Handwritten Tamil Character Recognition

Dr.Amitabh Wahi
Professor
Dept of Information Technology
Bannari Amman Institute of Technology
Sathyamanagalam, Erode,Tamilnadu
Email Id: awahi@bitsathy.ac.in

Mr.Sundaramurthy.S
Associate Professor
Dept of Information Technology
Bannari Amman Institute of Technology
Sathyamanagalam, Erode,Tamilnadu
Email Id:
sundaramurthys@bitsathy.ac.in

Poovizhi P
PG Scholar
Dept of Information Technology
Bannari Amman Institute of Technology
Sathyamanagalam, Erode,Tamilnadu
Email Id:
Poovizhip.se12@bitsathy.ac.in

Abstract— **Optical Character Recognition systems have been effectively developed for the recognizing the printed characters of many non-Indian languages like English, Chinese. At early stages few research works were carried out for recognizing the handwritten characters and now varies efforts are on the way for the development of efficient systems for recognizing the Indian languages, especially for Tamil, a south Indian language widely used in Tamilnadu, Pudhucherry, Singapore, Srilanka. In this paper, an OCR system is developed for the recognition of basic characters in handwritten Tamil language, which can handle different font sizes and font types. Hu's invariant moments and Zernike moments which have been used in pattern recognition are used in this system to extract the features of handwritten Tamil characters. Neural classifiers have been used for the classification of Tamil characters.**

Keywords—**Handwritten; Feed Forward Back Propagation Classifier; Feature Extraction; Neural Network; Tamil Characters**

I. INTRODUCTION

Handwritten character recognition is the area of research for the past few decades and there is a large demand for OCR on handwritten documents. Even though, sufficient studies have performed in foreign scripts like Chinese, Japanese and Arabic characters [1], only a very few work can be traced for handwritten character recognition of Indian scripts. Even now no complete hand written text recognition system is available in Indian scenario and it is difficult due to large character set of Indian languages and the presence of vowel modifiers and compound characters in Indian script. Some reports have appeared for isolated handwritten characters and numerals of a few Indian languages. Majority of them was based on Bangla and Devanagiri script [2]. Nowadays, Government of India is taken initiation towards development of language technology. Commercial systems are developed for some Indian scripts namely Assamese, Bangla, Devanagiri, Malayalam, Oriya, Tamil and Telugu, but that can handle only printed text, not handwritten manuscript. This study focuses mainly on offline handwritten character recognition of South Indian languages, namely, Tamil [21].

India has 22 languages. These languages are written using only twelve scripts. Devanagiri script used to write Hindi, Konkani, Marathi, Nepali, Sanskrit, Bodo, Dogri and Mathili. Sindhi is written using Devanagiri script in India and Urdu script in Pakistan. Assamese, Manipuri and Bangla languages are written using Bengali script. Gurmukhi script is used to write Punjabi language. All other languages have their own script. The upper and lower case is not present in Indian language scripts. Most of the Indian languages are derived from Ancient Brahmi and are phonetic in nature and hence writing maps sounds of alphabets to specific shapes. All these languages, except Urdu, are written from left to right. The basic characters comprises of vowels and consonants. Two or more basic characters are combined to form compound characters.

Moment based features are a widely used tool for character recognition. The moment invariants are under translation, rotation and scaling is first introduced by Hu in 1962. However, Hu's moments contain much redundant information about a character's shape. Nowadays, Zernike moments are becoming popular for character recognition. Because of local-tuned neurons, Feed Forward Neural Network has fast training/learning rate. Due to this advantage, Zernike moments are used in the field of character recognition. The motivation of this work is purely based on the application of moment features and neural classifiers in the character recognition. In this paper we have presented an OCR system for basic Tamil characters, feature extraction is performed using moments and Feed Forward neural networks are used as a classifiers. Zernike moments are considered for feature extraction.

II. LITERATURE SURVEY

Diagonal feature extraction scheme for off-line character recognition is proposed by Pradeep.J et al [5]. This paper tells that each character of size 90*60 pixels is divided into 54 equal zones, each zones have the size of 10*10 pixels. By moving along the diagonals, the features are extracted from each of the zone. Each zone has 19 diagonal lines and there will be some foreground pixels. The diagonal lines are summed to get a single sub-feature. These 19 sub-features are averaged to form a single feature. This process is repeated for all the zones. Totally, 54 features are extracted for each character. Taking average on row-wise (9) and column-wise (6), as a result, every character is represented by 69 features. A feed forward back propagation neural network having two hidden layers with architecture of 54-100-100-38 is used to perform the classification.

