International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-1, November 2019

Line Segmentation Challenges in Tamil Language Palm Leaf Manuscripts

R. Spurgen Ratheash. M. Mohamed Sathik

Abstract: ***The process of an Optical Character Recognition (OCR) for ancient hand written documents or palm leaf manuscripts is done by means of four phases. The four phases are 'line segmentation', 'word segmentation', 'character segmentation', and 'character recognition'. The colour image of palm leaf manuscripts are changed into binary images by using various pre-processing methods. The first phase of an OCR might break through the hurdles of touching lines and overlapping lines. The character recognition becomes futile when the line segmentation is erroneous. In Tamil language palm leaf manuscript recognition, there are only a handful of line segmentation methods. Moreover, the available methods are not viable to meet the required standards. This article is proposed to fill the lacuna in terms of the methods necessary for line segmentation in Tamil language document analysis. The method proposed compares its efficiency with the line segmentation algorithms work on binary images such as the Adaptive Partial Projection (APP) and A* Path Planning (A*PP). The tools and criteria of evaluation metrics are measured from ICDAR 2013 Handwriting Segmentation Contest.***

Keywords: line segmentation, Tamil palm leaf manuscripts, connected component, historical documents, Tamil character recognition.

I. INTRODUCTION

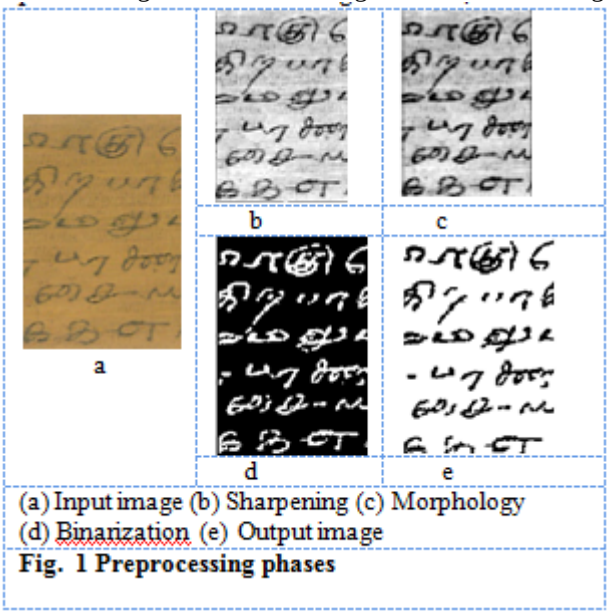
In digitizing palm leaf manuscripts, there are various challenges in terms of reading and understanding the scripts. Only scholars with knowledge in old scripts could read and understand the palm leaf manuscripts. However, reading this poses great challenge for the general people and researchers concerned.

Those eminent scholars who could read the palm leaf manuscripts have incorporated the information in the palm leaf manuscripts in printed form as books. In spite of the effort taken so far in preserving the palm leaf manuscripts and the information in it, there is still a long way to go in accumulating the entire essential information from the ancient archives. In order to grasp all the information from palm leaf manuscripts, it is of vital importance to recognize the information from the manuscripts in question. Automatic recognition is a process that helps in reading the scripts with suggestions when the scripts are not recognizable. It is important to implement automatic recognition of the palm leaf manuscripts as the structure or form of the modern day letters is different from that of the ancient scripts.

Revised Manuscript Received on November 30, 2019.
***** Correspondence Author
R. Spurgen Ratheash*, Assistant Professor, Department of Information Technology, MCA, Department of Computer Applications, Bishop Heber College, Bharathidasan University, Trichy, India.
M. Mohamed Sathik, PG and Research, Department of Computer Science, Sadakathullah Appa College, Tirunelveli, Tamilnadu, India. Manonmaniam Sundaranar University, Tamilnadu, India.

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/

Optical Character Recognition is a process that recognizes the character in the ancient scripts automatically with high accuracy. Basically an OCR has five major phases such as 'preprocess', 'line segmentation', 'word segmentation', 'character segmentation' and 'character recognition'. The first and foremost phase Pre-process does the initial work such as noise removal, morphology, binarization to separate the dark background and text foreground as shown in Fig.1.



