A Novel Method on Malayalam Handwritten Character Recognition

Anish S, Preeja V

Abstract— Handwritten Character Recognition (HCR) is one of the most challenging and active areas of research in the field of pattern recognition. It has a wide range of applications like preservation of documents into digital form, managing rare books etc. HCR is a difficult process due to the variants of handwriting styles of different individuals. Thus the success rate of any HCR system greatly depends upon the language that these systems are working on, and the amount of character sets in each language. Malayalam, a south Indian language and official language in the state of Kerala has a rich amount of character sets. Recognizing all those characters is a difficult task. In any types of character recognition systems, recognition rates play a vital role in the overall efficiency of the system. Several researches are going on this field to improve recognition rates. This paper deals with texture extraction model for character recognition process. In this model co-occurrence matrix and Euclidean distance are used to recognize the characters in an image.

Index Terms—Binarization, Co-occurrence Matrix, Euclidean Distance, Segmentation

I. INTRODUCTION

Handwritten Character Recognition (HCR) is the process of recognizing handwritten characters from an image. It is the conversion of image representation of a document into a digital format. HCR can be classified into 2 types, based on how the characters are taken as input. One method is writing set of characters on a paper and the scanned image of that paper becomes the input to the recognition systems. Other method is to write in a digitizer using a digital pen and that becomes input to the recognition systems. The former method is called Offline Handwritten character recognition and the latter is called Online Handwritten Character Recognition. This paper discuss on Offline Character Recognition Systems in Malayalam language. Generally, Offline HCR consists of four stages: Pre-processing, Segmentation, Feature Extraction and Classification. Pre-Processing is the removal of irrelevant information from the image. Thus the images obtained after pre-processing is free from all kinds of noises. Segmentation isolates individual characters from the handwritten text. Feature Extraction is the process of extracting relevant features from the image. Classification is the process of grouping into appropriate characters based on the features that are obtained from the extraction phase. Thus the success rate of any character recognition systems greatly depends upon the selection of efficient feature extraction method. There are several researches going in this field in order to find out efficient feature extraction-classification pair.

Revised Version Manuscript Received on August 18, 2015.

Anish S, M.Tech Student, Department of Computer Science and Engineering, Sree Chitra Thirunal College of Engineering, Pappanamcode, Trivandrum, India.

Preeja V, Assistant Professor, Department of Computer Science and Engineering, Sree Chitra Thirunal College of Engineering, Pappanamcode, Trivandrum, India.

II. RELATED WORKS

Malayalam Handwritten Character Recognition is very challenging and active area in the field of pattern recognition. In recent years, a lot of algorithms have been proposed to improve the recognition rates. Most of the algorithms were concentrated to recognize simple Malayalam characters. As considering the complexity and nature of Malayalam characters, it is indeed a difficult task to recognize all those characters. Thus all those works reported was not able to recognize complex and conjunct characters.

M S Rajasree, M Abdul Rahiman [1] proposed a method for Malayalam character recognition using Daubechies Wavelet for feature extraction and Back Propagation Neural Networks for recognition. Here, they extracted features using multi-resolution wavelets that give better recognition rates. Recognition is done through feed forward Back propagation neural networks and gives better learning phenomenon through back propagation. This method of recognition focuses on recognizing popular typefaces not on complex characters. This method achieved an accuracy of 92%. A neural network based model for character recognition proposed by Gaurav Kumar, Pradeep Kumar Bhatia [2] uses Fourier Transform method for feature extraction and for the recognition of characters neural network [3] is used. It achieved an accuracy of 93%. Another method of character recognition using Wavelet transforms [4][5] and Support Vector Machine was proposed by Jomy John, Pramod K.V, Kannan Balakrishnan [6]. This method achieved 90.25% accuracy. Another method was proposed by M S Rajasree, M Abdul Rahiman [7] using Vertical and Horizontal Analyzer Algorithm for feature extraction and Decision tree for classification. It recognizes basic Malayalam character sets. An accuracy of 91% was achieved in this method.

