International Journal of Advanced Technology in Engineering and Science
Vol. No.3, Special Issue No. 01, September 2015
www.ijates.com
ijates
ISSN 2348 - 7550

OFFLINE MALAYALAM CHARACTER RECOGNITION USING GENETIC ALGORITHM

Meenu Alex[1], Smija Das[2]

[1,2]Department of Computer Science, St. Joseph's College of Engineering & Technology Palai, Kerala, (India)

ABSTRACT

Offline Malayalam Character Recognition is an emerging area of pattern recognition and computer vision. This method deals with recognizing 33 isolated characters in Malayalam. The stages are preprocessing, feature extraction, classification and recognition. The dataset for the work was collected from people of different age group and professions. Mainly the data was collected from the staff and students of St. Joseph's College of Engineering, Palai affiliated to M.G University Kerala. The collected data undergoes preprocessing stage to remove as much as distortions as possible. For feature extraction, diagonal feature extraction scheme was employed. In the classification stage, a neural network along with the power of genetic algorithm was used. Because genetic algorithm give more productive result. The system was tested on 2 schemes: writer independent and writer dependent scheme. The system showed an overall accuracy of 81.73% for 33 Malayalam character classes.

Keywords: Character Recognition, Diagonal Feature, Neural Network

I. INTRODUCTION

Handwritten Character Recognition (HCR) is an emerging area in the fields of pattern recognition and computer vision. It has a wide range of applications in many fields. These include postal automation, automatic number plate recognition, CTS scanning, preservation of degraded documents, bank cheque processing etc. The aim of character recognition is to convert human readable characters which are present in a digitized or photographed sheet of paper and convert it into a machine editable form. The character recognition system can be of two types : online or offline. The online character recognition system is dynamic. The data are captured at the same time user writes on a digitizer with a stylus. It does the real time conversion of characters to their Unicode values. In an offline system, the data is captured by a scanner after the writing process.

There are so many challenges in the field of character recognition. There is variation in writing styles in between people. It may also vary in accordance with the emotions of the writer, the current situation and the writing condition. Another important feature of Malayalam language is its enormous character set. The identification of characters may be posing another challenge in the form of similarity between the characters. Adding up to the scene is the similarity in writing styles of different people. Handwritten character recognition is matured for foreign languages like English, Japanese, Chinese, Arabic etc. The reality is that recognition of scripts is a tedious process for South Indian languages like Malayalam, Kannada, Tamil, Telugu etc. This is

mainly due to the large character set, presence of compound characters and so on. In this study, we aim at identifying the different phases in Malayalam Character recognition and the methods employed for the process itself.

II. RELATED WORK

A lot of works are reported in foreign languages [2] in the domain of Handwritten Character Recognition. Among Indian languages like Devanagari, Tamil, Oriya and Bangla many works were occurred [3,4]. But in case of South Indian languages especially Malayalam, only few works were reported. A recognition system that can identify the complete character set of Malayalam is not developed till now. The main difficulty in Malayalam character recognition system is the lack of availability of a benchmarking database for comparison.

The first work in Malayalam Character Recognition was reported in 2007 by Lajish V.L [5]. It uses fuzzy-zoning method and normalized vector distance measures for the recognition of 44 Malayalam characters. Renju John et al [6] came forward with the concept of 1D Wavelet Transform of Projection Profiles for Isolated Handwritten Malayalam Character Recognition. Handwritten character recognition by applying Daubechie wavelet coefficients was proposed in [7]. M. Abdul Rahiman et al [8] proposed an HLH intensity pattern based method for recognition of Malayalam characters. A recognizing system for Malayalam characters using discrete features was introduced by Binu P. Chacko and Babu Anto in [9]. Jomy John et al [10] proposed a chain code histogram based method for recognizing vowels of Malayalam. Bindu S. Moni et al [11] invented a handwriting recognition system based on run length count (RLC). Vidya V. et al [12] proposed a method for handwritten character recognition based on Probabilistic Simplified Fuzzy ARTMAP (PSFAM). Features like Zernike moment features, cross feature, distance feature and fuzzy depth are extracted from the character. This method gained an accuracy of 79.48% for 142 Malayalam characters. In 2014, Shanjana C et al [13] proposed a method for Malayalam character segmentation. In this work, segmentation of characters is performed by combining Vertical projection profile method along with connected component analysis method.

