades.

Clustering is a mostly unsupervised procedure and the majority of the clustering algorithms depend on certain assumptions in order to define the subgroups present in a data set. Moreover, they may behave in a different way

[1] Associate Professor, Department of C.S.E, S.R.K.I.T,Vijayawada. email : profkiran@yahoo.com.
[2] Professor & Senior IEEE Member, C.S.E,Acharya Nagarjuna University, Guntur.
[3] Graduate Student of J.N.T.University.

depending on the features of the data set and the MIR input parameters values. Therefore, in most applications the resulting clustering scheme requires some sort of evaluation as regards its validity.

Aiming towards the development of a general clustering theory, addressing issues that are common to the different clustering paradigms, we wish to initiate a systematic study of measures for the quality of a given data clustering. A clustering quality measure is a function that, given a data set and its partition into clusters, returns a non-negative real number representing the quality of that clustering. We analyze what clustering quality measures should look like by introducing a set of requirements ('axioms') of clustering quality measures... Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters). The clustering problem has been addressed in many contexts and by researchers in many disciplines; this reflects its broad appeal and usefulness as one of the steps in exploratory data analysis. However, clustering is a difficult problem combinatorial, and differences in assumptions and contexts in different communities have made the transfer of useful generic concepts and methodologies slow to occur.

2. Cellular Automata (CA) and Fuzzy Cellular Automata (FCA)

A CA [4], [5], [6], consists of a number of cells organized in the form of a lattice. It evolves in discrete space and time. The next state of a cell depends on its own state and the states of its neighboring cells. In a 3-neighborhood dependency, the next state qi (t + 1) of a cell is assumed to be dependent only on itself and on its two neighbors (left and right), and is denoted as

$$qi(t+1) = f(qi-1(t), qi(t), qi+1(t)) \quad \text{-----(1)}$$

where qi (t) represents the state of the ith cell at tth instant of time, f is the next state function and referred to as the rule of the automata. The decimal equivalent of the next state function, as introduced by Wolfram, is the rule number of the CA cell [9],[10],[11]..In a 2-state 3-neighborhood CA, there are total 256 distinct next state functions.

2.1 FCA Fundamentals

FCA [2], [6] is a linear array of cells which evolves in time. Each cell of the array assumes a state qi, a rational value in the interval [0, 1] (fuzzy states) and changes its state according to a local evolution function on its own state and the states of its two neighbors. The degree to which a cell is in fuzzy states 1 and 0 can be calculated with the membership functions. This gives more accuracy in finding the coding regions.In a FCA, the conventional Boolean functions are AND , OR, NOT.

2.2. Dependency Matrix for FCA

Rules defined in equations 1, 2 should be represented as a local transition function of FCA cell. That rules are converted into matrix form for easier representation of chromosomes.

Table 1: FA Rules

Non-complemented Rules		Complemented Rules	
Rule	Next State	Rule	Next State
0	0	255	1
170	q_{i+1}	85	$\overline{q}_{i+1}$
204	q_i	51	$\overline{q}_i$
238	$q_i + q_{i+1}$	17	$\overline{q_i + q_{i+1}}$
240	q_{i-1}	15	$\overline{q}_{i-1}$
250	$q_{i-1} + q_{i+1}$	5	$\overline{q_{i-1} + q_{i+1}}$
252	$q_{i-1} + q_i$	3	$\overline{q_{i-1} + q_i}$
254	$q_{i-1} + q_i + q_{i+1}$	1	$\overline{q_{i-1} + q_i + q_{i+1}}$

Example 1: A 4-cell null boundary hybrid FCA with the following rule < 238, 254, 238, 252 > (that is, < (qi+qi+1), (qi-1+qi+qi+1), (qi + qi+1), (qi-1 + qi) >)

applied from left to right, may be characterized by the following dependency matrix

While moving from one state to other, the dependency matrix indicates on which neighboring cells the state should depend. So cell 254 depends on its state, left neighbor, and right neighbor fig (1). Now we represented the transition function in the form of matrix. In the case of complement [5[,[6],[8],FMACA we use another vector for representation of chromosome.

