www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume – 5 Issue -02 February, 2016 Page No. 15741-15743

Malayalam Character Recognition using Discrete Cosine Transform

Saidas S. R.[1], Rohithram T[1], Sanoj K. P[1], Manju Manuel [2]

[1]Assistant Professor, Dept of EIE
College of Engineering, Vadakara
mail2saidas@gmail.com

[2]Associate Professor, Dept of ECE
Rajiv Gandhi Institute of Technology, Kottayam,
manju.manuel@gmail.com

Abstract: ***This paper describes a feature extraction method for optical character recognition system for handwritten documents in Malayalam, a South Indian language. The scanned image is first passed through various preprocessing stages of operations like noise removal, binarization, thinning and cropping. After preprocessing projection profiles of each character is found. 1- D Discrete Cosine Transform (DCT) of projection profiles used as a feature. A multilayer artificial neural network (ANN) with logsig activation function is used for classification. The promising feature of the work is that successful classification of 44 handwritten characters.***

Keywords: Malayalam, Optical Character Recognition (OCR), Handwritten, Segmentation, Discrete Cosine transform, Projection profile, ANN (Artificial Neural Network)**.**

1. INTRODUCTION

Malayalam is one of the 22 officially recognized languages of India. It ranks eighth in its number of speakers. Malayalam consists of a total of 53 characters comprising of 37 consonants and 16 long and short vowels. Malayalam is a Dravidian language with about 35 million speakers. It is spoken mainly in the south western India, particularly in Kerala. Until the 16th century, Malayalam was written in the vattezhuthu script. Modern Malayalam script is derived from the Grantha script, a descendant of the ancient Brahami script. Malayalam is written from left to right [1]. As a result of the difficulties of printing Malayalam, a simplified or reformed version of the script was introduced. New style reduced the number of characters radically. The main change involved writing consonants and diacritics separately rather than as complex characters. These changes are not applied consistently so the modern script is often a mixture of traditional and simplified letters [2]. Some of the problems

Very few works have been reported so far Malayalam optical character Recognition. Malayalam language is rich in patterns while the combinations of such patterns makes the problem even more complex. NAYANA^{TM} OCR developed by CDAC one of the earliest one which capable of recognizing printed character [3]. Another work for recognition of unconstrained isolated handwritten character recognition using ANN which classify 33 characters in Malayalam [4] [5]. But the main drawback of the work is that it could classify only 33 classes. In this paper we present a method for classify 44 characters of Malayalam in to 44 classes. It make use of 1-D discrete cosine transform of projection profile of the character as feature and Neural network to classify 44 character.

The rest of the paper is organized as follows. In Section 2, Typical Optical Character recognition (OCR) system and various blocks are discussed. In Section 3, presented the use of preprocessing, feature extraction and classification. In Section 4 performance of the system for 44 handwritten characters are illustrated. Section 5 deals with applications of a typical OCR system. Finally Section 6 summarizes this paper with some concluding remarks.

2. THE OCR SYSTEM

Optical character recognition (OCR) is the conversion of scanned or photoed images of typewritten or printed text into machine-encoded/computer-readable text. A typical OCR consist of following stages as shown in Fig. 1

1. Image scanning
2. Preprocessing
3. Character extraction
4. Recognition

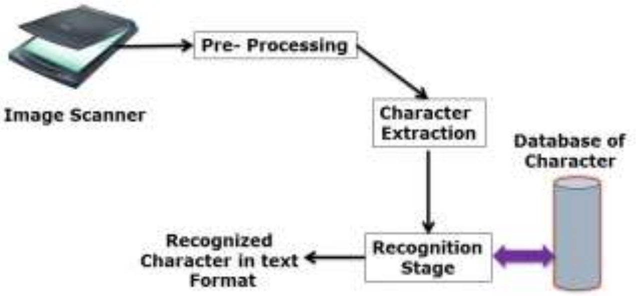

Fig. 1. Block diagram of typical character recognition system

2.1 Image scanner

Image scanner is a device used to transfer images or text into a computer. Image scanning is done to get the digital form of character in a page or paper. Resolution at which the scanning is done plays crucial role in accuracy of character recognition. If scanning is done at lower resolution performance of the system degrade at the same time storage size of data base is also low. If the scanning is done at higher resolution, performance of the system increases but overall storage size for database increases. So, scanning should be at optimal resolution. In this paper scanning is done at 300 dpi resolution. There are a number of other factors that affect the accuracy such as scanner quality, type of printed documents (laser printer or photocopied), paper quality, fonts used in the text, etc. Accuracy of an OCR system for black text on a white background higher than a grey-level document image with poor illumination and a document image with complex background. Fig. 2 shows sample of scanned image.