H. Swethalakshmi et al used sequences of strokes for the feature extraction [6]. The feature extraction is performed for the devanagiri and telugu scripts. Single Recognition Engine Approach, Multiple SVM, HMMs; these three approaches have

978-1-4799-3448-5/13/$31.00 © 2013 IEEE

been used for the feature extraction. The classification was performed using Support Vector Machine.

Tiji M Jose et al [7] illustrated the wavelet decomposition technique for the extraction of the features from the Tamil characters. The feed forward back propagation network classifier is used for the intention of classification. The recognition rate achieved in this paper was about 89% using the level 4 db2.

The scanned image is segmented into words using spatial space detection technique, at first paragraph is segmented to lines using vertical histogram, then the lines are segmented into words using horizontal histogram and finally the word to character image glyphs using horizontal histogram [9]. The extracted features for the character are the height, width of the character, number of horizontal lines present, Pixels in the various regions. After feature extraction, the output is given to the classifier. The accuracy achieved with this technique is 97%.

Indra Gandhi et al proposed a new approach of using Kohonen SOM (Self Organizing Map) for recognizing the online Tamil character [25]. The vectors of the binary image are created. When the segmentation of the character is over, then the images are scaled to unique height and weight. Some unwanted portions are included, but it can be removed by sobel edge detection. The median filter is used to increase the efficiency. The SOM is not applicable to the cursive characters which are used in this paper. The median filter is not suited for the offline Tamil characters. So the Zernike moments are used for feature extraction.

Jagadeesh Kannan et al [26] used Octal Graph method for the recognition of the Tamil Handwritten characters. Here, the character return on the octal graph's pixel is converted into the node of the graph. Each node has eight fields, that's why called as octal graph. Each node is connected to the other node based on the threshold value. The image is converted to the octal graph by the steps such as normalization, conversion, Identification of weighing factors and feature extraction. If the character is tedious and if it contains many curves, then octal graph method is not suitable.

III. INDIAN LANGUAGE CHARACTERISTICS

Tamil is one of the most popular South Indian languages. Tamil has 12 vowels and 18 consonants. By combining the vowels and consonants, the other letters are derived. *uyire<u>l</u>uttu* is vowels and *meyye<u>l</u>uttu* is consonants.

Fig.1. Tamil Characters-Vowels

Fig.2. Tamil Characters-Consonants

IV. PRE-PROCESSING

There are numerous tasks to be completed before performing character recognition. A handwritten document must be scanned and converted into a suitable format for processing. Preprocessing consists of a few types of sub processes to clean the document image and make it appropriate to carry the recognition process accurately. The sub processes which get involved in pre-processing are illustrated below:

- Binarization
- Noise reduction

A. Binarization

Binarization is a method of transforming a gray scale image into a black and white image through Thresholding [4][5]. Another approach, Otsu's method may be used to perform histogram based thresholding [6] [7] to get binarized image automatically. Otsu's method has been extended for multi level thresholding, called Multi Ostu method [8]. Normally, most researchers use thresholding concepts to extract the foreground image from background image [9][10][11]. In this method, the threshold value is fixed by taking any value between two foreground gray code images. Histogram based thresholding approach can also be used to convert a gray-scale image into a two tone image. In contrast, Adaptive Binarization method can also be used to identify the local gray value contrast of Image. This will help to extract text information from low quality documents. Another approach named Two-Level Global Binarization Technique represents the output using global thresholding technique [8].

B. Noise Removal

Digital images are prone to many types of noises. Noise in a document image is due to poorly photocopied pages. Median Filtering [9], Wiener Filtering method [12] and morphological operations can be performed to remove noise [6].

Median filters are used to replace the intensity of the character image [8], Where as Gaussian filters can be used to smoothing the image [13].