A method proposed by M S Rajasree and Abdul Rahiman was taken as a study to find the performance of that algorithm that uses Vertical and Horizontal Line Positional Analyzer Algorithm for feature extraction and for Classification it uses decision tree. After testing this method, it is found that it only works for basic Malayalam character sets. In order to improve recognition rates, a method that uses co-occurrence matrix and Euclidian distance is used for the recognition process. It is found that it has achieved high recognition rates compared to the above method of recognition.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

III. METHODOLOGY

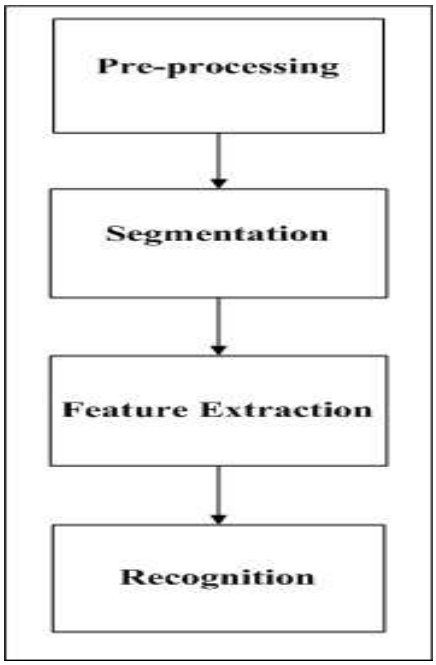

Fig. 1: Stages in Handwritten Character Recognition Process

First, a handwritten character image is given as an input to the recognition system. Then the pre-processing technique is applied on to the image to get itself prepared for the recognition process. It is then followed by segmentation process, which isolates each character from the image. Then, feature extraction process is carried out to extract suitable features from the image and finally based on the features extracted it is classified into appropriate characters. The step wise implementation of the algorithm is as follows.

Step 1:- Pre-processing

It is the process of removing irrelevant information present in the image. The input to the recognition system is the scanned image of the handwritten text. It contains lot of noise, so removing all those kinds of noises is the role of pre-processing technique [8][9]. It includes two processes.

1. Gray conversion

It is the conversion of an image having 32 bits/pixel into 8 bits/pixel.

2. Binary conversion

Binarization is the process of converting greyscale images to binary images. It is done in order to identify the objects of interest from the image. It separates the foreground pixels from the background pixels.

Step 2:- Segmentation

It isolates individual characters from the handwritten text. Segmentation process includes two steps:-

1. Line Segmentation

In this process, the written document is separated into line of characters. Thus, within those lines the entire characters are located.

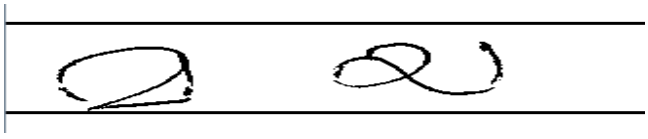

Fig. 2: Line Segmentation

2. Character Segmentation

In this process, each character in a line is subjected to the character separation process. Here the characters in the line are separated into individual ones that simplify further process of character extraction process.

Fig. 3: Character Segmentation

Step 3:- Feature Extraction

Feature extraction is done to find the set of parameters that can be used to define each character uniquely and precisely. Feature extraction plays an important role because of its effect on the capability of classifiers and ultimately on final accuracy. Texture based feature extraction model is used in the feature extraction phase.

Steps to extract feature from the image using Texture Extraction Method

1. Generate Co-occurrence Matrix
2. Extract feature values from the matrix and generate distance values for each character
3. Calculate Euclidean distance between the train and test characters and select the one those having least Euclidean distance.

Co-occurrence Matrix

Co-occurrence Matrix method is a statistical feature extraction method for global feature extraction. The Co-occurrence Matrix functions characterize the texture of an image by calculating how often pairs of pixel with specific values and in a specified spatial relationship occur in an image, creating a Co-occurrence Matrix, and then extracting statistical measures from this matrix.

Co-occurrence Matrix directions of Analysis

1. Horizontal(0^0)
2. Vertical(90^0)
3. Diagonal(45^0 and 135^0)

The direction of analysis [10] for different angular orientation is shown in Fig.4.

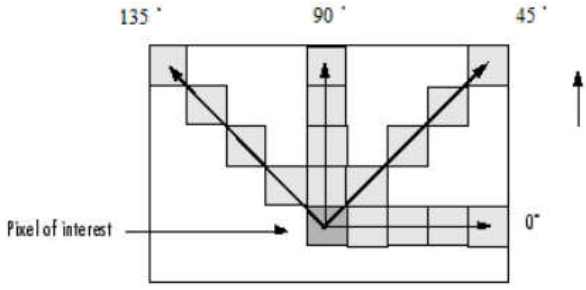

Fig. 4: Co-occurrence matrix direction of analysis

A sample image and the structure of co-occurrence matrix are shown in Fig.5 (a) and Fig.5 (b) respectively. The sample image taken is of a binary image. Its values are of 0's and 1's. The Fig.5 (b) shows the way to generate co-occurrence matrix. The '#' value indicates the number of times that particular pair of value appears in the image (i.e. co-occurrence).