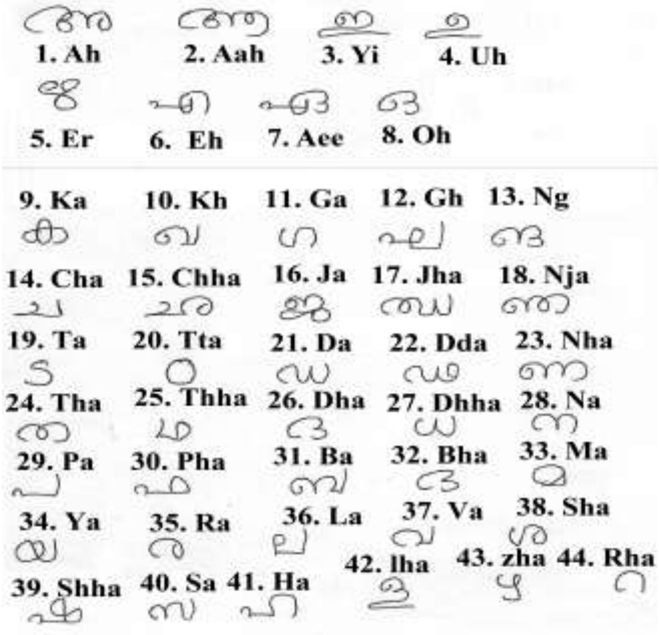

Fig. 2. Scanned Image of 44 character

2.2 Preprocessing

The scanned image is under go several preprocessing stages to suppress noise and to enhance the image. The preprocessing have ability to remedy some of the problems or noise that arises during digitalisation of document. The scanned image is usually passed through the following preprocessing stages

1) Binarization
2) Noise removal
3) Thinning

Usually the output of preprocessing is a binary image

2.3 Character Extraction

Usually scanned document consist of large number of words and sentence. For recognition we have to extract each character. Character extraction involve 3 stages. Firstly detected each line of text from the scanned document. From each line Words are detected and from each word Character is segmented. Each character in a scanned image is extracted and further used for feature extraction.

2.4 Data Base Creation

Prior to the usage of the character recognition system, a set of sample images have to be stored in database as template. OCR system involves two phases of operation namely training and testing. During training phase each character or some unique features extracted are stored in the data base with some ’label’ of a character. When testing is done test sample is compared with the data base and character is identified.

2.5 Recognition Stage

In the recognition stage, the extracted character is identified. Extracted character are compared with set of character in the data base. In the proposed method, ANN (Artificial neural network) is used for classification and is detailed in next section.

3. PROPOSED METHOD

The first step is creation of a data base for Malayalam characters. Since no standard data base is available a set of samples are collected from persons having different age and sex group. Data base is created such a way that 44 characters to be recognized are written over a white A4 size paper. 44 characters consist of 8 vowels and 36 consonants. Vowels are numbered from 1 to 8 and consonant are numbered from 9 to 44 in Fig. 2. These samples are scanned using Canon MF4820D. Scanning is done at a resolution of 300 - 400 dpi. Samples are made with a white background. Fig. 3 shows scanned image. Then each character is segmented and stored in different folder with some unique id’s to represent each character.

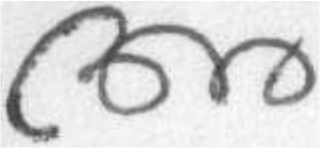

Fig. 3. Scanned image of malayalam letter ’ah’

3.1 Pre-processing

Scanned image usually suffers from noise, usually Gaussian noise and salt and paper noise. Pre-processing is done to suppress the effect of noise and to enhance quality of scanned images

The scanned image of Malayalam character ‘Ah’ is shown in Fig. 3 which is intensity (Gray Scale) image and is further passed through following preprocessing stages. The noise present in the scanned images are Gaussian noise and salt and paper noise. Gaussian noise can be suppressed by median filtering.

Next step of pre-processing is binarization. The filtered scanned image is a Gray scale which has 256 intensity levels, which is converted to binary image having 2 level intensity i.e., white and black using Otsu’s global threshold method. The binary image after thersholding have pixel value such that white pixel have logic ‘1’ and black pixel have ‘0’. For further steps the image is then inverted to make pixel value corresponds to character to ’1’. Form the inverted image connected pixels less than some threshold (say 30 pixel) is removed which corresponds to noise as shown Further bounding box is created which touches the character in four sides. It removes unnecessary black spaces. Then the image is resized to 32 × 64. Then perform thinning operation is performed to make the each character pixel to ’1’ - pixel width. Fig. 4 shows thinned image of Malayalam character ’a’.