V. SYSTEM IMPLEMENTATION MODEL

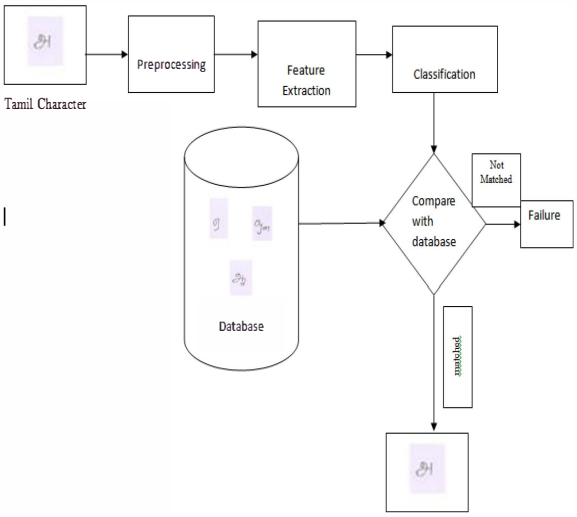

Fig.3. System Model

A. Feature Extraction

Individual image glyph is considered and extracted for features such as character height, width, horizontal lines, vertical lines, slope lines, circles, arcs etc.

The Selection of appropriate feature extraction method is probably the single most important factor in achieving high recognition performance. In [14] several methods of feature extraction for character recognition have been reported.

1) Zernike moments

Zernike moments are introduced by Zernike in 1934 and these moments are due to Zernike polynomials [20]. Zernike polynomials are a set of complex polynomials in which a complete orthogonal set is formed over the interior of the unit circle.

The orthogonal radial polynomial $R_{nm}(r)$ is defined as:

$$R_{nm}(r) = \sum_{s=0}^{(n-|m|)/2} (-1)^s g. \frac{(n-s)!}{s!\left(\frac{n+|m|}{2}-s\right)!\left(\frac{n-|m|}{2}-s\right)!} g^{r^{(n-2s)}} \quad (11)$$

The Zernike moment with repetition m and the order of *n* of a continuous image function *f (x, y)* is given by:

$$Z_{nm} = \frac{n+1}{\pi} \sum_{x} \sum_{y} f(x,y)[V_{nm}(x,y)]^* \quad (12)$$

Zernike moments have minimum information redundancy compared to Hu moments because it is orthogonal.

TABLE 1: Total number of moment's up to 10th order

Order (n)	Zernike moment of order n with repetation m (Anm)	Total number of moments up to order 10
0	A0,0	36
1	A1,1	
2	A2,0 A2,2	
3	A3,1 A3,3	
4	A4,0 A4,2 A4,4	
5	A5,1 A5,3 A5,5	
6	A6,0 A6,2 A6,4 A6,6	
7	A7,1 A7,3 A7,5 A7,7	
8	A8,0 A8,2 A8,4 A8,6 A8,8	
9	A9,1 A9,3 A9,5 A9,7 A9,9	
10	A10,0 A10,2 A10,4 A10,6 A10,8 A10,10	

B. Neural Classifier

A few models that have been applied for the HCR system include motor models [15], structure-based models [16] [17], stochastic models [18] and learning-based models [19]. Learning-based models have received wide attention for pattern recognition problems. Neural network models have been reported for the achievement of better performance than other existing models in many recognition tasks. Support vector machines have also been observed to achieve reasonable accuracy, especially in implementations of handwritten digit recognition and character recognition in Roman, Thai and Arabic scripts.

VI. EXPERIMENTAL RESULTS

The Proposed system contains the following modules: Preprocessing, Feature Extraction, Classification. Preprocessing is the conversion of color images to gray scale and then to binary images.

The features are extracted from those images using the moments and Back Propagation is used as the classifier.

A. *Preprocessing*

The original image is converted into binary image.

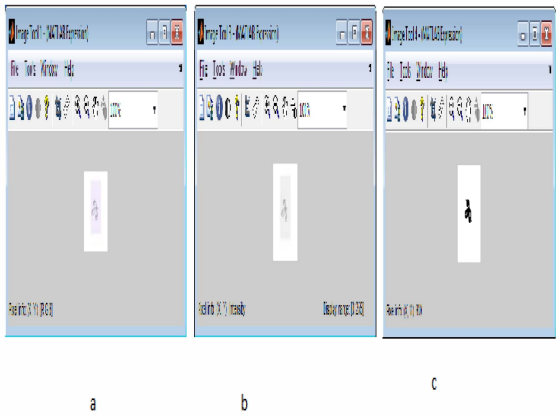

Fig.4. a) Original Image b) Gray Scale Image c) Binary Image

The image is cropped from its background. The Fig.5 shows that the image is cropped from the background.

