NOVATEUR PUBLICATIONS
JournalNX- A Multidisciplinary Peer Reviewed Journal
ISSN No: 2581 - 4230
VOLUME 4, ISSUE 7, July -2018

HANDWRITTEN CHARACTER RECOGNITION OF SIMILAR LOOKING LETTERS USING ANGULAR FEATURE EXTRACTION METHOD

For Malayalam Script

APARNA S MENON
Research Student, Electronics and Communication Dept.
NSS College of Engineering, Palakkad, Kerala, India.
Email: thanalmenon@gmail.com

KALA L
Associate Professor, Electronics and Communication Dept.
NSS College of Engineering, Palakkad, Kerala, India.
Email: kalalalitha97@gmail.com

ABSTRACT:

Classification of similar looking handwritten characters in Indian languages is challenging. Especially in the case of Malayalam, a Dravidian language from the Brahmic script which is characterized by its highly curvy and looped nature. So by considering such a nature of the language, a new feature extraction method in the field of HCR (Handwritten Character Recognition) named as Angular Method is used by using fuzzy logic classifiers with k –NN algorithm.

KEYWORDS: HCR, k –NN, Fuzzy logic and machine learning.

I. INTRODUCTION:

HCR is one of the most demanded and most needed technologies in this smart world which moves more and more paper less. Recognition of handwritten character is not only a need for Industries and Educational establishments but also for a common man to make his works easier and in saving time wastage on unproductive paper works. Though these works are

Highly efficient and highly appreciated in a developed or majorly mono or bi – lingual society, in developing and multi – lingual population like India, the work done is comparatively lesser and even complex due to the huge amount of scripted and non-scripted languages and due to its local dialects. Previous works on various Indian languages are as follows which uses various types of feature extraction and classification methods for character recognition purpose are on, Bangla [4] [6] [7] [8] [10] [16] [18] [19] [20], Devanagari [2] [4] [5] [6] [7] [8] [9] [12] [10] [15] [16] [17] [18] [19] [20], Gujrati [7] [11] [20], Gurmukhi [7] [14] [18] [20], Kannada [5] [9] [7] [8] [15] [20], Kashmiri [13], Malayalam [4] [5] [9] [15] [18] [20] [21], Odiya [4] [7] [8] [18] [20], Sanskrit [11], Tamil [5] [7] [8] [15] [20], Telugu [5] [7] [8] [9] [18] [20] and Urdu [4] [7] [10] [16] [18] [20].

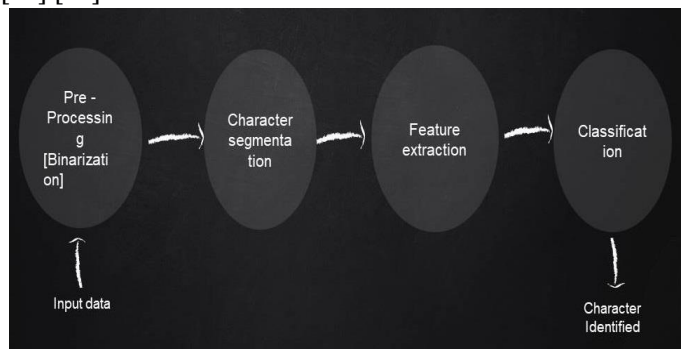

Figure 1: Layout of a character Recognition system

A basic character recognition system consists of four parts that are, pre – processing, segmentation of the character, feature extraction and classification as given in Figure 1. So this paper is a work on a newly implemented feature extraction method called 'Angular method' which considers angular nature of each letter as its feature and also uses comparatively lesser amount of feature set for recognition purpose. In this work the highly complex work of categorizing the similar looking single character letters on handwritten data of Malayalam language script is done using fuzzy logic with k -NN and also its recognition accuracy is found out.

The paper is arranged as follows, Data collection and pre –processing techniques are discussed in section II, where feature extraction techniques explained in section III, classification algorithms used are discussed in section IV, results and its analysis are explained in section V and conclusion and future works are explained in section VI.

II. DATA COLLECTION AND PRE – PROCESSING:

Since a standard handwritten database is not available for Malayalam, 21 samples of, two pair of data on similar looking Malayalam characters is collected from people of various age groups for all the languages

implemented in the work. Where, a pair consists of two similar looking letters Malayalam.

The image data are scanned and binarized into a global threshold value and is proceeded for character segmentation. Each data set is then segmented into individual letters using the bounding box method, which segments each letter into same size rate. Also the unintelligible sets of segmented characters are removed from the character set. Figure 2 shows the Pre – Processing flow diagram of an HCR.

