ISSN No. 0976-5697

Volume 8, No. 3, March – April 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Typed Malayalam Character Recognition

Amalu Johns
Department of Computer Science
Christ University
Bangalore, India

Jibrael Jos
Department of Computer Science
Christ University
Bangalore, India

Abstract: Malayalam is an Indian language .It is mainly using in Kerala one of the states in India. It is very difficult that to recognize Malayalam characters because it contains large number of curls, curves, loops etc... Many characters seem similar to each other. Many OCR systems are already available which can convert the printed or handwritten characters into machine encoded form but comparatively less number of systems are there to identify Malayalam characters. Here the recognition is performed through many stages. Starting from image acquisition for taking the input images, then the main task is that to remove noise from the images and thinning them. Thinning is performed to easily recognize each pixel from the character. Segmentation and feature extraction are the main procedures used for this. Based on the extracted features, characters are classified. The main application of this procedure is that it will help in effortlessly understanding the pronunciation of a word in different languages. This system obtains typed Malayalam characters from the internet and converts them into images which are then analyzed to identify the character.

Keywords: pre-processing, feature extraction, classification, segmentation

I. Introduction

This paper is focused on printed Malayalam character recognition. Recognition in Malayalam language is a tough task because each character is rich in features, complex shape and sizes. This technology changes over images of machine-printed characters into machine-readable characters. It aims at enabling machines to recognize optical symbols without human mediation. The stages which are involved in this recognition are segmentation, pre-processing, feature extraction, classification. The proposed system has been checked with a number of different Malayalam characters. There are so many identical characters present in Malayalam character sets. These are differentiated using their dissimilarity in some features. In Malayalam script the space between the sub-characters and the core character is equal and same as the space between the characters inside a word rendering the character segmentation process very intricate as conventional profiling methods fail.

This system is built with the intention of recognizing typed character images. The major use of this technique is that it will help in easily understanding the pronunciation of a word in other languages. For example the image of a bus board or placard in Kerala which is written in Malayalam can be converted into English or any of our understandable languages. It would be helpful for tourists in Kerala who are not familiar with the language.

