Offline Tamil Handwritten Character Recognition using Zone based Hybrid Feature Extraction Technique

L. Anlo Safi
(P.G Scholar)
Dept of CSE-PG,
National Engineering College
Kovilpatti, Tamil Nadu, India.

K.G.Srinivasagan, PhD.
(Professor &Head)
Dept of CSE-PG,
National Engineering College
Kovilpatti, Tamil Nadu, India.

ABSTRACT
Character recognition is the most important research area in today's world. Many researchers have focused on recognizing handwritten digits, numerals and characters in so many languages. To the best of our knowledge, little work has been done in the area of Tamil handwritten character recognition but they did not achieve better accuracy. Feature extraction is the important phase of character recognition. Feature extraction increases the recognition accuracy. This paper presents an overview of Feature Extraction techniques for off-line recognition of Tamil characters. The proposed method is Zone based hybrid approach for feature extraction. The 55 features which are extracted from the character image are the number of horizontal, vertical, diagonal lines along with their total length for each zone. ANN classifier is used for classification and recognition purpose. We obtained a better accuracy when compared with the previous approaches for offline Tamil handwritten character recognition.

Keywords
Offline handwriting recognition, Tamil Scripts, OCR, Euler Number, ANN.

1. INTRODUCTION
Recognition means automatic conversion of image into a form that can be processed by the computer. Recognition of handwritten character is a central problem in today's world. It is able to solve more complex problem and make human job in an efficient way. One of the most important research areas in today's world is pattern analysis and machine intelligence. There are two types of character recognition are present they are handwritten character recognition and typed character recognition. In handwritten character recognition the way of writing a single text normally vary from one user to another and different styles may occurs for n number of users. The large amount of noise will be occurs in the handwritten character recognition during the writing of the text and scanning of the document. In the Typed character recognition noise will be reduced while compared to handwritten character recognition and same styles may occur for different users. Two types of handwritten character recognition are present based on the image acquisition. If the input is a scanned document from a sheet of paper is referred as offline handwritten character recognition. In this only the pixels information are present, since we cannot have constrain styles. While in the on-line handwritten character recognition the user writing has to be carried out in an electronic device. A special pen is used to write text on the special computer screen. In this writers pen movement is captured for recognition process. The constrain style are used for the easy recognition of the text. The constrain style of the input images are the starting, ending, direction of the text image. Due to less amount of information, off-line handwriting recognition is considered more difficult than on-line handwriting recognition.

Optical character recognition (OCR) is the most important tool for character recognition. The basic idea behind the OCR is to identify and analyze the input image by dividing into line and further subdivided into word and then in character. These individual characters are compared with the image pattern to predict the correct character. There are four different steps involved in OCR techniques they are Pre-processing, Segmentation, Feature extraction, Classification and Recognition. In this paper we describe each step in a detailed way. Tamil language is the most popular in more than 3countries such as Malaysia, Singapore, Srilanka etc….of the world. Tamil handwritten character recognition has always been a challenging task in pattern recognition due to the complexity in the letter structure. The Tamil language is a combination of 12 vowels and 18 consonants and one special character called aytam. So the Tamil script totally consists of 247 letters. Nowadays so many researchers have focused on handwritten character recognition in all the languages. But in Tamil handwritten character recognition only a limited research has happened and no research has produced 100% accuracy, only 80-90% achieved.

2. RELATED WORK
A number of methodologies have been developed for character recognition in past few years. Preprocessing is used to clean the document and several preprocessing techniques where explained below: Binarization is the first step in preprocessing. Karthik *et al* (2012) [3] introduced level set theory for binarization process by using the threshold values. Karthik *et al* (2012) [13] attempt to remove noise by using the run length count method. And the results compared with other noise filtering model like Gaussian, Poisson and Speckle noises. Sandeep *et al* (2012) [2] used the moment normalization to find the centre of gravity or centroid and it used to resize the image into some fixed size and then translate it to the center of the image frame.The output of preprocessing is used as input for image segmentation process. Karthik *et al* (2012) [3] proposed the active contour. Sigappi *et al* (2011) [14] proposed line segmentation based on the intensity values of the pixels.Abdullah *et al* (2012) [1] proposed the word segmentation method in which words are first divided into segments with specific width. Rajakumar *et al* (2011) [21] suggest clustering concepts for the *al* (2011) [21] suggest clustering concepts for the segmentation of character.