Figure 1: Matrix Representation

$$T = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

2.3 Transition From One State To Other

Once we formulated the transition function, we can move form one state to other. For the example 1 if initial state is P (0) = (0.80, 0.20, 0.20, 0.00) then the next states will be

P (1) = (1.00 1.00, 0.20, 0.20),
P (2) = (1.00 1.00, 0.40, 0.40),
P (3) = (1.00 1.00, 0.80, 0.80),
P (4) = (1.00 1.00, 1.00, 1.00).

FMACA Based Pattern Classifier

An n-cell FMACA with k-attractor basins can be viewed as a natural classifier. It classifies a given set of patterns into k distinct classes, each class containing the set of states in the attractor basin.

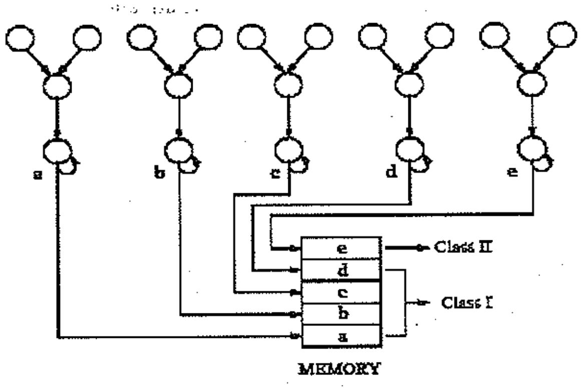


Figure 2: FMACA Based Classification Strategy With 5 Attractor Basins Classifying The Elements Into Two Classes.

Note : (i) An attractor basin covers the elements belonging to one class only.

(ii) Each attractor points to the memory location that stores the class information.

Suppose, we want to design a FMACA based pattern classifier to classify a training set S = {S1, S2, · · · , SK} into K number of classes. First, a FMACA with k (k ≥ K) number of attractor basins is generated. The training set S gets distributed into k attractor basins (nodes). Let, ′ S be the set of elements in an attractor basin. If ′ S belongs to only one class, then label that attractor basin as that class. Otherwise, this process is repeated recursively for each attractor basin (node) until all the patterns in each attractor basin belong to only one class.

3. Non Linear Cellular Automata

The linear/additive CA is amenable to detailed characterization with linear algebraic tools. Due to the absence of such a mathematical tool, there has been varied effort with different parameters to characterize non-linear CA. We detail the characterization of each of the categories separately. However, some very interesting works simulating non-linear CA from product of linear CA are recently reported.

These works are aimed at taking the advantage of linear algebraic tools to characterize the wide variety of non-linear CA state transition. One of the major thrust has been to study the non-linear CA dynamics as it evolves in successive time steps. The emergent patterns in the decentralized systems give rise to some form of globally coordinated behavior. A detailed study of CA dynamics helps us to understand the emergent behavior and analyze its computational power .CA classification based on the study of its dynamics was a major interest for the researchers.

A special class of non-linear CA, termed as Generalized Multiple Attractor CA (GMACA), has been proposed to develop the model. Theoretical analysis, reported in this chapter, provides an estimate of the noise accommodating capability of the proposed GMACA based associative memory model. Characterization of the basins of attraction of the proposed model establishes the sparse network of non-linear CA (GMACA) as a powerful pattern recognizer for memorizing unbiased patterns. It provides an efficient and cost-effective alternative to the dense network of neural net for pattern recognition. Detailed analysis of the GMACA rule space establishes the fact that the rule subspace of the pattern recognizing/classifying CA lies at the edge of chaos. Such a CA, as projected , is capable of executing complex computation.