(a) Input image (b) Sharpening (c) Morphology
(d) Binarization (e) Output image

Fig. 1 Preprocessing phases

The second phase, line segmentation segment the text lines if it is touching and overlapping with the subsequent lines. The third phase, separate the words according to the language from the segmented lines. The character segmentation separates the characters from the words and leads to the final phase of character recognition. Amongst the umpteen of methods for line segmentation, this article concentrates on two methods such as Adaptive Partial Projection (APP) and A* Path Planning (A*PP) which are considered the best in Thai and Khmer scripts respectively. The process of those methods is implemented in Tamil palm leaf manuscripts and it compares the result with the proposed segmentation method.

In section II provide the details of widespread line segmentation methods used in various languages by Literature survey. The comparison of Thai and Khmer language manuscripts line segmentation methods with Tamil language manuscripts explains in Section III, and the evaluation results in section IV with the conclusion in section V.

II. LITERATURE SURVEY

The text line segmentation is performed in three ways: the connected component estimation method is used to group the component for top-down method, vertical projection provides bottom-up approach, nearest neighbour detection is used in the third method on historical documents [1]. Connected components are calculated and classified large components that identify the long ascenders or descenders. The later connect multiple lines and smooth the histogram with Gaussian Kernel with the mean of its height. Further, it derivates and removes the lower intensity components than threshold in historical documents [2][3]. The projection profile partitioning is used to segment the lines by separating the Vertical strips and Horizontal run for each strip of an image [4][5]. The Hidden Markov Model is built by dividing the text line image and apply the Viterbi algorithm to detect possible paths to segment the lines [6][7]. The competitive learning algorithm applied on center of mass of y-coordinates derived from 1-D vectors [8]. The connected components are labelled to form a bounding box around the text with the measure of height and width. The threshold height value is fixed by mean and deviation and is compared with the height of the text area and the greater value has broken into two lines [9]. The Fringe Map is generated on the binary images which lead to identify Peak Fringe Numbers (PFN). All the identified PFNs are put together to separate the text lines [10]

III.COMPARATIVE STUDY

Many of the algorithms are efficient to segment the lines in historical documents and palm leaf manuscripts of Thai, Khmer, and Hindi languages. However, umpteen of algorithms are available to segment the lines in Tamil language printed documents incompetent for Tamil palm leaf manuscripts. This chapter summarizes the comparison of two efficient line segmentation methods in Thai and Khmer manuscripts with the proposed method in Tamil manuscripts.

A. ADAPTIVE PARTIAL PROJECTION

In Adaptive Partial Projection line segmentation method for Thai language manuscripts, the line numbers, average line position, average height of the line is calculated from the horizontal projection profile that derived from the piece wise projection method [11]. The peaks are identified by applying the histogram on the image and defined the number of lines as against the peaks. The image divides into vertical column. In each column, the horizontal projection profile is calculated and the average filtering is applied more than one time to smooth the histogram as well as eliminate the spurious peaks and valleys in the projection. The lowest value between two peaks is known as valleys that define the base line of characters placed in the column. The segmentation line is formed by joining all the characters' base line. If a base line overlaps one or more connected components, the column is divided into two and processed further when the width of the column is less than the width of the character. [12].

B. A* PATH PLANNING

A* Path Planning has been applied in artificial intelligence to find the path for robotic system, route planners, and games. A* Path Planning uses heuristic function for quick computation to find out an optimal solution to reach the end state. A*PP algorithm cannot reach the end when all sides are covered by obstacles. In handwritten historical documents of Khmer language, the path identification is effective when two lines do not touched or overlap. An algorithm malfunctions when are touched or overlapped two lines. It is solved by cost functions in the path planning algorithm such as distance cost function, map obstacle cost function, vertical cost function, the path planning is successful in line segmentation only and neighbour cost function. In spite of A*PP has the above-mentioned cost functions the path planning is successful in line segmentation only when they are partially overlapped in the handwritten historical documents [13].

C. PROPOSED METHOD

Line segmentation in Tamil palm leaf manuscripts is a Herculean task. In order to simplify the process, the text lines are partitioned three zones according to their characteristics namely 'text zone', 'upward elongated zone', and 'downward elongated zones' as in Fig. 2.

Fig. 2. Text line characteristics

Text zone is an actual character existing zone [14]. The strokes of Tamil letters, in the course of writing, go beyond the text zones such as 'upward' and 'downward' elongated zones.