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

0	0	1	1
0	0	1	1
1	0	1	1
0	1	0	1

Fig.5 (a): Sample Image

i/j	0	1
0	#(0,0)	#(0,1)
1	#(1,0)	#(1,1)

Fig.5 (b) Structure of co-occurrence matrix

The corresponding co-occurrence matrix of the above sample image is shown in the Fig.6. Here the notation h (d,θ) denotes the co-occurrence matrix with distance 'd' and angle' θ'.

4	6
7	6

$h_{1,0}^{0}$

Fig. 6: Co-occurrence matrix with d= 1 and θ = 0^0

Step 4:- Recognition

Recognition of characters is done through co-occurrence matrix and Euclidian distance between the training set data and the testing data. For each character in the training data set, 48 feature values are generated using co-occurrence matrix. These values are the intensity distribution around the character image in all possible direction of analysis and for the distance of computation are equal to 3. The distance of value 3 gives the more intensity distribution of the character that itself gives more accuracy. In the testing phase, Euclidian distance is computed between the testing data and each trained ones. Euclidean distance is usually the right measure for comparing cases. The basis of many measures of similarity and dissimilarity is Euclidean distance. Thus after obtaining the Euclidian distance between the trained one and the testing data, those characters having least Euclidian distance is selected as the recognised one.

IV. Results and Discussions

Training Phase and Testing Phase are done to evaluate the performance of the system. In the training phase, 44 characters are taken into consideration. The intensity distribution of each of the 44 characters are calculated and stored. Thus for each characters in the training phase a total of 48 values are generated. In the testing phase, the testing data's intensity values are compared with the trained data's values to find out the similarities between those characters. Euclidian distance computation is used to find out the similarities between the trained one and the testing one. Similarities are measured by using Euclidian distance.