The analysis and experimental results reported in the current and next chapters confirm this viewpoint. A GMACA employing the CA rules at the edge of chaos is capable of performing complex computation associated with pattern recognition. The design of non-linear CA based associative memory has so far mostly concentrated around uniform CA with same rule applied to each of the CA cells. This structure of uniform rule restricts the CA to evolve as a general purpose pattern recognizer although it may be able to memorize some specific patterns. Though some works on non-uniform CA has been reported , but none of the published literature has so far dealt with the evolution of CA as a general purpose pattern recognizer. Some of the recent works has succeeded in evolving the class of CA referred to as GMACA (Generalized Multiple Attractor CA) capable of simulating the associative memory model.

The simple, regular, modular, cascadable local neighborhood structure of CA serves as an excellent sparse network. It has been shown that the evolved CA displays encouraging results for pattern recognition. Also, extensive experimental results reported in these papers confirm the comparative advantage of CA technology.

4. K-Means Algorithm

K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume K clusters) fixed a priori.

The main idea is to define K centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate K new centroids as bar centers of the clusters resulting from the previous step. After we have these K new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the K centroids change their location step by step until no more changes are done. In other words centroids do not move any more. Finally, this algorithm aims at minimizing an objective function.

The objective criterion function[2]:

$$E = \max \sum_{r=1}^{K} \sum_{d \in S} \cos(d_i, c_r) = \sum_{r=1}^{K} \|D_r\| \qquad (2)$$

The algorithm is composed of the following steps:

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

Although it can be proved that the procedure will always terminate, the k-means algorithm does not necessarily find the most optimal configuration, corresponding to the global objective function minimum. The algorithm is also significantly sensitive to the initial randomly selected cluster centers. The k-means algorithm can be run multiple times to reduce this effect.

Actually, text clustering is to find a partition letting criterion function optimum. That is a combinatorial optimization problem, so we can use Cellular automata based local search[3] to do text clustering. According to text clustering features, a text clustering algorithm based on Cellular automata based local search is given in this section.

Definition:

$S = \{d_1, d_2, ..., d_N\}$: A text set includes N texts

$P = (S_1, S_2, ..., S_K)$: A K-partitioning (that is a cluster result) of text set S, where $S = Y_{i=1,2,...,K} S_i$, $S_i \text{ I } S_j = \varnothing$, when $i \neq j$. $n_i = |S_i|$: the number of texts in text set, $i = 1,2,..,K$ $D = \sum_{d \in S} d$: Sum of all text vectors in S. $c = D / \|D\|$: Central vector of text set S.

4.1 Medical Image Processing system evaluation

To measure ad hoc MIR effectiveness in the standard way, we need a test collection consisting of three things:

1. A image collection
2. A test suite of information needs, expressible as queries
3. A set of relevance judgments, standardly a binary assessment of either relevant or non relevant for each query-image pair.

4.2 Number of Classes

The number of classes is defined at the start of the process. It cannot grow but can be reduced when a class empties. In next figures, the indicated numbers of clusters correspond to the values initially chosen. The images that are not assigned to a class at the end of the classification are allocated tp a new one (at the last position in the ranked list of clusters). By choosing the same number of classes from 2 to 13 for all queries, the levels of the average precision over all relevant images are lower than those without classification with lists.

The differences between results indicated in Figure 3 and Figure 4 measure how much the above defined distance 4 We choose ni > ni+1 to favor the fMIRst ranked classes ranks the clusters. The average precision decrease is about 5% when clusters are ranked according to the computed distances and not according to the number of relevant images they contain.