An extension from the text zone by upward or/and downward characters in Tamil palm leaf manuscript is known as an 'obstacle'. It is categorised by 'presence of' and 'absence of' an obstacle. The first category makes the line segmentation as challengeable because considering other language manuscripts, the presence of obstacles can be identified only in vowels or in consonants. In Tamil language manuscripts, the presence of an obstacle may exist in any characters. Prior to the line segmentation process, the presence of an obstacle in text lines are classified in two ways by their nature of extension. If it touches with the subsequent lines, they are called 'touched obstacle text lines' and when the obstacle reaches up to the middle of the following or succeeding lines, they are defined as 'overlapped obstacle text lines' as shown in Fig.3. The second category is known as 'independent text lines' that segments the lines automatically without any complication by an identification of the character using connected component.

Fig. 3.	Touched obstacle text line	Overlapped obstacle text line	Independent text line

The APP method described in section 3.1 is one of the best methods for Thai language palm leaf manuscripts. This provides accuracy in independent characters that are placed in the text line. If the lines are touched and there is less space between the two lines, the result of APP line segmentation provides a by and large result. The A*PP method in section 3.2 is one of the best heuristic ways of approach in line segmentation of handwritten documents.

The path planning algorithm is applied on Bali, Sunda, Khmer palm leaf manuscripts provide a better result in 'touched characters' with minimum time. Both the algorithms are applied on binary images provide notable results in touched lines. Considering the overlapping lines, the above-mentioned methods are not effective as such. When the above-said methods are applied in Tamil palm leaf manuscripts, the result of APP is fairly reported in the image of independent scripts as in Fig.4. The character '□□/Lu/', 'ன/La/', and 'ன/La/' are placed first, second and third lines of the image respectively. The APP method segments the first line as three line spaces such as S1, S2, and S3. It furthers the cutting edge that breaks and changes the structure of the character '□□/Lu/' as 'ன/La/'. The 'cutting edge' is defined as the point where the lines sever from its subsequent lines. S4 denotes the line space between the second and the third text lines. An 'absence of' obstacle in APP provides fair results. Besides, the method is not much effective in touched and less line space images of Tamil palm leaf manuscripts. The A*PP method provides good results with the independent lines as shown in Fig.5.

Fig. 4. a. Original Image **b. APP segmentation Line**

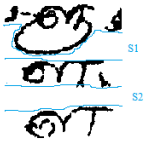


Fig. 5 A*PP Line segmentation for independent lines

However, in touched lines of Tamil palm leaf manuscripts, the path planning falls short in reaching the end state. The cutting point of an algorithm changes the characters structure. It also creates a continuous process without reaching the end state when more than two lines are touching each other in Fig.6. The character '□□/thy/', 'ந/na/' and 'த/tha/', '□□/ye/' are in the first, second, third lines respectively. ES1 specified the 'End State' of the first line and ES2 for the second. S1 and S2 are the line space between the text lines. The A*PP segmentation proceeds the first line character 'thy' can segment without any difficulties. Considering the second line, the path planning creates a loop and it reaches the end state ES2 without segmenting the text line. If an APP proceeds, the cutting edge segments the line and changes the character 'ந/na/' into unpredictable structure and 'த/tha/' into wrongly predicted 'க/ka/'.

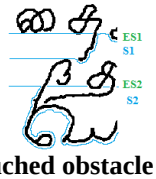


Fig.6 A*PP for Touched obstacle text lines

The proposed method provides one of the best results in three ways such as independent text lines, touched and overlapped obstacle text lines. The Connected Component (CC) algorithm analyses the quality of extracted group and the quantitative measures of grouping quality within certain constraints. It relates to a random image model, and set figure and background. The probability of the features of the image belongs to the figure is P_f and the background is P_{bg}. The process considers binary cues. Value 1 is assigned where P_f and P_{bg} are in the same group, when otherwise it is 0. The probability of finding the connected point in the background pixel is $P_c(L)$ with the distance L, that may be connected with the figure and the background. The density of such points implies with the error probability E_f [14].

$$P_{bg}^{(1)}(L) = P_{bg}\left[1 - \sum (1 - E_f)^m P_f^m (1 - P_f)^{k(L)-m} C_{k(L)}^m\right]$$