The next important process is the feature extraction where feature vectors are extracted for each partition. Many feature extraction techniques has been adopted and it is broadly classified into Global Feature Extraction Methods and Local Feature Extraction Methods. Rajib *et al* (2012) [5] classify three global features. Ashutosh *et al* (2012) [10] and Deepika *et al* (2012) [12] proposed Gradient Features, Projection Features and Curvature Features. Many researchers have used direction feature vector techniques [4], [6], [7], [8] in which the image is divided into vertical and horizontal windows or frames. Sigappi *et al* (2011) [14] proposed profile features which are primarily used to extract features from Tamil handwritten word images and it covers vertical projection profile, word profiles, and background-to-ink transitions. Zone based hybrid approach [9] [11] is the combination of image centroid zone and zone centroid zone of numeral/character image. A new method, called diagonal based feature extraction; is introduced by Srinivasan et *al* (2012) [17] for extracting the features of the handwritten alphabets. YusufPerwej *et al* (2012) [20] introduced Characteristic Loci Feature extraction to extract feature. It is used to reduce the dimensions of the character. A new feature extraction technique proposed by Dayashankar *et al* (2010) [22] is used to calculate only twelve directional feature inputs depending upon the gradients.

[1] [2] [4] [5] [7] [14] used Hidden Markov Models for the final phase of recognition. Peyarajan *et al* (2011) used Kohonen's Neural Network which is also known as Self Organizing Map. Rakesh *et al* (2012) [8] used a special classifier for recognizing the handwritten Devanagari vowels by means K-NN (K- Nearest Neighbour). K-NN classifier is functioned for the learning and the testing phases. Ashutosh et al (2012) [10] and Deepika et al (2012) [12] used Support Vector Machine (SVM) for classification process. Support vector machines (SVM) are a group of supervised learning methods that can be applied to classification or regression. Several more researches used the designed multi-layer back propagation neural network (BPNN) classifier [11] [17] [20] [22] [23] [24][25] consisting of three layers- the input, hidden, and output. Multilayer neural network employs back propagation algorithm. Ranpreet *et al* (2011) [18] proposed a hybrid algorithm of Back Propagation and Genetic algorithm to train and test the network.

3. IMAGE ACQUISITION AND PREPROCESSING

The input image have collected from 50 persons with different age groups and stored in .tiff/.jpg format. More than 5000 image are present in our database. In offline handwritten character recognition the input image is the scanned document from a sheet of paper. The main objective of pre-processing is to clean the document image. The input image is the colored RGB image. In Binarization process the RBG image is converted into the gray scale images and further it converted into binary images.

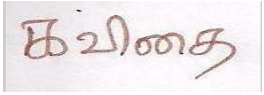
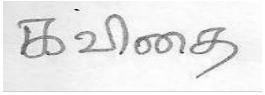

Fig 1: Original image **Fig 2: Gray scale image**

The samples of Tamil handwritten character are presented in Figure 1.The colored image is in the range of 0 to 255. The following condition check whether it is colored image or not.

Size (Original image, 3) ==3.

The pixel values may in the range of 0 to 255 in Figure 2, where 0 represents the black image and 255 represent the white image. The conversion of the grayscale image to a binary image is shown in Figure 3. The most common method is to fix the threshold value based on the image. Replace all pixels in the input image with luminance greater than level with the value 1 and replaces all other pixels with the value 0. The threshold value should be in the specified range of [0, 1]. To compute the level argument by use the functions gray thresh. Level is not mentioned then the im2bw use 0.5 by default.

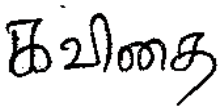
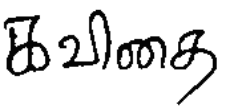

Fig 3: Binary image **Fig 4: Denoised image**

The noise will occur during the scanning of the input image and while writing the text. So many filtering techniques are present in modern world. Here the Median filtering is used to remove the unwanted patterns. Figure 4 shows the input image after applying the filter. The median filter is better while compared with other filters in preserving the edges of the character. The way of writing the character should be varied from one pen to another so it will difficult to segment the character. But Morphology based thinning algorithm is used to produce a single pixel width image and can remove the irregularities in character. The single pixel image is presented in Figure 5.

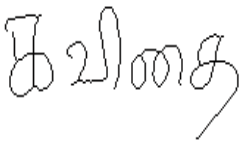

Fig 5: Thinning

4. SEGMENTATION

After the pre-processing steps, the difficult task in OCR techniques is Script Segmentation. It is a process which is used to split the word image into individual character. Segmentation of handwritten character is more complex than the typed character. Region probe algorithm is used to get individual character from the image. In region probe algorithm the boundary values are calculated based on continuous ranges of the intensity value. Figure 6 shows the segmented result with different size based on the character size. Before going into the feature extraction phase convert the segmented character into a normalized size. The standard size provided by is 55*55 for all the segmented character.