II. LITERATURE REVIEW

Segmented character recognition using single value decomposition method is the main focus in Anil R et al. [1]. Euclidean distance measure is used for discovering the closest character class of the segmented character image during testing. It has reduced the comparison in the classification by training the image. The general terms which are used in this paper are optical character recognition (OCR), singular value decomposition and classification of the character nayana, a project of centre for development of advanced computing which is an optical character recognition system used for Malayalam characters. The nearest matching characters in Malayalam causes a mismatch. For improving this OCR, a specialized pattern classifier for close matching character has been developed. The contour method is used here for the segmentation. Here each line, character and word is segmented one by one and each segmented characters are tested with datasets. The accuracy gets reduced when the system takes the mean. Without considering mean, the accuracy is improving but that time the number of comparison is increasing. The misclassifications are mainly happening because of the similarity of the character shapes. Freeman code, back propagation neural network and Unicode are the main keywords which are used here in Amritha Sampath et al. [2]. Here in this system it is mainly dealing with the handwritten Malayalam characters. So here the handwritten characters are changing into machine understandable code. The handwritten characters are not the same and it may change according to the mood and urgency of the person. So compared to the printed characters, it is a difficult task. Malayalam language contains old script and new script. Therefore, the recognition task is difficult in Malayalam. The steps which are used here are pre-processing, feature extraction, recognition, post processing. In the feature extraction phase it is removing all the noise present in the character. And in the feature extraction step there are 2 types; extracting low level feature like length, width etc. and the high level features are number of curves, number of loops and its position. And a special feature used here in this system is the identification of the direction information based on the freeman code. Direction like a pen tip move is recorded in a single dimensional array 1 for NE, 2 for E, 3 for SE and it is stored thus. The training set consists of different feature sets which will be identified by the network. The neural network contains so many links and nodes which are arranged in layers. In this system it has used neural network for classification. After classification the next step is post processing where the Unicode of each character is identified and Disambiguation of characters can be done based on the positioning and meaning each character gives to the word, hence making the system more efficient. For example in the case of ‘va’ and ‘pa’ both are similar so it is distinguished by the count of pixels above and below the horizontal line. Manik Varma et al. [3] focuses upon the recognition of the characters which are taken from the images taken from a real scene by a standard camera. So the images of notices, bus board, placards also can be identified. The performances of various features are accessed by nearest neighbor and SVM classifier. Here street images from different parts of India are the data sets. In English both upper and lowercase letters are taken separately and there are around 62 classes in English. But in kannada there are no upper and lower case letters. But consonants and vowels are combined to give 600 classes in kannada. Obtaining natural images for training purposes can be expensive and time consuming. The outcomes got via training in hand-printed characters were not empowering. This could be because of the constrained changeability among the composition styles and additionally the moderately little size of the training set that the capturing of the image is enabled. Coates Reading text from photographs is a tough task. The methods which are used in Adam Coates et al. [4] are recently developed in machine learning. Therefore, learning features automatically and it also shows how it is helpful to develop classifiers to be recognized in a high accuracy rate. The main features which are used in this system are photo OCR, feature learning, robust reading, character recognition. The results point out that it may be possible to achieve high performance using more scalable and sophisticated feature learning algorithms currently being developed by machine learning researchers. Trupti R.Zalke et al. [5] mainly focusing on the work done in Indian languages like devangiri scripts. Devangiri is the basic script used all over India. Here first the input images are taken from the scanner. Each image is preprocessed and segmented. Then according to the feature extraction result the classifier classifies each of the character. Finally post processing is happening. This paper will surely helpful for other research communities in India in the fields like social sciences, economics and other recognition system. The Malayalam characters are rich in complexity of patterns. That makes it complex to recognize each one .In Bindhu philip et al. [6] they are doing it by using OCR engines. The stages presented in the development of OCR engine are image acquisition, segmentation, pre-processing, normalization, feature extraction and classification. Pre-processing means removing the back ground noises and gray scale images. And in the case of segmentation it is segmenting it into lines, words, characters and sub characters. It is taking the left and right consonant and conjunct and also taking the vowel sign too. In order to the left, right, left and right vowel signs here it getting the sound of the word. The classification of Malayalam characters is done by using SVM classifiers. In feature extraction it is identifying each and every feature. In feature extraction method it uses several lateral features like frequency capture, average gap analysis and absorption. Absorption captures the large number of vertical strokes of Malayalam character. Average gap analysis refers gaps between the numerous twists in the characters. And one more distinct feature of Malayalam character is the character with multiple loops that is recognized by the frequency capture process. Because of the rich pattern in Indian scripts it causes many problems owing to its complexity. Hisham P.M et al. [7] describing about the methodologies which are used for the character recognition. The methodologies are singular value decomposition (SVD) and Euclidean distance measurement. The contour based segmentation task is using for the segmentation operation. The classification is done with Euclidean distance measurement in a lower dimensional feature space. Paul ClarkIn et al. [8] present two diverse ways to deal with the location and recuperation of text in pictures of genuine scenes. The keywords which are utilized as a part of this paper are point of view recuperation of document, edge angle distribution, oriented text, text texture. The principal approach utilizes page edges and other rectangular limits around content to find a surface containing content, and to recuperate a fronto-parallel view. This is performed utilizing line discovery, perceptual grouping, and comparison of potential text regions using a confidence measure. The second approach utilizes low level surface measures with a neural system classifier to find locales of content in a picture. At that point this framework recovers a fronto-parallel perspective of each found section of content by isolating the individual lines of content and deciding the vanishing purposes of the content plane. The outcomes are outlined by number of pictures.

Some researches are mainly for helping the people who are visually impaired. Xiangrong Chen et al. [9] use a number of images of city as a dataset. From those images it manually extracts the text regions. In practice, we trained a cascade with 4 strong classifiers contains 79 features. An extensive algorithm and adaptive binarization are applying on the region which is selected by the cascade classifier. Here a commercial OCR is utilized to read the test and in addition dismiss in the event that it is a non-text area.