The total number of connected features is

$$N_f^{(i)} = N_{bg} N_f^{(i-1)} \qquad N_{bg} < 1$$

The presence of character in the text lines are measured by $N_f^{(i)}$ through the binary cues of an image. In image matrix, when the value is drastically changed and extends with minimum value, it is defined as an obstacle of the character. If an extended minimum value ends with greater value that identified as touched obstacle text lines otherwise independent text lines. The process also predicts an overlapped obstacle text lines when the value is maximum than all other characters are present in the text lines. The proposed method step by step process algorithmic way in the following Algorithm.1

Algorithm.1

```
// Input Image
Input:
    In_Img : Binary Image
//Output Image
Output:
    Out_Img : Line Segmented Image
//Variables
VT ← Vertical Space Track
HT ← Horizontal Space Track
M ← In_Img Width
N ← In_Img Height
//Sum of column to calculate Vertical space
for ( k = 1; k<= N; k++) do begin
    Vs ← sum of zero from In_Img(k, 1:20) Location
    if (Vs > Threshold)
            VT_k = assign 1
    else
            VT_k = assign 0
    end
end for
//Sum of row to calculate Horizontal space
for ( k = 1; k<= N; k++) do begin
    if (VT_k ==1)
            Hs ← sum of zero from In_Img(k, 1:M)
    if (Hs ==0)
            HT_k = assign 1 (Space Found)
    else if (Hs < Threshold)
            HT_k = assign 2 (Obstacle Found)
    else
            HT_k = assign 3 (space Not Found)
```

31. **end**
32. **end**
33. **end for**
34. $Out_{Img} \leftarrow HT_k$
35. Return (Out_{Img})

In image matrix, sum of the value k presents in the row to predict whether the background or character exists within the 20 positions against N and assign the sum value into Vs. The vertical space identifies by zero which is otherwise identified as the horizontal space by the same way against M and assign the sum value into Hs as shown in Fig. 7.

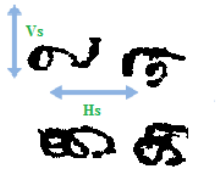


Fig. 7. Vertical and Horizontal space

The vertical space identifies by VT_k, horizontal space HT_k to predict three categories such as 'space between two lines', 'obstacle is present in the space' and 'no space'. An optimality of the performance for touched and overlapped obstacle text lines of Tamil palm leaf manuscripts are implemented by fixing the cutting edge in the place of minimum value exists in HT_k with the consideration of VT_k between the text lines.

IV. EXPERIMENTAL EVALUATION

The line segmentation algorithms specified in section 3 is applied on 750 lines of the Tamil palm leaf manuscripts of 225 images. This matches the resultant images with the ground truth images to evaluate the metrics of DR, RA, and FM based on MM, NN, and o2o. These evaluation criteria and tools are provided by ICDAR 2013 Handwritten Segmentation Contest.

An evaluation of this article by the metrics of one-to-one (oTo) match score, NN and MM [15]. oTo is computed for a region pair based on the evaluator's acceptance threshold. Let NN be the total number of ground truth elements and MM be the total number of result elements. An above said three metrics are calculated with the oTo score. The Detection Rate (DR) is defined by

$$DR = \frac{o2o}{NN}$$

Recognition Accuracy (RA) is

$$RA = \frac{o2o}{MM}$$

and F-measure (FM) is calculated by

$$FM = \frac{2.DR.RA}{DR+RA}$$

In Tamil palm leaf manuscript images the text lines have the challenges such as Low contrast (LC), Damaged Letters (DL), Overlapped Lines (OL), Low Space (LS), Cross Lines (CL), Touched Lines (TL), and Independent Lines (IL). The performance on various challenges of text lines are shown in Table.1 – Table. 3. The performance metrics of DR, RA, FM on Hundred various kinds of images shows in Fig.8 – Fig.10.

V. CONCLUSION AND FUTURE WORK

The present article sheds light on the line segmentation work of 'touched' and 'overlapped text lines' of Tamil palm leaf manuscripts. In this article, the comparison of line segmentation methods are working on binary images among umpteen of methods. The proposed method provides optimality about the ins and outs of line segmentation in the Tamil palm leaf manuscripts with 95% of Recognition Accuracy while APP and A*PP provides 51% and 89% from the same value of NN. The Table.4 provides an overall performance of evaluation metrics with the line segmentation methods such as APP, A*PP and Proposed method. The structure of the Tamil character may break when the proposed line segmentation is implemented on Tamil palm leaf manuscripts. All said and done, the proposed method is successful in bringing out the very existence of lines.