Let Lq be the list of images constructed from the succession of the clusters ranked according to the number of relevant items they contain.

$$\Delta_{P'}E(P)=\sum_{d\in S'_i} d\cdot(c'_i-c_i)+\sum_{d\in S'_j} d\cdot(c'_j-c_j)+\Delta_{k} \quad (3)$$

$$\Delta_{p'}E(P)\geq\Delta_k \quad (4)$$

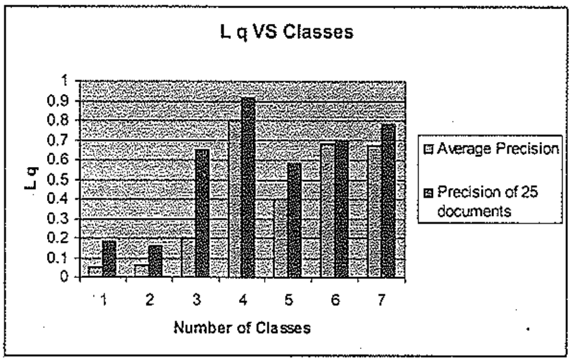


Figure 3 : L_q with different numbers of classes (precision at 10 and average precision without classification are respectively indicated)

Let L_c be the list of images constructed from the succession of the clusters ranked according to the MIR distances with the query using CA classifier $L_c = C1\times C2\times C3\times K$

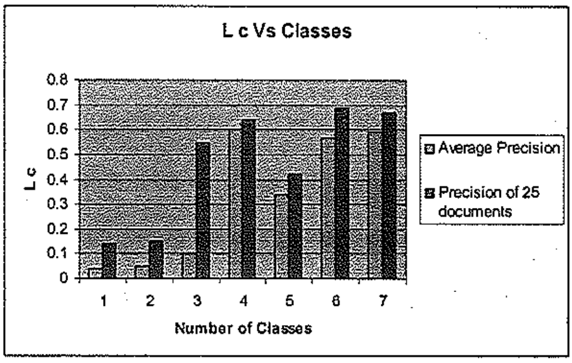


Figure 4 : L_C with different numbers of classes

By choosing the same number of classes from 5 to 25 for all queries, the levels of the average precision over all relevant images are lower than those without classification with lists Lq and LC (Figure 3 and Figure 4). The decrease rate varies from 2.2% to 3% . Figure 3 and Figure 4 show that those lists do not allow to globally improving results of the retrieval. The average precision decreases since the relevant images that are not in the fMIRst cluster are ranked after all items of that one. The differences between results indicated in Figure 3 and Figure 4 measure how much the above defined distance 4 We choose ni > ni+1 to favor the fMIRst ranked classes ranks the clusters. The average precision decrease is about 5% when clusters are ranked according to the computed distances and not according to the number of relevant images they contain.

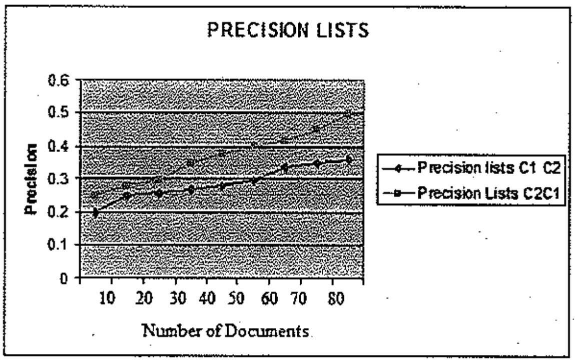


Figure 5 : Precision of lists C1.C2 and C2.C1 (Pure Clustering)

However, the fMIRst ranked cluster according to the distances to the queries is very often better than the next ones as shown in Figure 5 where we have compared lists C1.C2 and C2.C1 . With this second list, the relative decrease of the average precision over the 145 queries equals 28% (from 0.21 to 0.089). In Figure 6, we can see that the fMIRst ranked cluster is the best 3 times for the 5 queries indicated among 6 clusters.