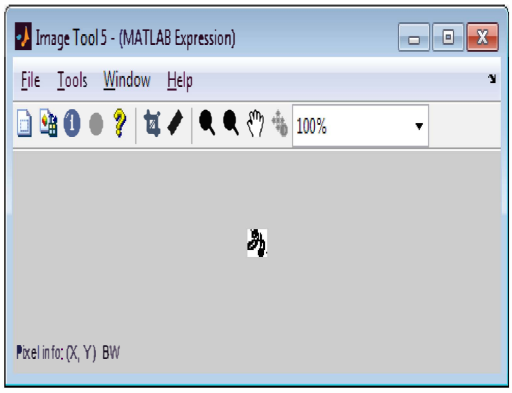

Fig.5.The Extracted Character from the background

The cropped image is placed at the center of the blank matrix is shown in the Fig.6.

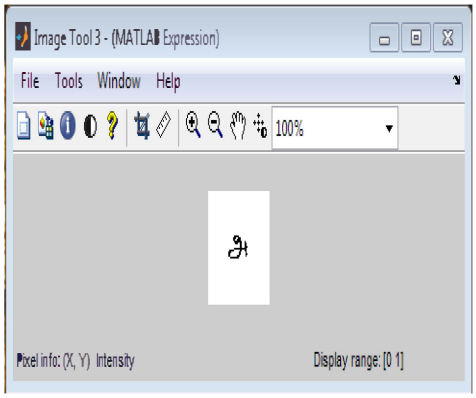

Fig.6. Placing the Cropped Image in the White Template

B. *Feature Extraction*

The feature is extracted using the Zernike moments.

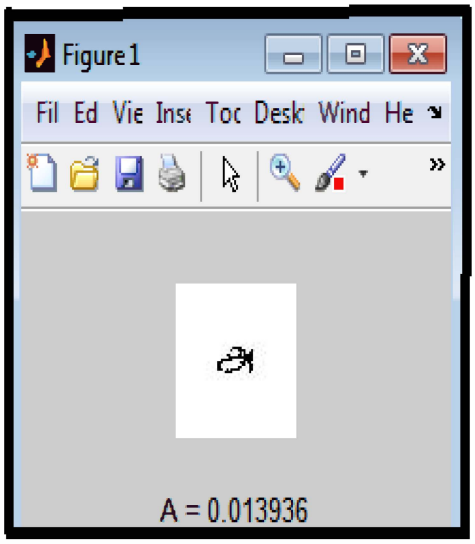

Fig.7. Feature Extraction performed for the character

C. *Classification and Recognition*

The output from the feature extraction is given as the input to the feed forward neural network classifier.

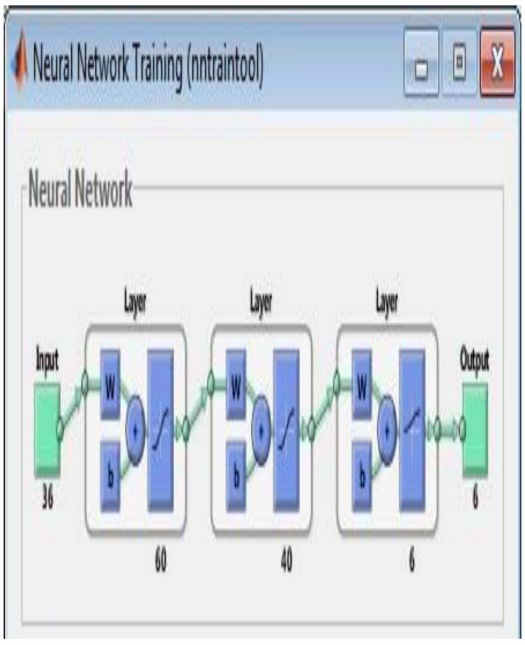

Fig.8.Feed Forward Neural Network Classifier

The classification of the Tamil characters has been done by the two-layer feed-forward networks. The training/learning is very fast.