Fig 6: Segmentation

5. FEATURE EXTRACTION

The objective of feature extraction is to capture the essential symbols from the handwritten document. But it is one of the important problems in pattern recognition. Feature

extraction is a process of extracting a set of feature from the individual character. These methods provide the ease of implementation and good recognition. The feature may be the height, width, horizontal line, vertical lines etc of the character. The feature vector explained here is based on the direction and it should be as follows:

- The number of horizontal lines
- The total length of horizontal lines
- The number of right diagonal lines
- The total length of right diagonal lines
- The number of vertical lines
- The total length of vertical lines
- The number of left diagonal lines
- The total length of left diagonal lines
- The number of intersection points

So if there are N zones, there will be 9N elements in feature vector for each zone. For the system proposed, the original image was first divided into 1x3 zones in vertical direction. Then 9 features were extracted for each zone. Here 27 features are extracted from the 1x3 zones. Again the original image was divided into 3x1 zones by dividing in the horizontal direction and obtain 27 features were extracted for each such zone. So 27+27=54 features are extracted for each character set. The number of the line type is calculated by using the formula is

Value = (1-(number of lines/10)*2)

The total length is calculated by using

$$\text{Length} = \frac{\text{Total pixel present in that line type}}{\text{Total pixels belonging to skeleton}}$$

The 54 feature values for the single character are [0.8, 0.6, 0.8, 1, 0.2545, 0.5455, 0.0727, 0…, and 0.0963]. Here first value is the Number of horizontal line type, is calculated as [1((1/10)*2)] =0.8 i.e. number of horizontal lines present in that zone is 1. The value 1 represent there is no line type is present and the length is denoted as zero for such line type. In some cases the number of line in the exacting zone is greater than 5 then negative value get display for the particular line type. The intersection point is that it should have more than one neighbor and it calculated by the same procedure for the line type. After zonal feature extraction, one feature are extracted for the entire image based on the regional properties namely

- Euler Number: It is defined as the difference of Number of Objects and Number of holes in the image. The Euler number for each character is shown as


1-2 = -1 (number of holes two)


1-0 = 1 (no holes is present)


1-2 = -1 (number of holes two)


1-2 = -1 (number of holes two)

The total feature extracted from the zone based hybrid technique is 27+27+1=55.

6. CLASSIFICATION AND RECOGNITION

The classification stage is the decision making part of a recognition system and it uses the features extracted in the previous stage. A feed forward back propagation artificial neural network classifier is used to classify the unknown character into a known character.The neural net approach contains two phases. The first phase is used to train the network by providing various input samples. Based on the trained data the accuracy will get increased. The second phase is used to test the input samples with the use of the trained data. The number of input nodes is chosen based on the number of features and the total numbers of characters n determines the number of neurons in the output layer. The number of neurons in the hidden layers is obtained by trial and error. The hidden layers use log sigmoid activation function.

7. EXPERIMENTAL RESULTS

The recognition system has been implemented using Matlab7.1. The proposed algorithms are tested on handwritten character database written by 50 users. The database used in the experiment has 200 samples and is divided into training and testing each consists of 100 samples. The structure of neural network includes an input layer with 55 inputs, two hidden layers each with 100 neurons and an output layer with 4 neurons. The gradient descent back propagation method with momentum and adaptive learning rate and log-sigmoid transfer functions is used for neural network training.

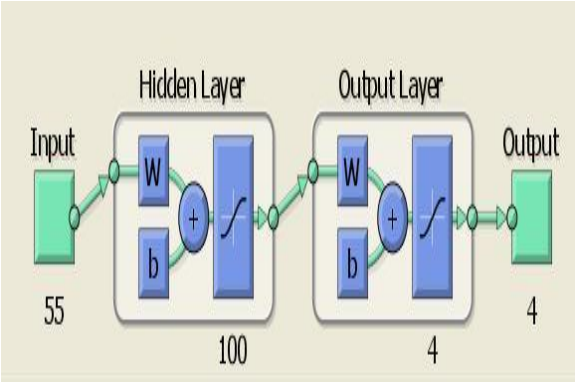

Fig 7: Feed forward back propagation neural network architecture

The network training parameters are:
Input nodes: 55
Hidden nodes: 100 each
Output nodes: 4 (character)
Training Algorithm: Gradient Descent with Momentum
Training and Adaptive Learning Perform function: Mean Square Error Training Goal Achieved: 0.000001
Training Epochs: 100000
Training Momentum Constant: 0.9
Training Learning Rate: 0.01

Figure 8 shows the Error (MSE) vs. Training Epochs performance of the network with 55 features obtained though horizontal, vertical and diagonal extraction. It can be noted that it requires 48 epochs to reduce the mean square error to the desired level. The experimental results have tested with different hidden layer and the performance measures have decreased while the hidden layer increases.

