Asian Journal of Information Technology 15 (20): 4124-4128, 2016
ISSN: 1682-3915

Handwritten Tamil Character Recognition Using Geometric Feature Extraction Approach

[1]S. Kowsalya and [2]P.S. Periyasamy
[1]Department of CSE, United Institute of Technology, Coimbatore, Tamail Nadu, India
[2]Department of ECE, KSRCE, Tiruchengode, Tamil Nadu, India

Abstract: Character recognition is one of the most fascinating and challenging researches currently in the area image processing. It has been receiving considerable attention due to its versatile range of real-time application which includes reading aid for the blind, postal automation, processing of cheque and digitization of historical documents. Now a days different methodologies for different language are in widespread use for character recognition. Character recognition from a scanned document page involves difficult task due to the free-flow nature of handwritten. In this study a geometric feature extraction approach is implemented with efficient learning mechanism for training and testing using neural network for Tamil handwritten script. After selective preprocessing steps for constrained inputs, the document is split into paragraph and then segmented to line, word and individual character for further recognition. The geometric features for each character are trained in an Effective Learning Machine (ELM) with almost information. With this information each testing character is analyzed for recognition. This procedure results more than 90% accuracy for individual characters.

Key words: Character recognition, preprocessing, geometric feature extraction, ELM, extraction

INTRODUCTION

In today's fast growing technology, digital recognitions are playing wide role and providing more scope to perform research in OCR techniques. Many of today's document scanners for the PC come with the software that performs a task of character recognition. OCR software allows you to scan a printed document and then convert the electronic text in word format. Though a very small part of digital image processing we find that the implementations of OCR in the corporate world it seems to be huge. It can be used in banking and other financial institutions, libraries, convert existing books into computer format so the books can be taken on a CD-ROM or directly uploaded on the internet.

Tamil character recognition is one of the challenging tasks in optical character recognition because of its various characters and combinations of characters which are not exist in any other languages. Tamil script has 30 basic shapes namely 12 vowels and 18 consonants also 1 special character named as Aayutha ezhuthu which is classified in Tamil grammar as being neither a consonant nor a vowel and 216 composite letters. There are many works in character recognition by Bharath and Madhvanath (2007), Seethalakshmi *et al.* (2005), Aparna *et al.* (2002) and Chinnuswamy and Krishnamoorthy (1980). In this study, yet another methodology for character recognition is carried with geometric features which trace out the different structural shapes of Tamil characters and the classifier is trained with quite a number of possible combinations to attain the best network. Optical Character Recognition, OCR can be for printed and handwritten documents and further handwritten documents can be of online or offline. In offline handwritten texts, many automated reading handwritten characters are in practical applications like reading aid for people who are visually handicapped which is combined with a speech synthesizer, mail-order forms, banks, credit cards, library ledger, automating reading for sorting of postal mail, signature recognition, multimedia, etc (Pal *et al*., 2003).

Preprocessing: Preprocessing is defined as cleaning and enhancing the document image and making it appropriate for input to the OCR system. It avoids the unwanted noise, background information and inputs the needed character as input to the classifier. The major steps considered under preprocessing are noise removal , binarization , segmentation and bounding box . The noise is mainly due to optical scanning devices in the input, leads to poor system performance. These imperfection must be removed prior for recognition. Noise can also be included in the image during image acquisition. There are

Corresponding Author: S. Kowsalya, Department of CSE, United Institute of Technology, Coimbatore Tamal Nadu, India