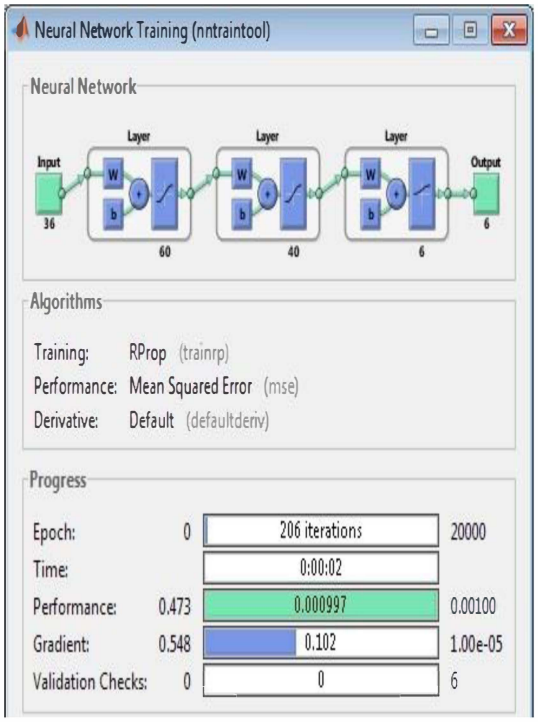

Fig.9. Training using Feed Forward Classifier

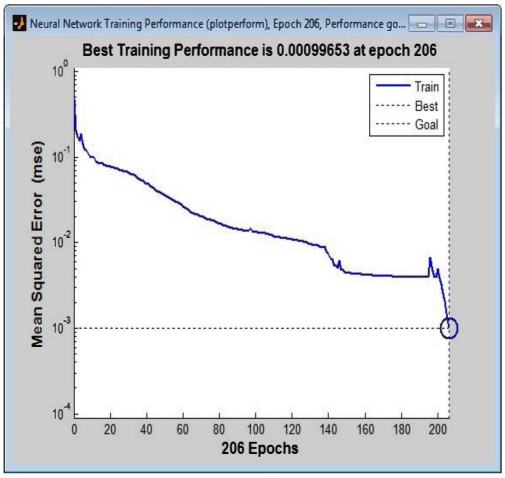

Fig.10. Performance Plot

VII. RESULTS AND DISCUSSION

25% of the data set is used as the testing data and the rest 75% is used as training data. 100 samples are collected from different persons.

The 36 Input are given to the neural network. The training is done in RProp algorithm, because it trains the given set within the short period of time. The iteration used for training is 20000. The time required for the training is about 02 seconds. The goal is put to 0.001. The accuracy is maintained based on the value that is set for the performance to meet. The two hidden layers are set for training. The first hidden layer consists of 60 neurons and the second hidden layer consists of 40 neurons.

The feature extraction performed using Hu moment is inefficient when compared to Zernike and polynomial moments. When the order of the moments is increased, the efficiency of the recognition rate is increased.

VIII. CONCLUSION AND FUTURE WORK

In this paper, the Zernike moments are used instead of the Hu moments. In the Hu moment, the image is defined over the real plane, but in Zernike the image is mapped to the unit circle. Zernike moments are invariant to translation, scale and rotation. To achieve scale invariant and translation invariant, the normalization method is essential. The translation invariant can be achieved by moving the image to the center.

Hu's invariant moments and Zernike moments which have been used in pattern recognition are used in this system to extract the features of handwritten Tamil characters. Neural classifiers (Feed Forward Neural network) have been used for the classification of Tamil characters.

The computations of the Zernike moments are quite difficult. The reason behind this is due to the normalization of the image. In future, different feature extraction methods like legendre polynomial and classifiers like support vector machines (SVM) can be used to improve the efficiency.

REFERENCES

[1] R. Plamondan, S.N. Srihari, "Online and offline handwriting recognition: A comprehensive survey", IEEE Transaction on PAMI, Vol22 (1) pp 63 – 84, 2000.

[2] U. Pal and B.B. Chaudhuri, "Indian script character recognition: A survey", Pattern Recognition, Elsevier, Vol. 37, pp. 1887-1899, 2004.