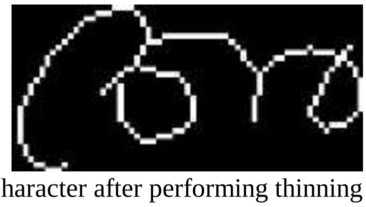

Fig. 4. Character after performing thinning operation

So, after preprocessing a thinned image of character resized to 32 × 64 is obtained. Which is suite for further processing

3.2 Feature Extraction

Feature extraction is most important step in Character recognition. Feature extraction is a special form dimensionality reduction. It reduces amount of resources required to describe a character. In this case the final output is pre-processing of size 32 x 64 (=2048 bits) by finding out feature we can represent same character by lesser number of bits.

1) Obtain horizontal and vertical projection of a character Projection profiles are found by counting number pixel having pixel value 1. In horizontal projection, count number of pixel having pixel value '1' is in each row. Similarly vertical projection, count number of pixel having pixel value '1' in each column. Fig. 5 shows the projection profile of a Malayalam character 'a'.

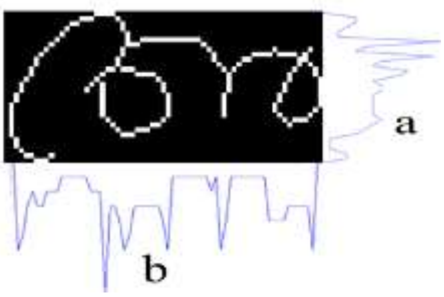

Fig. 5. Projection profile of character 'a' a) Horizontal projection b) Vertical Projection

2) Obtain discrete cosine transform (DCT) for each projection profile

3) Obtain feature vector by choosing First five DCT coefficients for each projection profile. So total of 10 DCT coefficients are used as feature vector for recognizing each character.

3.3 Classification

In the proposed system classification is achieved by using a feed forward network with 1 hidden layer with 10 nodes is used. 10 features after discrete cosine transform in feature vector is used as a input to the Neural network. Since system classifying 44 Malayalam character we use a 44 nodes in output layer as shown Fig. 6

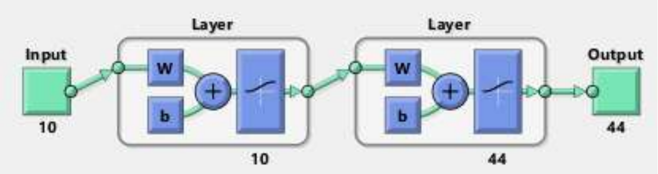

Fig. 6. Multi Layer Feed Forward Network

4. RESULTS AND DISCUSSION

The experiment is done using Matlab R2013a on a computer having Intel i3 processor and 6 Gb RAM. In the experiment trained the network using 40 samples of each 44 classes and a total of 1760 (40 x 44) samples. The trained network is tested using 10 samples of each classes. The recognition accuracy proposed system of 85.02with 44 class problem. To compute recognition accuracy the neural network is trained 10 times using the database. Each time the recognition accuracy is noted and average of all these is taken as recognition accuracy. Table. I shows the recognition rate for each character. From the table it is clear that lowest recognition rate that is obtained is 78.57% and highest is 89.88%. Most of the characters in 44 classes have recognition accuracy greater than 80%. Whole system recognition accuracy obtained is 85.02% with a feature vector of dimension 10.

5. APPLICATION

There are wide range of applications for an Optical Character recognition (OCR), some of the common application are listed bellow

1) Reading aid for the blind
2) Automatic text entry into the computer for Desktop publication, Library cataloging, Ledgering Automatic reading for sorting Postal mail, Bank cheques, Other documents
3) Document data compression, from document image
4) scanned to ASCII format
5) Automatic number plate recognition
6) Make electronic images of printed documents searchable

6. CONCLUSION

In the paper a malayalam character recognition system is proposed. The proposed makes use of 1-D wavelet transform of projection profile as feature. Then accuracy of recognition of the system for various neural network and for various wavelet based feature are found. it have been found, that neural network using softmax activation function at the output layer has maximum efficiency. All the reported literature in malayalam literature have claimed up to 33 classes only . But in this work 44 classes have been identified.

References

[1] [Online]. Available: http://scriptsource.org/scr/Mlym

[2] [Online]. Available: http://www.omniglot.com/writing/malayalam.htm

[3] CDAC, "Viswabharat@tdil," Journal of Language Technology, July 2003.

[4] G. Raju, "Recognition of unconstrained handwritten malayalam characters using zero-crossing of wavelet coefficients," Advanced Computing and Communications, pp. 217–221, 2006.

[5] R. John, G. Raju, and D. S. Guru, "1d wavelet transform of projection profiles for isolated handwritten malayalam character recognition," International Conference on Computational Intelligence and Multimedia Applications, pp. 481–485, 2007.

[6] M. A. Rahiman and M. S. Rajasree, "Printed malayalam character recognition using back-propagation neural networks," IEEE International Advance Computing Conference, pp. 197–201, 2009