Neural Network Based Offline Tamil Handwritten Character Recognition System

J. SUTHA, M.E
Asst. Professor, CSE Dept.,
Sethu Institute of Technology, Kariapatti.
Email id : sutha_skad@yahoo.co.in

N. RAMARAJ, M.E.,Ph.d.,
Principal,
G.K.M. College of Engg. And Techmology.
Chennai.

Abstract

In this paper we propose an approach to recognize handwritten Tamil characters using a multilayer perceptron with one hidden layer. The feature extracted from the handwritten character is Fourier Descriptors. Also an analysis was carried out to determine the number of hidden layer nodes to achieve high performance of backpropagation network in the recognition of handwritten Tamil characters. The system was trained using several different forms of handwriting provided by both male and female participants of different age groups. Test results indicate that Fourier Descriptors combined with backpropagation network provide good recognition accuracy of 97% for handwritten Tamil characters.

Keywords : *Indian Language, Feature Extraction, Handwritten character Recognition, Backpropagation network, Fourier Descriptors.*

1. Introduction

Handwritten character recognition is a difficult problem due to the great variations of writing styles, different size and orientation angle of the characters. Among different branches of handwritten character recognition it is easier to recognize English alphabets and numerals than Tamil characters. Many researchers have also applied the excellent generalisation capabilities offered by ANNs to the recognition of characters [3],[4],[6]. Many studies have used fourier descriptors and Back Propagation Networks for classification tasks. Fourier descriptors were used in [3] to recognise handwritten numerals. In [6], fourier descriptors and a Back Propagation Network were used to classify tools.

There have been only a few attempts in the past to address the recognition of printed or handwritten Tamil Characters. However, less attention had been given to Indian language recognition. Some efforts have been reported in the literature for Tamil[1],[5],[7],[9] scripts. Although most of these handwriting recognition applications concentrate on on-line handwriting[1],[7], there have been attempts on offline Tamil handwritten characters [5],[9]. In this study, we exploit the use of neural networks for off-line Tamil handwriting recognition. Neural networks have been widely used in the field of handwriting recognition [3] [6].

The present work describes a system for offline recognition of Tamil script, a language widely spoken in South India. It is also one of the official languages in countries such as Singapore, Malaysia, and Sri Lanka apart from India. Recently, the Indian Government recognized it as a classical language.

In this paper, we propose a recognition system for handwritten Tamil characters. The organization of the paper is as follows. Section 2 gives an introduction into the Tamil language. Preprocessing steps are described in section 3, and the feature extraction is described in Section 4, the recognition and structure analysis are described in Section 5, Results of the experimentation are presented in section 6. Finally in section 7, we present our conclusions and future work.

0-7695-3050-8/07 $25.00 © 2007 IEEE
DOI 10.1109/ICCIMA.2007.86

2. The Tamil Language

Tamil which is a south Indian language, is one of the oldest language in the world. The Tamil alphabet consists of 12 vowels, 18 consonants and one special character (AK). Vowels and consonants are combined to form composite letters, making a total of 247 different characters. A subset of the Tamil alphabet, which consists of the most commonly used 30 letters, is used for the study.

3. Preprocessing

The handwritten character data samples were acquired from various students and faculty members both male and female of different age groups. Their handwriting was sampled on A4 size paper. They were scanned using flat-bed scanner at a resolution of 100dpi and stored as 8 bit grey scale images. Some of the common operations performed prior to recognition are: Smoothing, thresholding and skeletonization.

3.1 Image Smoothing

The task of smoothing is to remove unnecessary noise present in the image. Spatial filters could be used. To reduce the effect of noise, the image is smoothed using a Gaussian filter.

3.2. Thresholding

The task of thresholding is to extract the foreground from the background. A number of thresholding techniques have been previously proposed using global and local techniques. Global methods apply one threshold to the entire image while local thresholding methods apply different threshold values to different regions of the image. The histogram of gray scale values of a handwritten character image typically consists of two peaks: a high peak corresponding to the white background and a smaller peak corresponding to the foreground. So the task of determining the threshold gray-scale value is one of determining as optimal value in the valley between the two peaks. Here Otsu's method of histogram based global thresholding algorithm is used[8].

3.3. Skeletonization

Skeletonization is the process of peeling off a pattern as any pixels as possible without affecting the general shape of the pattern. In other words, after pixels have been peeled off, the pattern should still be recognized. The skeleton hence obtained must be as thin as possible, connected and centered. When these are satisfied the algorithm must stop. A number of thinning algorithms have been proposed and are being used. Here Hilditch's algorithm is used for skeletonization[8].

4. Feature Extraction

Feature extraction plays important role in the success of handwritten character recognition system. In this proposed system, the features are extracted from closed boundary trace. There are many features that can be used to describe closed boundary trace, the Fourier coefficients was chosen so that they are invariant with respect to translation, rotation and size of similar characters.