III. CHARACTERISTICS OF MALAYALAM SCRIPTS

Malayalam is one among the regional languages in India which owes its origin to Sanskrit. Designated as a classical language in 2013, it has the reputation of being the second most difficult language to be proficient with. It is mainly used in the state of Kerala, Union territory of Lakshadweep and Mahe. Malayalam script is derived from Grantha script.

The letters of Malayalam script consists of curves and loops. The vastness of character set is yet another distinguishing factor of Malayalam. Malayalam characters can be basically categorized as vowels, consonants. It also contains 9 rarely used numerals. Another division of Malayalam Character comprises of the conjunct consonants and consonant diacritics.

Figure 3.1: Malayalam Vowel Set

ക ഖ ഗ ഘ ങ

ച ഛ ജ ഝ ഞ

ട ഠ ഡ ഢ ണ

ത ഥ ദ ധ ന

പ ഫ ബ ഭ മ

യ ര ല വ ശ

ഷ സ ഹ ള ഴ റ

Figure 3.2: Malayalam Consonant Set

IV. PROPOSED METHOD

The proposed system consists mainly of the stages given below.

- Image Acquisition
- Preprocessing
- Feature Extraction
- Classification
- Recognition

The layout of the system is given in figure 4.1. Initially the samples are collected from different people. The scanned image is first subjected to preprocessing to remove as much distortions as possible. After preprocessing, features are extracted and fed as input to neural network classifier. The output of this network is compared with those in the database and provides the correct recognition results. Diagonal based feature extraction method is used to extract features of each character. This method also attempted to use the power of genetic algorithm to recognize the character.

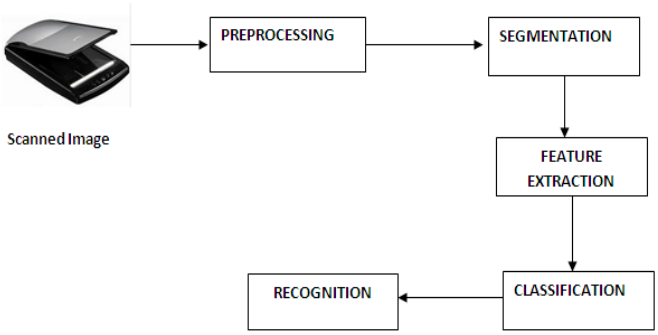

Fig 4.1: Proposed System

4.1 Image Acquisition

The images for the experiment are captured by using a scanner at 300 dpi resolution. The images can be in any format like JPEG, BMP, PNG etc. These images are given to the system as input for further steps. A sample scanned image is shown in figure 4.2.

ijates
ISSN 2348 - 7550

Fig 4.2: Original Image

4.2 Preprocessing

The aim of preprocessing is to remove as much as distortions as possible from the scanned image. Degraded documents or poor quality of scanner are responsible for these distortions. At first, the scanned image is converted to grayscale if it is in RGB or any other format. After that the grayscale image is converted to binary using Otsu's method of global thresholding. A Gaussian filter is used to remove noise. Then we are performing some morphological operations erode and dilate on the image. Erode performs binary erosion. Dilate performs binary dilation. The characters are cropped by placing a bounding box around it. The cropped characters are finally normalized to 100x100 size.

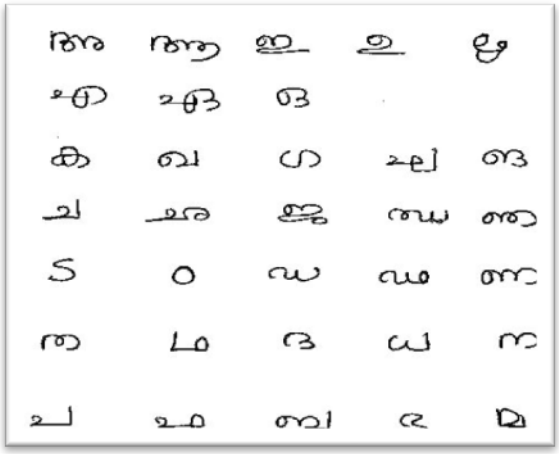

Figure 4.3: Binarized Image

4.3 Feature Extraction

The idea behind performing feature extraction is to extract the salient characteristics of the image. The success of a character recognition system relies on effective feature – classifier combination. Here we are extracting the diagonal feature from the character image. Diagonal feature extraction scheme for recognizing off-line handwritten characters is used in this work. Every image of size 90x60 pixels is divided into 54 equal zones, each of size 10x10 pixels. The features are extracted from each zone pixels by moving along diagonals of its respective 10x10 pixels. Each zone has 19 sub-features values are averaged to form a single feature value and

placed in the corresponding zone. This procedure is sequentially repeated for the all the zones. There could be some zones whose diagonals are empty of foreground pixels. The feature values corresponding to these zones are zero. Finally, 54 features are extracted for each character.