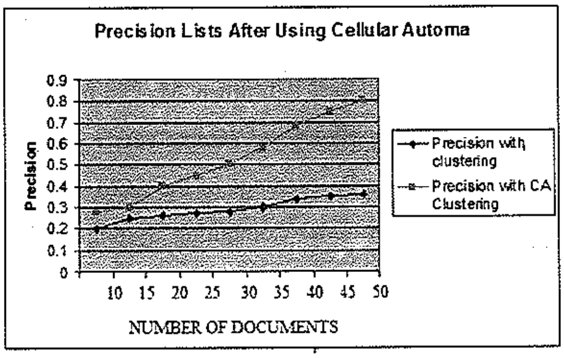


Figure 6 : Precision Using CA Classification

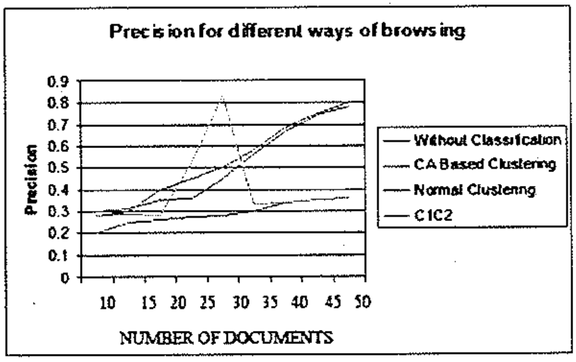


Figure 7: Precision For Different Browsing Aspects

Figure 6,7 show some results obtained with 203 clusters. One can see that precision at low level of recall with lists Ln are better than those of list LC (succession of each cluster's contents). However, only list Lq allows to obtain better results than without classification. At recall 0.21, the relative increase of precision of list L5 over list LC equals 19.5% (from 0.297 to 0.3892). Figure 6,7shows the results obtained by using the title field as queries and then the whole topic (list LCn with 2 clusters). Not surprisingly, the best results are obtained with longer queries even if in some cases the narrative field contains words not wanted.

5. Experimental Results

In order to evaluate the classification without taking into account the position of clusters, we use the list of relevant images supplied by NIST for TREC-7 and, for each query, select the best clusters according to the number of relevant images they contain .Hence, we can measure how much the classification groups relevant images. Let Lq be the list of images constructed from the succession of the clusters ranked according to the number of relevant items they contain.

6. Conclusion

We see that clustering with CA can greatly improve the effectiveness of the ranked list. In fact it can be as effective as the interactive relevance feedback based on query expansion. Surprisingly this high performance can be achieved by following a very simple strategy. Given a list of clusters created by the CA based local search algorithm starts at the top of the list and follows it down examining the images in each cluster. The experimental results proves the improvement of clustering quality with addition of Cellular Automata.

We have also shown how a Cellular automaton with clustering is used to retrieve images which help to regroup the relevant ones. It increases the effectiveness of retrieval by providing to users at least one cluster with a precision higher than the one obtained without using CA. We have examined, with TREC-7 corpora and queries, the impact on the classification results of the cluster numbers and of the way to browse them. We have shown that a variation of the number of clusters according to the query size improves the results. By automatically constructing a new ranked list according to the distances between clusters and queries, the precision is lower than without CA. The evaluation of other distances is in progress.

References

[1] Peter G. Anick & Shivakumar Vaithyanathan, *"Exploiting Clustering and Phrases for Context-Based Medical Image Processing"*, in Proceedings of ACM/SIGMIR'97, PP. 314-323, 1997.

[2] Ellen M. Voorhees, Donna Harman, *"Overview of the Seventh Text REtrievalConference (TREC-7)"*, NIST special publication, 1998.

[3] J. Allan, J. Callan, W. B. Croft, L. Ballesteros, D. Byrd, R. Swan and J. Xu, *"Inquery does battle with TREC-6"* In Sixth Text Retrieval Conference (TREC-6), PP 169-206, 1998.

[4] Cohen. A.M and Hersh. W, *"A survey of current work in biomedical text mining"*, Briefings in Bioinformatics, 6, 57-71, 2005.

[5] P. Kiran Sree, I.Ramesh Bababu,*"Identification of Protein Coding Regions in Genomic DNA Using Unsupervised FMACA Based Pattern Classifier"*, in International Journal of Computer Science & Network Security with ISSN: 1738-7906,Vol.8, No.1, 305-308.