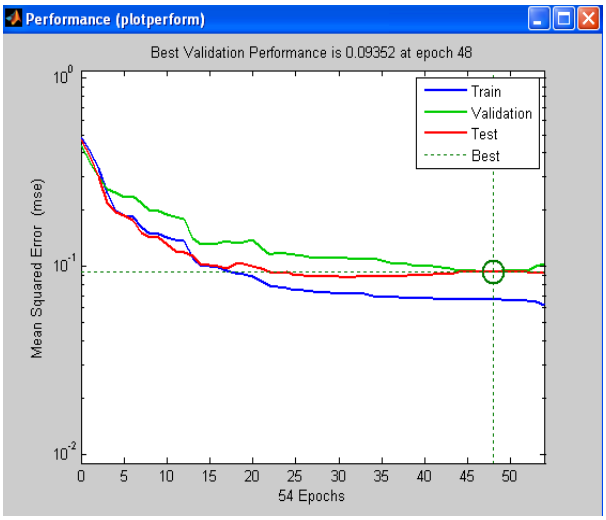

Fig 8: The variation of MSE with training Epochs for 55 features.

Here, Sample1=  , Sample2= 

Sample3=  sample4= 

TABLE I

RESULT OBTAINED

Character	**Recognition %**	**Mis-Recognition%**
Sample 1	100	0
Sample 2	98	2
Sample 3	96	4
Sample 4	98	2
Total	**98**	**2**

TABLE II

COMPARISION ANALYSIS OF FEATURE EXTRACTION

S. No	**Feature**	**Classifier**	**Accuracy**
1.	Boundary Tracing	MLP	94.1%
2.	Stroke Feature	MLP	90.5%
3.	Zernike Moments	MLP	95.3%
4.	**Zone based directional information**	**MLP**	**98%**

We have experimental in our system through 50 samples for each character such as 'f' 'tp' 'i' 'j' (support by Bamini font).The Table 1 illustrates that the detailed recognition results for 200 samples. The recognition rate of 100%, 98%, 96%, and 98% was achieved for the Sample1, Sample2, Sample3 and sample 4. Te error rate of 0%, 2%, 4%, 2% was occurred for sample 1, sample2, sample3 and sample4. The best performance measures is 0.09362 has been obtained in 48iteration with the minimum computation time of 0:00:02 sec. In Table 2 shows that the detailed comparison of feature extraction.

8. CONCLUSION

Digital recognitions are playing wide role and providing great scope to perform research in handwritten character recognition. Still there have been many challenges and issues. So it is obvious need to strengthen the existing one. Many researchers have contributed many solutions for resolving the issues. The proposed work has the novel solution for performing the character. This work is mainly used the zone based method for feature extraction. The feature values are used to identify the edge point of the character. Based on the feature extraction only the accuracy may get increased. The experimental results reveal that 55 features give better recognition accuracy. From the test results it is identified that the directional information of feature extraction yields the highest recognition accuracy of 98 % for 55 features. This work has surely to significant enhancement than the existing work.

9. REFERENCES

[1] Manal A.Abdullah,Lulwah M. Al-Harigy, and Hanadi H. Al-Fraidi "Off-Line Arabic Handwriting Character Recognition Using Word Segmentation"- journal ofcomputing, volume 4, issue 3, March 2012, ISSN 2151-9617.

[2] Sandeep B. Patil, G.R. Sinha and Kavita Thakur "Isolated Handwritten Devnagri Character Recognition using Fourier Descriptor and HMM"- International Journal of Pure and Applied Sciences and Technology,volume 8(1) (2012), 69-74.

[3] S. Karthik, Hemanth.V.K, V. Balaji, K. P. Soman "Level Set Methodology for Tamil Document Image Binarization and Segmentation"-International Journal of Computer Applications Volume 39– No.9, February 2012.

[4] M. Amrouch, Y. Es-saady, A. Rachidi, M. El Yassa, D. Mammass "Handwritten Amazigh Character Recognition System Based on Continuous HMMs and Directional Features"- International Journal of Modern Engineering Research Vol.2, Issue.2, Mar-Apr 2012.