Fig. 1: Noise removal

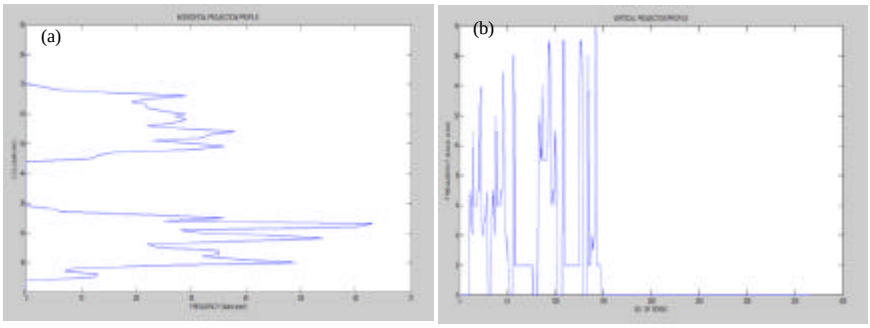

Fig. 2: Projection profile

several types of noise such as Gaussian noise, Rayleigh noise, salt and pepper noise. The noise present in the input image can be removed by using filters. The input image of handwritten character in from the local data base is converted to gray scale format. Binarization is the important image processing step in which pixel value is separated into two groups white as foreground and black as background. The goal of binarization is to remove unwanted information and thus protecting the useful information from the image. Script segmentation is an important task for character segmentation system. Segmentation is the process of splitting the document image into text lines and then splitting lines to word and then to individual character.

Noise removal: Noise removal is the process of removing the unwanted black pixel present in the input image. The scanned document often contains noise that may arise due to printer, scanner, print quality, documents age (Roy *et al*., 2004) etc. Noise present in the input image is removed using filter and output passed to next stage of the preprocessing. Noise removal for a single character is shown in Fig. 1.

Binarisation: The binarization algorithm converts grey scale to binary by determining proper threshold value to separate the handwritten from the background. Otsu (1979)'s method of binarization is the most optimized method for binarization technique, in which it determines its threshold value automatically depending upon the input data and the output data will be a matrix of zeros and ones.

Segmentation: Se gmentation is the process of splitting the whole handwritten document into individual characters. Broken characters are not considered (Pal, 1997). The optimized segmentation method is projection profile method. In which the horizontal projection profile used for paragraph and line segmentation and vertical projection profile used for word and character segmentation works as follows: Lines are segmented from a document by finding the valleys of the horizontal projection profile computed by counting the number of black pixels in each row. The trough between two consecutive peaks in this profile denotes the boundary between two text lines. In vertical projection profile the column wise scanning results in multiple separate words. The same methodology is used for character segmentation from the word segmentation. From the each segmented letter the efficient feature extraction provides better recognition accuracy. Vertical projection graph is shown in Fig. 2.

Bounding box: Bounding box is finding the smallest matrix size which fits the entire skeleton of the character. This process should be done because each zone of the image is having different line segment, so each character matrix should be independent of image size. The bounded output of the character image is given to the feature extraction stage and is shown in Fig. 3.

Feature extraction: Feature extraction is the process of defining the rules and unique feature of an object which is the character's individual identity. Here, geometry of

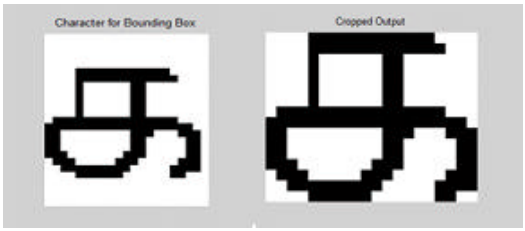

Fig. 3: Bounding box

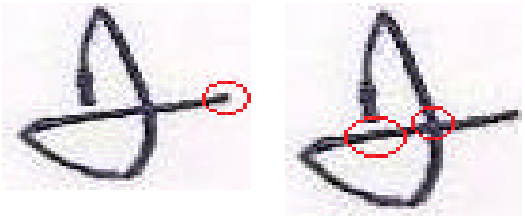

Fig. 4: a) Starters and b) Intersections

the character is taken into consideration to define the unique features of the character. The various partitions attained to extract the features are:

Zoning: Zoning is defined as the splitting of character image into several windows of equal size. The image of character is zoned into 9 equal sized windows. The occurrence of line segment in each windows of the character skeleton is determined because each character has particular line segment occurring in particular zone.