[5] P.Kiran Sree, I.Ramesh Babu, *"Towards an Artificial Immune System to Identify and Strengthen Protein Coding Region Identification Using Cellular Automata Classifier"*, in International Journal of Computers and Communications, Vol, Vol. 1, Issue 1, Volume 1, Issue 2, PP. (26-34), 2007.

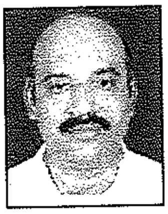
[6] P.Kiran Sree, Dr I. Ramesh Babu, N.S.S.N Usha Devi, *"Investigating an Artificial Immune System to Strengthen the Protein Structure Prediction and Protein Coding Region Identification using Cellular Automata Classifier"*, in International Journal of Bioinformatics Research and Applications , ISSN : 1744-5493.

[7] P. Kiran Sree, I. Ramesh Babu,*"NPCRIT: A Novel Protein Coding Region Identifying Tool using Decision Tree Classifier with Trust-Region Method and Cellular Automata Based Parallel Scan Algorithm"*, International journal of Advances in Computer Science and Engineering, with ISSN: 0973-6999

[8] Allen. R.B, Obry. P, Littman. M, *"An interface for navigating clustered image sets returned by queries"*, In Proceedings of the ACM Conference on Organizational Computing Systems, PP. 166-171. Milpitas, CA, 1993

Author's Biography

P. Kiran Sree received his B.Tech in Computer Science & Engineering, from J.N.T.U and M.E in Computer Science & Engineering from Anna University. He is pursuing Ph.D in Computer Science from J.N.T.U, Hyderabad. He has published many technical papers; both in international and national Journals& Conferences .His areas of interests include Cellular Automata, Parallel Algorithms, Artificial Intelligence, Compiler Design and Computer Networks. He also wrote books on Analysis of Algorithms, Theory of Computation and Artificial Intelligence. He was the reviewer for many International Journals and IEEE Society Conferences in Artificial Intelligence and Networks. He was also member in many International Technical Committees.He was the Associate Editor for Asian Journal of Scientific Research (ISSN: 1992-1454), Journal of Artificial Intelligence (ISSN: 1994-5450), Information Technology Journal (ISSN: 1812-5646), Journal of Software Engineering (ISSN: 1819-4311), and Research Journal of Information Technology. He was also invited speaker and organizing chairman for special sessions in WSEAS international conferences. He is the member of C.S.I, I.E.T.E, I.S.T.E (India), ICST (Europe) and IAENG (U.S.A). He is now associated with S.R.K Institute of Technology, Vijayawada.



Inampudi Ramesh Babu received his Ph.D in Computer Science from Acharya Nagarjuna University, M.E in Computer Engineering from Andhra University, B.E in Electronics & Communication Engg from University of Mysore. He is currently working as Professor in the department of computer science, Nagarjuna University. Also he is the senate member of the same University from 2006. He held many positions in Acharya Nagurjuna University as Head, Director - Computer Centre, Chairman- Board of studies. He was a special officer, convenor of ICET. He is also a member of Board of Studies for other universities. His areas of interest include Image Processing, Computer Graphics, Cryptography, Artificial Intelligence and Network Security. He is a member of IEEE, CSI, ISTE, IETE, IGISS, Amateur Ham Radio (VU2 IJZ). He is currently supervising 10 Ph.D students who are working in different areas of image processing & Artificial Intelligence. He has published 35 papers in international journals and conferences.

N.S.S.S.N Usha Devi was a graduate student of C.S.E from J.N.T.U. She has published 5 research papers in international conferences and two in international journals. Her interests include Cellular Automata and Adhoc Networks. She was the member of IAENG (U.S.A),. ICST (Europe). She was also Associate Editor of Journal of Software Engineering (ISSN: 1994-5450).