4.4 Classification

The decision making part as well as the final stage of a character recognition system is classification. In this stage, unique labels are assigned to each character image based on the extracted features.

1) *Neural Network* – It is a network that learns from observed data. Neural Network learns slow but have fast prediction capacity. This can perform tasks which cannot be performed by a linear program.

4.5 Recognition

The Chromosome bit string from the Chromosome generation function is used to recognize Malayalam character by comparing the fitness value of an unknown character with all the Malayalam character which is store in database during training process using neural network. The highest fitness value is the recognition result.

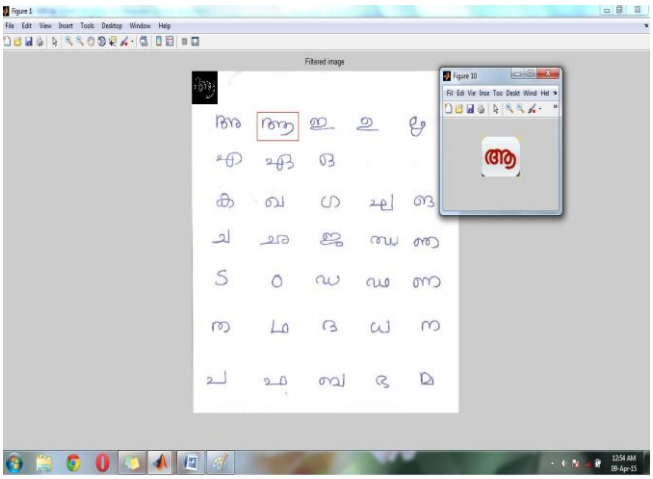

Figure 4.4: Correct Recognition

The fitness value is calculated as follows:

$$\text{Fitness value} = \sum_{i=1}^{n} S - L$$

V. RESULTS AND DISCUSSIONS

The recognition system has been implemented using Matlab R2012a. The experiment was conducted on 33 characters-8 isolated vowels and 25 consonants of Malayalam character set. A standard benchmark database is not available for Malayalam. The input dataset was collected from people at different age groups, different professions, sex etc. The scanned image is taken as dataset /input. The experiment is conducted on more than 1155 character samples. The testing characters are separated into data sets, the training data set and testing data set. The segmented characters are stored as 100x100 BMP images. The neural network used here consists of one input layer, three intermediate layer and one output layer. We have tested the system according to two different schemes: Writer dependant and writer independent schemes. In the writer dependant scheme, writing

samples of people which were used in training were subjected to testing whereas in the writer independent scheme, the writing samples of new users whose samples were not used for training were subjected to testing.

For the writer dependent scheme, the database consists of 825 samples of handwritten data collected from 25 people belonging to different age groups and professions and for the writer independant scheme, the database consists of another 330 samples collected from 5 different people. Compared to writer independent scheme, a higher recognition accuracy was obtained for the writer dependant scheme. An accuracy of 86.54% was obtained for writer dependent scheme and an accuracy of 69.69% was obtained for writer independent scheme. The overall accuracy obtained for this system is 81.73%.

VI. CONCLUSION

A method was deployed for handwritten Malayalam character recognition. This method uses neural network for classification. The success of any character recognition system depends on the feature and classifier which is used to classify the unknown input to well define class. The method employed here uses a diagonal feature extraction method. Character recognition in Malayalam is a very tedious process. This is mainly due to the presence of large character set.

The main highlight of this system is that this method attempts to use the power of genetic algorithm to recognize the character. Using genetic algorithm which is less use in Malayalam and any other Indian languages is more productive. This system is deployed as two schemes : writer dependent and writer independent scheme. This system showed an overall accuracy of 81.73 percentage for a set of 33 Malayalam character classes.