[5] Rajib Lochan Das , Binod Kumar Prasad and Goutam Sanyal "HMM based Offline Handwritten Writer Independent English Character Recognition using Global and Local Feature Extraction"- International Journal of Computer Applications Volume 46–No.10, May 2012.

[6] Mithun Biswas ,Ranjan Parekh "Character Recognition using Dynamic Windows"- International Journal of Computer Applications Volume 41–No.15, March 2012.

[7] S. V. Halse and Maheshwari S. Hiremath "Hand written Character recognition"-,Journal of Computer and Mathematical Sciences Vol. 3, Issue 1, 29 February 2012.

[8] Rakesh Rathi, Ravi Krishan Pandey and Mahesh Jangid "Offline Handwritten Devanagari Vowels Recognition using KNN Classifier"- International Journal of Computer Applications Volume 49– No.23 July 2012.

[9] Gita Sinha Rajneesh Rani Renu Dhir , "Gurmukhi Numeral Recognition using Zone based Hybrid Feature Extraction Techniques"- International Journal of Computer Applications Volume 47– No.21, June 2012.

[10] Ashutosh Aggarwal ,Rajneesh Rani, RenuDhir, "Reconition of Devanagari Handwritten Numerals using Gradient Features and SVM"- International Journal of Computer Applications Volume 48– No.8, June 2012.

[11] Ashoka H.N, Manjaiah D.H, Rabindranath Bera, "Feature Extraction Technique for Neural Network Based Pattern Recognition" International Journal on Computer Science and Engineering (IJCSE) Vol. 4 No. 03 March 2012.

[12] Deepika Wadhwa, Karun Verma "Online Handwriting Recognition of Hindi Numerals using Svm"- International Journal of Computer Applications Volume 48– No.11, June 2012.

[13] Karthik S, Mamatha H.R, Srikanta Murthy K "An Approach based on Run Length Count for Denoising the Kannada Characters"- International Journal of Computer Applications Volume 50– No.18, July 2012.

[14] AN.Sigappi,S. Palanivel,V. Ramalingam "Handwritten Document Retrieval System for Tamil Language"- International Journal of Computer Applications Volume 31– No.4, October 2011.

[15] Peyarajan ,R. Indra Gandhi, "On-line Tamil hand written character recognition using Kohonen Neural Network" - S An International Journal of Computer Systems Engineering, Vol 02, Issue 02, July 2011.

[16] J.ASHOK, DR.E.G.RAJAN, "Off-Line Hand Written Character Recognition Using Radial Basis Function"- Int. J. Advanced Networking and Applications Volume: 02, Issue: 04, Pages: 792-795 (2011)

[17] J.Pradeep, E.Srinivasan and S.Himavathi, "Diagonal based feature extraction for handwritten alphabets recognition system using Neural Network"- International Journal of Computer Science & Information Technology (IJCSIT), Vol 3, No 1, Feb 2011.

[18] Ranpreet Kaur, Baljit Singh, "A Hybrid Neural Approach For Character Recognition System"- International Journal of Computer Science and Information Technologies, Vol. 2 (2) , 2011.

[19] Nirase Fathima Abubacker, Indra Gandhi Raman, "An Approach for Structural Feature Extraction for Distorted Tamil Character Recognition"- International Journal of Computer Applications Volume 22– No.4, May 2011.

[20] Yusuf Perwej, Ashish Chaturvedi, "Machine Recognition of Hand Written Characters using Neural Networks"- International Journal of Computer Applications Volume 14– No.2, January 2011.

[21] S.Rajakumar, Dr.V.Subbiah Bharathi, "Century Identification and Recognition of Ancient Tamil Character Recognition"- International Journal of Computer Applications,Volume 26– No.4, July 2011.

[22] Dayashankar Singh, Sanjay Kr. Singh, Dr. (Mrs.) Maitreyee Dutta, "Hand Written Character Recognition Using Twelve Directional Feature Input and Neural Network"- International Journal of Computer Applications Volume 1 – No. 3, 2010

[23] Anita Pal,Dayashankar Singh, "Handwritten English Character Recognition Using Neural Network"- International Journal of Computer Science & CommunicationVol. 1, No. 2, July-December 2010.

[24] C.Sureshkumar, Dr.T.Ravichandran, "Handwritten Tamil Character Recognition and Conversion using Neural Network"- International Journal on Computer Science and EngineeringVol. 02, No. 07, 2010.

[25] Dharamveer Sharma, Deepika Gupta, "Isolated Handwritten Digit Recognition using Adaptive Unsupervised Incremental Learning Technique"- International Journal of Computer Applications Volume 7– No.4, September 2010