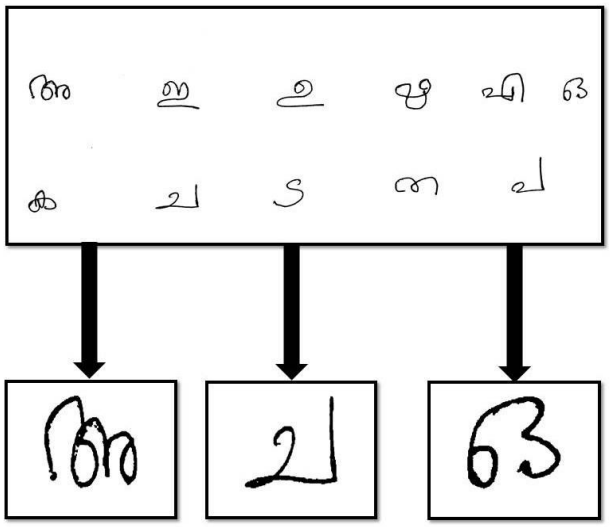

Figure 2: Pre - processing representation of Malayalam Characters

III. FEATURE EXTRACTION METHOD:

The performance of a HCR (Handwritten Character Recognition) system depends greatly on its extracted feature set. In digital image processing, Feature extraction can be explained as the representation of an image or image patch. This step in HCR helps in extraction of useful information for character recognition and for removal of any extraneous information from the considered image. Figure 3 shows the pictorial representation of angular method used. Steps used for feature extraction using Angular method are given below:

- Centroid of each of the character blob is found for both training and test data.
- Find the distance from the centroid to the maximum distant point on the boundary along following 16 directions for both training and test data using Euclidean distance method, i.e. 0 deg, 22.5 deg, 45 deg,90 deg, 112.5 deg, 135 deg, 157.5 deg, 180 deg, 202.5 deg, 225 deg, 247.5 deg, 270 deg, 292.5 deg, 315 deg & 337.5 deg.
- Feature-set threshold value is created by taking the average of the obtained distance value along all the angles for each of the training data characters.

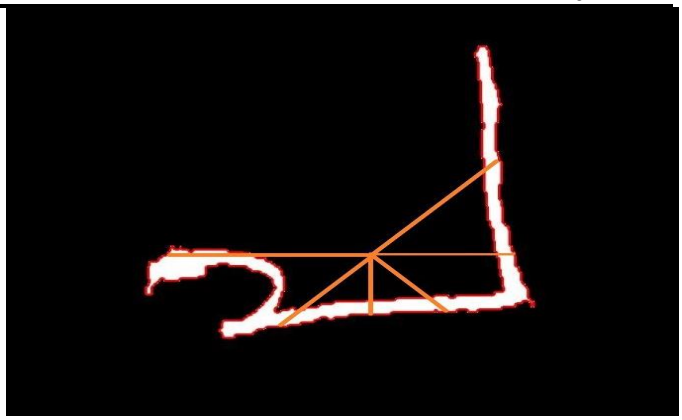

Figure 3: Representation of angular method

IV. CLASSIFICATION ALGORITHM:

Classification in an HCR is the method of categorizing a character from other with respect to extracted feature threshold value. Fuzzy logic with k – NN algorithm is used for classification in this work. A fuzzy logic algorithm is a true or false classifier which classifies the letter showing its feature set value above the threshold as true and else false. The steps used for classification is explained below.

- The study is conducted on 2 pairs of similar looking single character Malayalam letter.
- Each set of data is taken and is undergone with all the preprocessing and extraction methods.
- A matrix is calculated for the test data with the feature-set threshold value using k-NN classification method. The k – NN matrix is found by finding the RMS value between the pre –calculated threshold value and the feature set value of the input test data.
- Change in angular feature value of similar looking characters would be only varying for one or a few angle sets. So values in these angles are considered for differentiating the similar characters. The minimal feature angle to be fed into the fuzzy logic is set by:
  - Analyzing the unique properties of the selected characters along its feature value by comparing the threshold value to the test data value.
  - The most recognizable feature angle(s) are selected and used to set threshold value in fuzzy logic method.
  - Where, fuzzy logic method only considers critical angle values which have been previously selected as the characteristic angle of the special character.
- So considering that angle which distinguishes the characters and thus classifying these characters using fuzzy logic i.e. if the k –NN matrix crosses the threshold, character 1 is detected or else character 2 is detected and thus similar looking characters are recognized.

V. RESULTS AND DISCUSSION:

In this work 21 Malayalam datasets are used for analysis. The character detection accuracy for each of the similar letter of Malayalam is tabulated in TABLE I. Here thus it's visible that since Malayalam script is angular in nature, a good accuracy is attained and even similar letters could be easily distinguished from other just by considering 16 features for extraction purpose. This shows that it is a very suitable method for HCR of angular natured languages by using less parameter for classification.

TABLE I. CLASSIFICATION ACCURACY

S. No.	Character	Recognition Accuracy
1.	അ (*'A'*)	95.24%
	ആ (*'AA'*)	90%
2.	എ (*'AE'*)	75%
	ഏ (*'AEE'*)	100%

a. Recognition Accuracy percentage of similar looking Handwritten Characters in Malayalam.

VI. CONCLUSIONS AND FUTURE WORKS:

The presented work is done on similar looking letters of Malayalam language. The extraction of features are done using a new method in the field of HCR named Angular Method, which is used for classification using Fuzzy logic where recognition accuracy of each language is given in TABLE I. From that it is visible that the angular feature extraction method works well for curvy and highly angular natured language like Malayalam that even minute variations between similar looking handwritten characters could be identified just by considering 16 values as the feature of character, with a good percentage which increases the scope of the method in character recognition and in machine learning domains.