REFERENCES

[1] R. R. Plamondan, S.N. Srihari, “Online and offline handwriting recognition: A comprehensive survey”,IEEE Trans. On PAMI, Vol22(1) pp 63-84,2000.

[2] Jannoud, I.A.: Automatic Arabic Hand Written Text Recognition System. American Journal of Applied Sciences (2007)

[3] Ved Prakash Agnihotri “Offline Handwritten Devanagari Script Recognition”, I.J. Information Technology and Computer Science, 2012, 8, 37-42

[4] Pal, U., et al.: Handwritten Bangla Compound Character Recognition Using Gradient Feature. In: 10th International Conference on Information Technology (2007)

[5] Lajish V. L., “Handwritten character recognition using perpetual fuzzy zoning and class modular neural networks”, Proc. 4th Int. National conf. on Innovations in IT, 2007, pp 188-192

[6] R. John, G. Raju and D. S. Guru, “1D Wavelet transform of projection profiles for isolated handwritten character recognition”, Proc. Of ICCIMA07, Sivakasi, 2007, 481-485, Dec 13-15K.

[7] G. Raju, “Recognition of unconstrained handwritten Malayalam characters using zero-crossing of wavelet coefficients”, Proc. of 14th International conference on Advanced Computing and Communications, 2006, pp 217-221.

[8] M. A. Rahiman et. al., “Isolated handwritten Malayalam character recognition using HLH intensity patterns”, 2010 Second International Conference on Machine Learning and Computing.

[9] Binu P. Chacko, Babu Anto P, "Discrete Curve Evolution Based Skeleton Pruning for Character Recognition", Seventh International Conference on Advances in Pattern Recognition, 2009.

[10] Jomy John, Pramod K. V, Kannan Balakrishnan "Offline Handwritten Malayalam Character Recognition Based on Chain Code Histogram", Proceedings Of ICETECT 2011.

[11] Bindu S Moni, G Raju, "Modified Quadratic Classifier for Handwritten Malayalam Character Recognition using Run length Count", In International Conference IEEE, 2011.

[12] Vidya V, Indhu T R, Bhadran V K,R Ravindra Kumar, "Malayalam Offline Handwritten Recognition using Probabilistic Simplified Fuzzy ARTMAP", Advances in Intelligent Systems and Computing Volume 182, 2013, pp 273-283.

[13] Shanjana C, Ajay James, "Character Segmentation in Malayalam Handwritten Documents", IEEE International Conference on Advances in Engineering & Technology Research (ICAETR - 2014), August 2014

[14] B. Anuradha and B. Koteswarra; "An efficient Binarization technique for old documents", Proc.of International conference on Systemics,Cybernetics and Inforrmatics,Hyderabad, pp771-775,2006

[15] Bindu S Moni, G Raju, "Modified Quadratic Classifier and Directional Features for Handwritten Malayalam Character Recognition", IJCA Special Issue on Computational Science - New Dimensions Perspectives NCCSE, 2011.

[16] Jomy John, Pramod K. V., Kannan Balakrishnan, "Unconstrained Handwritten Malayalam Character Recognition using Wavelet Transform and Support vector Machine Classifier", In International Conference oncommunication Technology and System Design, ELSEVIER 2011.

[17] Lajish V. L., "Handwritten character recognition using gray scale based state space parameters and class modular NN",Proc. 4th Int. National conf. on Innovations in IT, 2007, 374-379.

[18] Abdul Rahiman M, M. S. Rajasree, Masha N, Rema M ,Meenakshi R, Manoj Kumar G, "Recognition of Handwritten Malayalam Characters using Vertical Horizontal Line Positional Analyzer Algorithm", IEEE International Conference 2011.

[19] Arica, N., Yarman-Vural, F.T.: An Overview of Character Recognition Focused on Off-Line Handwriting. IEEE Transactions on System, Man and Cybernetics –Part C: Applications and Reviews (2001)

[20] Pal, U., et al.: A System for Off-line Oriya Handwritten Character Recognition using Curvature Feature. In: 10th International Conference on Information Technology (2007)

[21] Meenu Alex, Smija Das: A Study on offline character recognition in Malayalam scripts. Proc. Of International Conference on Emerging Trends in Technology and Applied Sciences.p. 47,April-May 2015.