Table 1 Performance of DR in different kind of challenges

DETECTION RATE (DR)	LC	DL	OL	LS	CL	TL	IL	OVERALL
APP $_{DR}$	30.64	51.79	35.66	37.18	36.88	44.23	40.66	39.57
A*PP $_{DR}$	66.66	87.53	83.89	78.67	87.42	85.43	84.56	82.02
PROPOSED $_{DR}$	70.63	89.33	91.54	85.94	91.09	88.94	92.25	87.1

Table 2 Performance of RA in different kind of challenges

RECOGNITION ACCURACY(RA)	LC	DL	OL	LS	CL	TL	IL	OVERALL
APP $_{RA}$	45.7	65.16	49.69	51.02	51.97	60.12	56.98	54.38
A*PP $_{RA}$	79.17	86.15	89.04	85.28	87.45	88.9	90.34	86.62
PROPOSED $_{RA}$	98.86	90.69	94.64	94.73	91.72	93.68	96.46	94.4

Table 3 Performance of FM in different kind of challenges

F-MEASURE (FM)	LC	DL	OL	LS	CL	TL	IL	OVERALL
APP $_{FM}$	36.63	57.68	41.39	42.91	43.09	50.92	47.42	45.71
A*PP $_{FM}$	72.33	86.74	86.37	81.75	87.36	87.12	87.34	84.14
PROPOSED $_{FM}$	82.14	89.94	93.05	89.6	91.31	91.14	94.27	90.2

Table 4. Performance of evaluation metrics

	ALGORITHMS		
Metrics	**APP**	**A * PP**	**Proposed**
NN	27058	27058	27058
MM	18586	25650	25485
oTo	9762	22895	24347
DR	35.94	84.64	90.12
RA	50.68	89.05	95.46
FM	41.96	86.73	92.64

Fig. 8. Image wise performance of DR

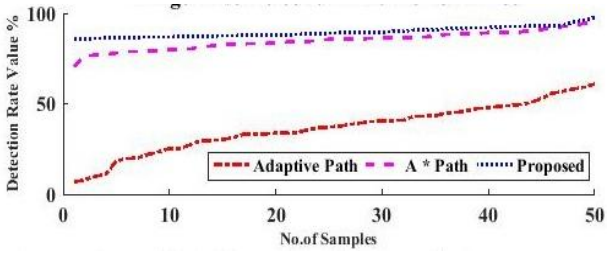


Fig. 9. Image wise performance of RA

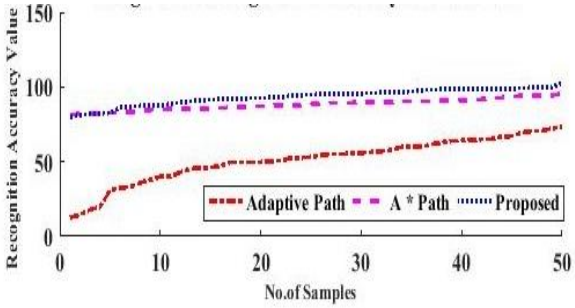
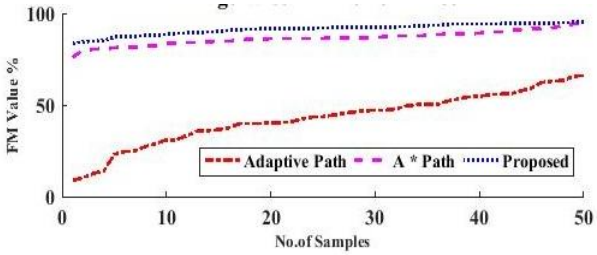