[3] Jomy John, Pramod K. V, Kannan Balakrishnan "Handwritten Character Recognition of South Indian Scripts: A Review", National Conference on Indian Language Computing, Kochi, Feb 19-20, 2011

[4] Shanthi N and Duraiswami K, "Performance Comparison of Different Image size for Recognizing unconstrained Handwritten Tamil character using SVM", Journal of Computer Science Vol-3 (9): Page(3) 760-764, 2007.

[5] S. Manke, U. Bodenhausen, .A connectionist recognizer for online cursive handwriting recognition., *Proceedings of ICASSP 94*, Vol. 2, 1994, pp 633-636.

[6] Shanthi N and Duraiswami K, "A Novel SVM -based Handwritten Tamil character recognition system", springer, Pattern Analysis & Applications,Vol-13, No. 2, 173-180, 2010.

[7] Ramanathan R, Ponmathavan S, Thaneshwaran L, Arun.S.Nair, and Valliappan N, "Tamil font Recognition Using Gabor and Support vector machines", International Conference on Advances in Computing, Control, & Telecommunication Technologies, page(s): 613 – 615, 2009.

[8] Sutha J and RamaRaj N, "Neural network based offline Tamil handwritten character recognition System" , International Conference on Conference on Computational Intelligence and Multimedia Vol : 2, page(s): 446 – 450, 2007.

[9] Jagadeesh Kumar R, Prabhakar R and Suresh R.M, "Off-line Cursive Handwritten Tamil Characters Recognition", International Conference on Security Technology, page(s): 159 – 164, 2008

[10] Suresh Kumar C and Ravichandran T, "Handwritten Tamil Character Recognition using RCS algorithms", Int. J. of Computer Applications, (0975 – 8887) Volume 8– No.8, October 2010.

[11] Bremananth R and Prakash A, "Tamil Numerals Identification", International Conference on Advances in Recent Technologies in Communication and Computing, page(s): 620 – 622, 2009.

[12] Stuti Asthana, Farha Haneef and Rakesh K Bhujade, "Handwritten Multiscript Numeral Recognition using Artificial Neural Networks", Int. J. of Soft Computing and Engineering ISSN: 2231-2307, Volume-1, Issue-1, March 2011.

[13] Paulpandian T and Ganapathy V, "Translation and scale Invariant Recognition of Handwritten Tamil characters using Hierarchical Neural Networks", Circuits and Systems, IEEE Int. Sym. , vol.4, 2439 – 2441, 1993.

[14] Anil.K.Jain and Torfinn Taxt, "Feature extraction methods for character recognition-A Survey," Pattern Recognition, vol. 29, no. 4, pp. 641 -662, 1996.

[15] L. R. B. Schomaker, H. L. Teulings, .A handwriting recognition system based on the properties and architectures of the human motor system., *Proceedings of the IWFHR, CENPARMI*, Concordia, Montreal, 1990, pp 195-211.

[16] K. H. Aparna, Vidhya Subramanian, M. Kasirajan, G. Vijay Prakash, V. S. Chakravarthy, Sriganesh adhvanath, .Online handwriting recognition for Tamil., *Proceedings of the Ninth International Workshop on Frontiers in Handwriting Recognition (IWFHR' 04)*, Tokyo, Japan, 2004, pp 438-443.

[17] K. F. Chan, D. Y. Yeung, .Elastic structural mapping for online handwritten alphanumeric character recognition., *Proceedings of 14th International Conference on Pattern Recognition*, Brisbane, Australia, August, pp 1508-1511, 1998.

[18] X. Li, R. Plamondon, M. Parizeau, .Model-based online handwritten digit recognition., *Proceedings of 14th International Conference On Pattern Recognition*, Brisbane, Australia, August, 1998, pp 1134-1136.

[19] Sigappi A.N, Palanivel S and Ramalingam V, "Handwritten Document Retrieval System for Tamil Language", Int. J of Computer Application, Vol-31, 2011.

[20] Sangeetha.S.K.B and Dr.Vijayachamundeeswari.V," Tamil OCR- A Survey", International Conference on Computing and Control Engineering (ICCCE 2012), 12 & 13 April, 2012.

[21] Bharath.A and Sriganesh Madhvanath, "*HMM-Based Lexicon Driven and Lexicon-Free Word Recognition for Online Handwritten Indic Scripts*", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol34, No.4, April 2012.