Starters, intersection and minor starters: To define the line segment in each zone the traverse of character skeleton should be defined. For this purpose certain pixels in the character skeleton is defined as starters having individual neighbouring, intersections having more than one neighbour and minor starters having multiple neighbouring pixels (Rajan *et al*., 2009) are as shown in Fig. 4a, b.

Character traversal: After zoning the character skeleton into equal sized windows, each zone of the character is subjected to character traversal algorithm. The algorithm starts by considering the starters first and then with minor starters. The entire line segment obtained during this process is stored with position of pixels in each line segment. Once all the pixels in the image of character are visited the algorithm stops.

MATERIALS AND METHODS

After extracting the line segment from the image, the category of line segment is defined from any one of

4	5	6
3	c	7
2	1	8

Fig. 5: The 3×3 matrixes

horizontal or vertical or right or left diagonal line. For determining the type of line segment, the convention is required to define the position of neighboring pixels with respect to the center pixel c of 3x3 matrixes under consideration (Fig. 5). If maximum occurring direction is 2 or 6 then the line type is right diagonal likewise, if its 4 or 8 then left diagonal or if 1 or 5 then the line type is vertical and finally if it is 3 or 7 then the line type is horizontal. After the line segment of each character is defined, the feature in the each zone is determined. The features that can be extracted in each zone are the number of horizontal, vertical, right diagonal, left diagonal lines and normalized length of all the four lines and finally the normalized Area of the Skeleton. The mber of any particular line type is normalized using the following method:

$$\text{Value} = 1 - ((\text{number of lines}/10) \times 2)$$

Normalized length of any particular line type is found using the following method:

$$\text{Length} = (\text{Total pixels in that line type})/(\text{Total zone pixels})$$

The above features are extracted individually for each zone. In this study the image of character is split into 9 equal sized windows. If the spitted zones are N then the number of features extracted will be 9*N for each character.

Classifier: The most successful application proposed by neural networks is the optical character recognition. NNs can simply cluster the feature vectors in the feature space or they can integrate feature extraction and classification stages by classifying characters directly from images. The number of nodes in the input layer varies according to the dimensionality of the feature vector or the segment image size. The number of nodes in the hidden layer governs the variance of samples that can be correctly recognised by this NN (Rajan *et al*., 2009). As the threshold functions

are nondifferentiable, the gradient descent learning algorithms for multilayer feed forward neural networks cannot be directly applied. Hence, a number of modifications to gradient descent methods have been proposed in the literature. The extreme learning machine ELM, algorithm is a new learning algorithm for the Single hidden Layer Feed forward Neural networks (SLFNs).

Algorithmic model: Corwin *et al*. (1994) has proposed an iterative method for training multilayer networks with threshold functions. The sigmoid function with a gain parameter is used in the training instead of the threshold function directly:

$$g(x)=1/(1+e^{-\lambda x})$$

If the training error is small the gain parameter λ is gradually increased during the training until the slope of the sigmoid is sufficiently large to allow a transfer to a threshold network with the same architecture. However, in many cases, the error may not be small enough to allow the λ to be increased.

Algorithmic flow: Given a training set X, Threshold activation function h(x) and number of hidden neurons L:

$$X = \{(xi,ti) \mid xi \in R^n, ti \in R^m\} \text{ and } h(x)=1_{x>=0} + 0_{x<0}$$

- Assign random input weight w_i and bias of hidden neurons b_i, i=1,...L
- Calculate the hidden layer output matric H
- Calculate the output weight $\beta = H'T$

Algorithmic to predict: The extreme learning machine based on rough sets is mainly used on the sample data sets classification prediction. For testing set Y, put into the trained network to predict and obtain a final classification result. Then compare the proposed algorithm and the traditional ELM with classification accuracy. There are two main standards to measure the pros and cons of the algorithm: Training time and prediction time; Testing set classification accuracy (Huang *et al*., 2006). On the comprehensive when training network and using network testing, the shorter time consumed by the better for the measure of the sample set classification accuracy using the following equation:

$$\text{Testing accuracy} = 1 - \frac{\text{Number of misclassification}}{\text{Total number of samples}}$$

The higher testing accuracy the better and the more effective the algorithm is.