The obtained recognition accuracy can be improved by increasing the amount of training data and thus achieving a better conclusion on idea about handwritten angular properties. Also applying the feature extraction method on various classification algorithms could also find a change in accuracy levels. This work can be applied on real time works since even a minute change in characters could be detected with ease. The authors hope that the paper would benefit researchers who are working on handwritten character recognition domain with their future works.

REFERENCES:

1) B. B. Chaudhuri and U. Pal, "An OCR system to read two Indian language scripts: Bangla and Devnagari (Hindi),'' in Proc. ICDAR, Ulm, Germnay, 1997, pp. 1011-1015.
2) S. A. Chaudhari and R. M. Gulati, ``An OCR for separation and identification of mixed English-Gujarati digits using kNN classifier,'' in Proc.
3) K. Ubul, G. Tursun, A. Aysa, D. Impedovo, G. Pirlo, and T. Yibulayin, "Script Identification of Multi-Script Documents: A Survey," IEEE, Vol.5, pp. 6546- 6559.
4) K. Roy, S. K. Das, and S. M. Obaidullah, ``Script identifcation from handwritten document,'' in Proc. NCVPRIPG, Dec. 2011, pp. 66-69.
5) M. Hangarge, K. C. Santosh, and R. Pardeshi, ``Directional discrete cosine transform for handwritten script identifcation,'' in Proc. ICDAR, Washington, DC, USA, Aug. 2013, pp. 344-348.
6) U. Bhattacharya and B. B. Chaudhuri, ``Handwritten numeral databases of Indian scripts and multistage recognition of mixed numerals,'' IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 3, pp. 444-457, Mar. 2009.
7) R. Pardeshi, B. B. Chaudhuri, M. Hangarge, K.C. Santosh, "Automatic Handwritten Indian Scripts Identification," in Proc. ICFHR, Heraklion, Greece, Dec.2014,pp.375-380.
8) U. Pal, N. Sharma, T. Wakabayashi, and F. Kimura, ``Handwritten numeral recognition of six popular Indian scripts,'' in Proc. ICDAR, Parana, Sep. 2007, pp. 749-753.
9) L. Pauly, R.D. Raj, Dr.B. Paul, "Hand written Digit Recognition System for South Indian Languages using Artificial Neural Networks ," in Proc. IC3, Noida, India, 2015.
10) S. M. Obaidullah, C. Halder, N. Das, and K. Roy, "Numeral script identification from handwritten document images,'' in Proc. 11th IMCIP, 2015, pp. 585-594.
11) A. Busch, W. W. Boles, and S. Sridharan,"Texture for script identication,'' IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 11, pp. 1720-1732, Nov. 2005.
12) M. S. Das, D. S. Rani, and C. R. K. Reddy, ``Heuristic based script identifcation from multilingual text documents,'' in Proc. Int. Conf. Recent Adv. Inf. Technol. (RAIT), Dhanbad, India, Mar. 2012, pp. 487-492.
13) R. Bashir and S. Quadri, "Identifcation of Kashmiri script in a bilingual document image,'' in Proc. ICIIP, Shimla, India, Dec. 2013, pp. 575-579.
14) R. Rani, R. Dhir, and G. S. Lehal, ``Script identifcation of pre-segmented multi-font characters and digits,'' in Proc. ICDAR, Washington, DC, USA, Aug. 2013, pp. 1150-1154.
15) S. A. Angadi and M. M. Kodabagi, ``A fuzzy approach for word level script identifcation of text in low resolution display board images using wavelet features,'' in Proc. ICACCI, Aug. 2013, pp. 1804-1811.

16) S. Chanda, S. Pal, K. Franke, and U. Pal, "Two-stage approach for word-wise script identification,'' in Proc. ICDAR, Jul. 2009, pp. 926-930.
17) S. Chanda and U. Pal, ``English, devnagari and urdu text identification,'' in Proc. Int. Conf. Cognit. Recognit., 2005, pp. 538-546.
18) S. Ghosh and B. B. Chaudhuri, ``Composite script identification and orientation detection for Indian text images,'' in Proc. ICDAR, Beijing, China, Sep. 2011, pp. 294-298.
19) U. Pal and B. B. Chaudhuri, ``Identification of different script lines from multi-script documents,'' Image Vis. Comput., vol. 20, nos. 13-14, pp. 945-954, Dec. 2002.
20) S. Chanda, K. Franke, and U. Pal, ``Identification of Indic scripts on torn documents,'' in Proc. ICDAR, Beijing, China, Sep. 2011, pp. 713-717.
21) Chacko A.M.M.O., Dhanya P.M. (2015), "A Comparative Study of Different Feature Extraction Techniques for Offline Malayalam Character Recognition," In: Jain L., Behera H., Mandal J., Mohapatra D. (eds) Computational Intelligence in Data Mining - Volume 2. Smart Innovation, Systems and Technologies, vol 32. Springer, New Delhi.