4.1. Boundary tracing

The purpose of the algorithm is to extract the information of the boundary of a handwritten character. Various boundary tracing methods are available. The “eight-neighbor” adjacent method is adopted in this system. The algorithm scans the binary image until it finds the boundary. The searching will follow according to the clockwise direction. For any foreground pixel p, the set of all foreground pixels connected to it is called connected component containing p. The pixel p and its 8-neighbors are shown in Figure 1.

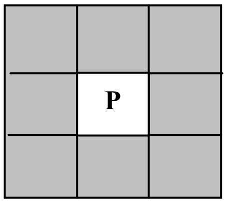

Figure 1. Pixel P and its 8-neighbors

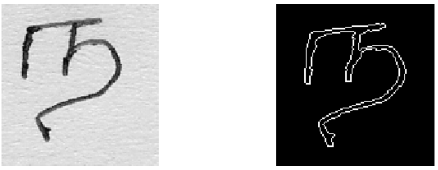

Figure 2. Boundary tracing of a character

Once a white pixel is detected, it will check another new white pixel and so on. The tracing will follow the boundary automatically. When the first pixel is found, the program will be assigned the coordinates of that position to indicate that this is an origin of the boundary. The tracer will search for the next nearby white pixels. The new found pixel will be assigned as a new reference point and starts the eight-neighbor searching. In this way, the coordinates of the initial point are varied according to the position. As the tracer moves along the boundary of the image, the corresponding coordinates will be stored in an array for the computation of Fourier Descriptors. During the boundary tracing process, the program will always check the condition whether the first coordinates of the boundary are equal to the last coordinates. Once it is obtained means that the whole boundary has been traced and completed the boundary tracing process. Figure 2 shows the boundary tracing result.

4.2. Fourier descriptors

Once a boundary image is obtained then Fourier descriptors are found. This involves finding the discrete Fourier coefficients a[k] and b[k] for $0 \leq k \leq L-1$, where L is the total number of boundary points found, by applying equations (1) and (2).

$$a[k] = 1/L \sum_{m=1}^{L} x[m] e^{-jk(2\pi/L)m} \tag{1}$$

$$b[k] = 1/L \sum_{m=1}^{L} y[m] e^{-jk(2\pi/L)m} \tag{2}$$

where x[m] and y[m] are the x and y co-ordinates respectively of the m^{th} boundary point. The Fourier coefficients derived according to equations (1) and (2) are not rotation or shift invariant. In order to derive a set of Fourier descriptors that have the invariant property with respect to rotation and shift, the following operations are defined. For each n compute a set of invariant descriptors r(n).

$$r(n) = \left[|a(n)|^2 + |b(n)|^2 \right]^{1/2} \tag{3}$$

It is easy to show that r(n) is invariant to rotation or shift. A further refinement in the derivation of the descriptors is realized if dependence of r(n) on the size of the character is eliminated by computing a new set of descriptors s(n) as

$$s(n)=r(n)/\ \ r(1) \qquad (4)$$

The Fourier coefficients a(n), b(n) and the invariant descriptors s(n), n = 1,2 .... (L-1) were derived for all of the character specimens. Then the sixteen invariant descriptors of s(n) are input into the neural network.

5. Recognition

Recognition of handwritten letters is a very complex problem. The letters could be written in different size, orientation, thickness, format and dimension. These will give infinity variations. The capability of neural network to generalise and be insensitive to the missing data would be very beneficial in recognising handwritten letters. The proposed Tamil handwritten character recognition system uses a neural network based approach to recognize the characters represented by scale and shift invariant features. Feed forward Multi Layered Perceptron (MLP) network with one hidden layer trained using back-propagation algorithm has been used to recognise handwritten Tamil characters.

5.1 Structure Analysis of Back Propagation Network

The recognition performance of the Back Propagation network will highly depend on the structure of the network and training algorithm. In the proposed system, Back Propagation algorithm has been selected to train the network. It has been shown that the algorithm has much better learning rate. The number of nodes in input, hidden and output layers will determine the network structure. The best network structure is normally problem dependent, hence structure analysis has to be carried out to identify the optimum structure. In the proposed system, the number of input and output nodes were fixed at 16 and 5 respectively, since the feature extracted from the handwritten character images were the sixteen invariant fourier descriptors and the target outputs are subset of 30 Tamil characters denoted in binary form. Therefore, only the number of hidden nodes need to be determined. No. of epochs taken to train the network and recognition efficiency will be used to judge.