Table 1: Results of testing tamil characters

Type of inputs	Number of inputs	Testing accuracy %
Individual characters	250	97
Words	100	94
Document	50	69
Overall resluts	-	87

RESULTS AND DISCUSSION

The handwritten data used for the experiment is taken from various people of different ages from individual characters to words and lines. The input data are scanned using a high DPI scanner. The database is divided into different subsets and testing is done on each individual subset while others are taken for training. The correct identification results for all the subsets are considered to get an average accuracy. The performance difference is tabulated and the difference is analyzed due to the various study qualities, scanner quality and poor writing style and also due to touching characters. The main advantages behind implementing neural networks in OCR is its faster development times comparing to other model and its ability to automatically retrain for the peculiarities of different writing and printing styles and finally it is capable to run on parallel processors. Also the drawback of choosing neural network is that, introducing a new shape to the network requires that the whole network be retrained, to form a different architecture show in Table 1.

CONCLUSION

Script identification is an important step for multi-lingual OCR development. The document was scanned and input was taken into system. Then the various Preprocessing techniques were done. Feature extraction is the important steps in character recognition system as each character has different feature that distinguish from each other characters. In this paper, an effective approach for Tamil character recognition is proposed. The different kinds of data sets for Tamil characters are considered. This work can be further improved with other different classifiers such as SVM, SOM, tree classifier, Fuzzy classifiers and comparing the accuracy.

REFERENCES

Aparna, K.H., S. Jaganathan, P. Krishnan and V.S. Chakravarthy, 2002. An optical character recognition system for Tamil Newsprint. Proceedings of the International Conference on Universal Knowledge and Language-2002, November 25-29, 2002, ICUKL, Goa, India, pp: 881-886.

Bharath, A. and S. Madhvanath, 2007. Hidden Markov Models for online handwritten Tamil word recognition. Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), September 23-26, 2007, IEEE, Bangalore, India, ISBN:0-7695-2822-8, pp: 506-510.

Chinnuswamy, P. and S.G. Krishnamoorthy, 1980. Recognition of handprinted Tamil characters. Pattern Recognit., 12: 141-152.

Corwin, E.M., A.M. Logar and W.J. Oldham, 1994. An iterative method for training multilayer networks with threshold functions. IEEE. Trans. Neural Netw., 5: 507-508.

Huang, G.B., Q.Y. Zhu and C.K. Siew, 2006. Extreme learning machine: Theory and applications. Neurocomputing, 70: 489-501.

Otsu, N., 1979. A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybernet., 9: 62-69.

Pal, U., 1997. On the optical character recognition of printed Bangla script. Ph.D Thesis, Indian Statistical Institute, Kolkata, India.

Pal, U., S. Sinha and B.B. Chaudhuri, 2003. Multi-script line identification from Indian document. Proceedings of the 7th International Conference on Document Analysis and Recognition (ICDAR'03), August 3-6, 2003, IEEE, Kolkata, India, pp: 880-884.

Rajan, K., V. Ramalingam, M. Ganesan, S. Palanivel and B. Palaniappan, 2009. Automatic classification of Tamil documents using vector space model and artificial neural network. Expert Syst. Appl., 36: 10914-10918.

Roy, K., U. Pal and B.B. Chaudhuri, 2004. A system for joining and recognition of Broken Bangla numerals for Indian postal automation. Proceedings of the 4th Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP'04), December 16-18, 2004, Indian Statistical Institute, Bangalore, India, ISBN 978-1-4503-3061-9, pp: 641-646.

Seethalakshmi, R., T.R. Sreeranjani, T. Balachandar, A. Singh and M. Singh *et al.*, 2005. Optical character recognition for printed Tamil text using Unicode. J. Zhejiang Univ. Sci., 6: 1297-1305.