Fig. 10. Image wise performance of FM



REFERENCES

1. Ines Ben Messaoud, Hamid Amiri, Haikal El Abed, Volker Margner "A Multilevel Text line Segmentation Framework for Handwritten Historical Documents", ICFHR, IEEE, pp. 515-520, 2012.
2. Xi Zhang, Chew Lim Tan "Text Line Segmentation for Handwritten Documents Using Constrained Seam Carving", ICFHR, IEEE, pp. 98-103, 2014.
3. MadeWindu Antara Kesiman, Dona Valy, Jean-Christophe Burie, Erick Paulus, Mira Suryani, Setiawan Hadi, Michel Verleysen, Sophea Chhun and Jean-Marc Ogier, "Benchmarking of Document Image Analysis Tasks for Palm Leaf Manuscripts from Southeast Asia" J.Imaging, pp. 1- 27, 2018.
4. Papangkorn Inkeaw, Atcharin Klomsae, Sanparith Marukatat, "Lanna Dharma Handwritten Character Recognition on Palm Leaves Manuscript based on Wavelet Transform", ICSIPA, IEEE, pp. 253-258, 2015.
5. Nagendra Panini Challa, R.Vasanth Kumar Mehta, "Applications of Image Processing Techniques on Palm Leaf Manuscripts - A Survey", Helix, pp. 2013-2017, 2017.
6. Ge Peng, Pengfei Yu, Haiyan Li and Lesheng He, "Text Line Segmentation Using Viterbi Algorithm For The Palm Leaf Manuscripts of Dai", ICALIP, IEEE, pp. 336-340, 2016.
7. Ge Peng, PengFei Yu, HaiYan Li, HongSong Li, XuDong Zhu, "A Character Segmentation Algorithm for the Palm Leaf Manuscripts", ICCIA, IEEE, pp. 354-358, 2017.
8. Dona Valy, Michel Verleysen, and Kimheng Sok, "Line Segmentation Approach for Ancient Palm Leaf Manuscripts using Competitive Learning Algorithm", ICFHR, IEEE, pp. 108-113, 2016.
9. Jija Das Gupta, Bhabatosh Chanda, "A Model Based Text Line Segmentation Method for Off-line Handwritten Documents", ICFHR, IEEE, pp.125-129, 2010.
10. Vijaya Kumar Koppula, Atul Negi, "Fringe Map Based Text Line Segmentation of Printed Telugu Document Images", ICDAR, IEEE, pp. 1294-1298, 2011.
11. N. Tripathy and U. Pal, "Handwriting Segmentation of Unconstrained Oriya Text", Sadhana, Springer, pp.755-769, 2006.
12. Rapeeporn Chamchong, Chun Che Fung, "Text Line Extraction Using Adaptive Partial Projection for Palm Leaf Manuscripts from Thailand", ICFHR, IEEE, pp. 586-591, 2012.
13. Olarik Surinta, Michiel Holtkamp, Faik Karabaa, Jean-Paul van Oosten, Lambert Schomaker and Marco Wiering, "A* Path Planning for Line Segmentation of Handwritten Documents", ICFHR, IEEE, pp.175-180, 2014.
14. Alexander Berennolts, Michael Lindenbaum, "On The Performance of Connected Components Grouping", IEEE, pp.189-192, 2000.
15. R.Spurgen Ratheash, and M. Mohamed Sathik, "A Detailed Survey of Text Line Segmentation Methods in Handwritten Historical Documents and Palm Leaf Manuscripts", IJCSE, pp 99-103 , 2019.

AUTHORS PROFILE

R. Spurgen Ratheash, an Assistant Professor of Information Technology, received his MCA degree in Computer Applications from Bishop Heber College, Bharathidasan University, Trichy, India in 2007. He received his MPhil degree in Computer Science and an M. Tech in Information Technology from Manonmaniam Sundaranar University, Tirunelveli, India in 2012 and 2014 respectively. He is a research scholar at Sadakathullah Appa College, affiliated to Manonmaniam Sundaranar University, Tirunelveli, India. His major research interests include Digital Image Processing, Document Image Analysis and Character Recognition of Tamil language palm leaf manuscripts.

Dr. M. Mohamed Sathik is the Principal of Sadakathullah Appa College, Tirunelveli, India. He received two Ph.Ds majored in Computer Science and Computer Science & Information Technology in Manonmaniam Sundaranar University, Tirunelveli, India. He has many more feathers in his cap by degrees such as M. Tech, MS (Psychology) and MBA. He is pursuing post Doctoral degree in Computer Science. Known for his active involvement in various academic activities, he has attended many national and international seminars, conferences and presented numerous research papers. With publications in many international journals, he has published two books besides having guided more than 40 research scholars. The prolific academician is a member of curriculum development committee of various universities and autonomous colleges in Tamil Nadu, India. His areas of specialization are Virtual Reality, Image Processing and Sensor Networks.