5.2 Number of Hidden Layer Nodes

The number of hidden nodes will heavily influence the network performance. Insufficient hidden nodes will cause underfitting where the network cannot recognise the numeral because there are not enough adjustable parameter to model or to map the input-output relationship. However, excessive hidden nodes will cause overfitting where the network fails to generalise. Thus, the network could not map the independent or testing data properly. One way to determine the suitable number of hidden node is by finding the minimum number of epochs taken to recognize a character and recognition efficiency of training as well as test character set as the number of hidden nodes is varied. The performance of the network is shown in Table 1. Number of Epochs Variation with different number of Hidden nodes is plotted in the Figure 3.

Table 1. Number of Epochs Variation with different number of Hidden node

No. of Hidden Nodes	Learning Rate	Momentum Factor	No. of Epochs	Recognition %	
				Training Set	Test Set
3	0.2	0.8	15	100	96.2
5	0.2	0.8	33	100	97
7	0.2	0.8	56	100	97
10	0.2	0.8	79	100	96
12	0.2	0.8	345	100	96.5
15	0.2	0.8	501	100	97

6. Experimental results

The invariant Fourier descriptors feature is independent of position, size, and orientation. With the combination of Fourier descriptors and back propagation network, a high accuracy recognition system is realized. The entire database of 40 users was divided into two categories. The training set consists of the writing samples of 25 users selected at random from the 40, and the test set, of the remaining 15 users. A portion of the training data was also used to test the system. In the training set, a recognition rate of 100% was achieved and in the test set the recognized speed for each character is 0.1sec and accuracy is 97%. Understandably, the training set produced much higher recognition rate than the test set. Structure analysis suggested that backpropagation network with 5 hidden nodes has lower number of epochs as well as higher recognition rate.

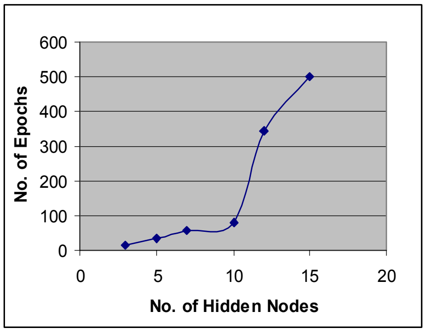

Figure. 3 No. of Epochs Variation with different number of Hidden nodes

7. Conclusions

In this paper we have presented a system for recognizing handwritten Tamil characters. Experimental results shows that Fourier descriptors with back propagation network yields good recognition accuracy of 97%. The results of structure analysis shows that if the number of hidden nodes increases the number of epochs taken to recognize the handwritten character is also increases. The methods described here for Tamil handwritten character recognition can be extended for other Indian scripts by including few other preprocessing activities like line segmentation and character segmentation.

8. References

[1] K. H. Aparna, Vidhya Subramanian, M. Kasirajan, G. Vijay Prakash, V.S. Chakravarthy and Sriganesh Madhvanath, "Online Handwriting Recognition for Tamil", *Proceedings of the 9th IEEE International Workshop on Frontiers in Handwriting Recognition,* IWFHR-9 2004.

[2] R. M. Bozinovic and S. N. Srihari, "Off-line cursive script word recognition", *IEEE Trans. on Pattern Analysis and Machine Intelligence.*, vol. 11, no. 1, pp. 68-83, Jan. 1989.

[3] Y.Y. Chung, and M.T. Wong, "Handwritten Character Recognition by Fourier Descriptors and Neural Network", *Proceedings of IEEE TENCON – Speech and Image Technologies for Computing and Telecommunications*, 1997.

[4] M. Hanmandlu, K.R.M. Mohan and H. Kumar, "Neural-based handwritten character recognition", *Proceedings of 5th IEEE International Conference on Document Analysis and Recogniton,* ICDAR'99.

[5] S. Hewavitharana and H.C. Fernando, "A Two Stage Classification Approach to Tamil Handwriting Recognition", *Proc. The Tamil Internet 2002 Conference*, California, USA, pp.118-124, 2002.

[6] Hongbong Kim and Kwanghee Nam "Object Recognition of One-DOF Tools by a Back-Propagation Neural Net", *IEEE Transactions on Neural Networks,* vol. 6, no. 2, pp. 484-487, March 1995.

[7] Niranjan Joshi, G. Sita, A.G. Ramakrishnan, and Sriganesh Madhvanath, "Comparison of Elastic Matching Algorithms for Online Tamil Handwritten Character Recognition", *Proceedings of the 9th IEEE International Workshop on Frontiers in Handwriting Recognition,* IWFHR-9 2004.

[8] N. Shanthi and K. Duraiswamy, "Preprocessing Algorithms for the Recognition of Tamil Handwritten Characters", *3rd International CALIBER – 2005*, Cochin, pp.77-82, Feb. 2005.

[9] R.M. Suresh, S. Arumugam and L. Ganesan, "Fuzzy Approach to Recognize Handwritten Tamil characters", *Published by IEEE, ICCIMA'99,* New Delhi, pp.459-464, 1999.