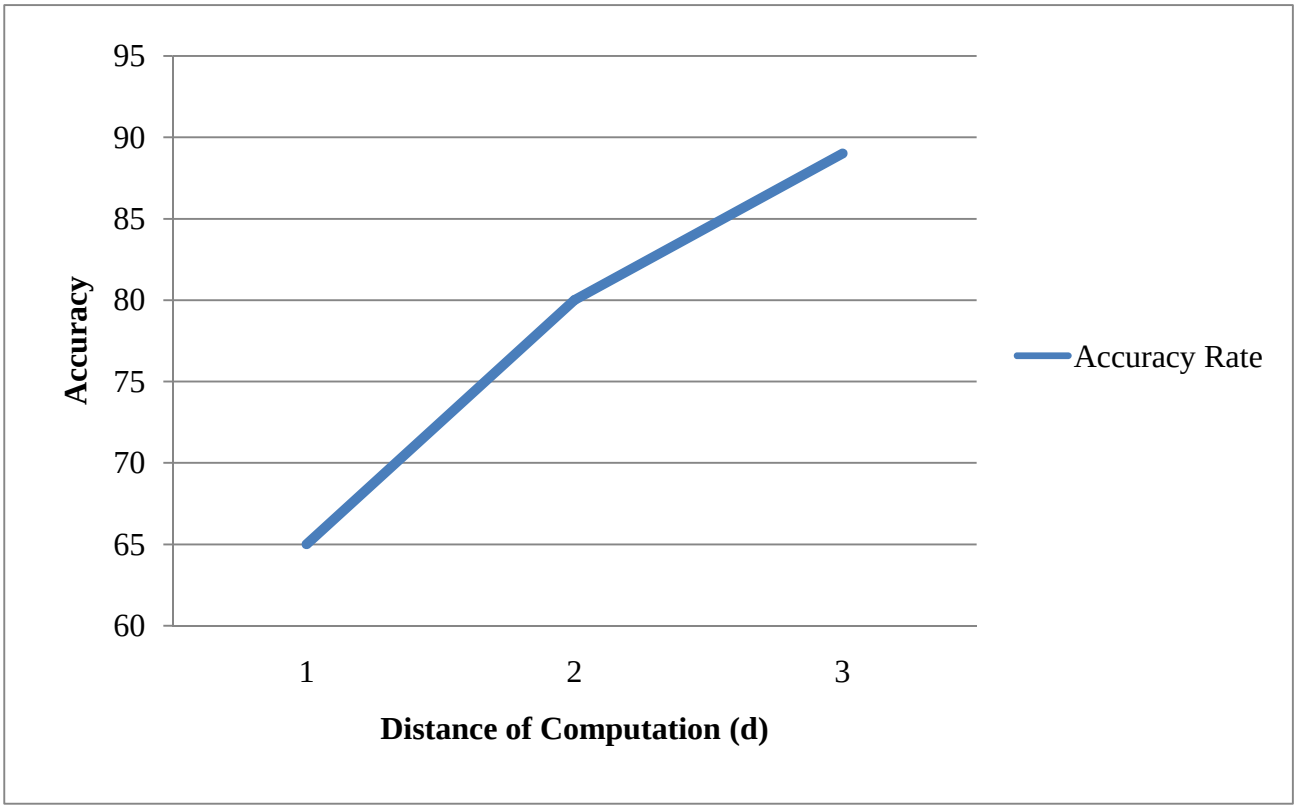

Fig. 7: Accuracy Rate Graph

The distance between vectors X and Y is defined as follows:

$$d(X,Y) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2} \quad \text{---------- (1)}$$

A graph showing the accuracy rate comparison is shown in the Fig.7. Accuracy is computed using the parameter distance of computation (d). If d=1, then the system provides an accuracy of 65% and if d=2, it gives an accuracy of 80% and finally if d=3, it gives an accuracy of 89%. Our system is tested with d=3 and has achieved an accuracy of 89%.

% Accuracy=

$$\frac{\text{No:of characters found correctly}}{\text{Total No:of patterns}} \times 100 \quad \text{-------- (2)}$$

V. Conclusion and Future Work

Recognition approaches heavily depend on the nature of data to be recognized. This paper deals with offline character recognition, thus appropriate recognition methods are to be used in order to improve recognition rates. Firstly, we have tested Vertical and Horizontal Line positional algorithm for feature extraction and Decision Tree for classification. But it is found that it deals with only basic Malayalam character sets

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

and also imposes several constraints. Thus it adds the complexity to the system. Thus we used texture based feature extraction model for recognition process. Texture based model performs better and also improves recognition accuracy as well. The great advantage about texture based model is that it can be used to handle complex characters as well. The greatest difficulty about character recognition is the presence of highly similar characters. Thus the recognition system sometimes gives unfair results for similar characters. Further research may be done to deal with complex characters.

REFERENCES

1. Abdul Rahiman M, M S Rajasree "Printed Malayalam character recognition using back propagation neural networks" International Advanced Computing Conference ,IACC 2009.
2. Gaurav Kumar, Pradeep Kumar Bhatia "Neural Network based approach for recognition of text images" International Journal of Computer Applications, 2013.
3. Lajish V.L "" Handwritten Character recognition using perpetual fuzzy zoning and class modular neural networks,"Proc.Of 4th Int.National Conf.on Innovations in IT,pp.188-192,2007.
4. G.Raju, "Wavelet transform and projection profiles in handwritten character recognition-a performance analysis" Proc.Of 16th International Conference on Advanced Computing and Communications,pp.309-314,2008.
5. G.R John, D.Guru "1-D wavelet transform of projection profiles for isolated handwritten character recognition" Proc.Of ICCIMA07, Sivakasi, pp.481-485, 2007.
6. Jomy John, Pramod K.V, Kannan Balakrishnan "Unconstrained Handwritten Malayalam Recognition using Wavelet Transform and Support Vector Machine Classifier" International Conference on Communication Technology and System Design 2011.
7. Abdul Rahiman M, M S Rajasree "Recognition of Handwritten Malayalam Characters using Vertical and Horizontal Line Positional Analyzer Algorithm" International Conference on Machine Learning and Computing(ICMLC 2011).
8. Ostu.N "A threshold selection method from gray level histograms" IEEE Trans.Systems, Man and Cybernetics, Vol.9, pp.62-66, 1979.
9. Lajish V.L "Handwritten character recognition using gray scale based state space parameters and class modular neural networks " Proc.Of 4th Int.National Conf.on Innovations in IT,pp.374-379,2007.
10. A. Materka and M. Strzelecki, "Texture analysis methods – a review," http://www.eletel.p.lodz.pl, 2010.

Anish S is currently doing his M.Tech. in Computer Science and Engineering at Sree Chitra Thirunal College of Engineering under University of Kerala, Trivandrum, Kerala, India. He received his B Tech Degree in Computer Science and Engineering from University College of Engineering, Karyavattom under University of Kerala in 2011. He concentrates mainly on Image Processing and Pattern Recognition.

Preeja V is working as Assistant professor at the department of computer science and engineering, Sree Chitra Thirunal College of Engineering, Trivandrum, Kerala. She did her B.Tech degree at College of engineering, Chengannur from CUSAT and M.Tech degree at NIT Calicut. She published her research works in many international journals and her area of interest mainly on Data structures and Database.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.