III. METHODOLOGIES USED

STEPS IN CHARACTER RECOGNITION

- Data Acquisition
- Pre Processing
- Feature Extraction
- Segmentation
- Classification
- Post Processing

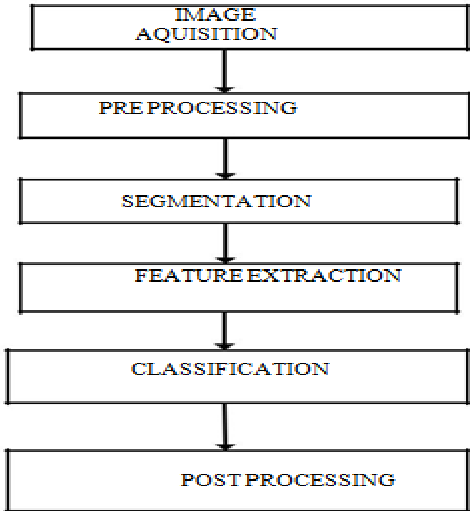


Fig 1: Character Recognition Steps

A. Data Acquisition:

This is the phase in which information are gathered as a part of the recognition process. The information might be caught online, in that case the printed Malayalam characters are taken from the web. In the disconnected case, the information is filtered by using a scanner..

B. Preprocessing:

RGB to gray:

It converts a color image to a grayscale image by wiping out the shade and immersion data while holding the luminance.

Gray thresh:

This method is used to convert grayscale image to a binary image with the use of im2bw function. The im2bw replaces all pixels in the input image with the luminance greater than level with the value one that is for white and all the other pixels with the value zero that is for black.

Complement the binary image:
When complimenting the binary image zeros will become one and once will become zeros. The black areas will become white and white areas will become black.

Size Normalization:

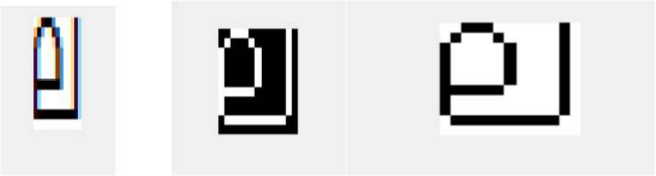
In size normalization the height and width of a specific image can be changed into a fixed height and width. Likewise, the image size can be changed into a size which is needed. It will helps to give same height and width to all images.

Morphological operations:

Morphological operations are applied on the binary images. There are many kind of operations that are present in bwmorph which are skel, shrink, thin, thicken, spur, remove etc. Here in this system we used skel and shrink operations for thinning the binary images.

Skel: In this operation the boundary pixels are removed from the object but the objects don‟t break apart. Here we can give the number of times the operation need to be applied. If „n‟ is infinity then that operation will apply infinite times until the character does not change.

Shrink: It shrinks object to points, the objects with holes shrinks to a connected ring and the objects without gaps shrink to a point. Here also, applying the number of times that operation needs to be applied.



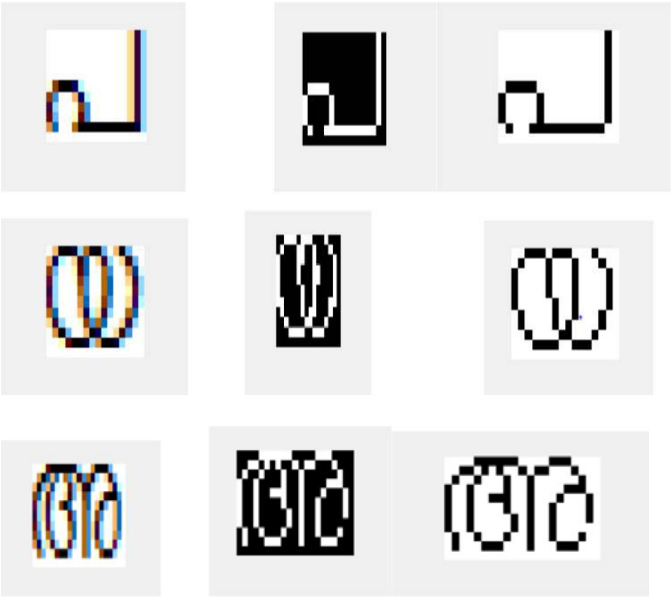


Fig 2: Preprocessing Steps

C. Segmentation:

Segmentation is an operation that isolates individual characters from the printed text. The segmentation in this system is done by using regionprops method. regionprops(BW,properties) returns estimations for the arrangement of properties determined by properties for each associated segment in the binary image. Here each letter is separated from a word with a bounding box which is used as a parameter inside the regionprops method. For cropping each character from the image, it has used the parameters like top, bottom left points and top, bottom right points. These four points will give a clear idea of from where to where one character should be cropped. These individual character images are saved into a folder by using the name a1, a2, a3…. The further steps like pre processing, feature extraction, classification are doing on these characters one by one.

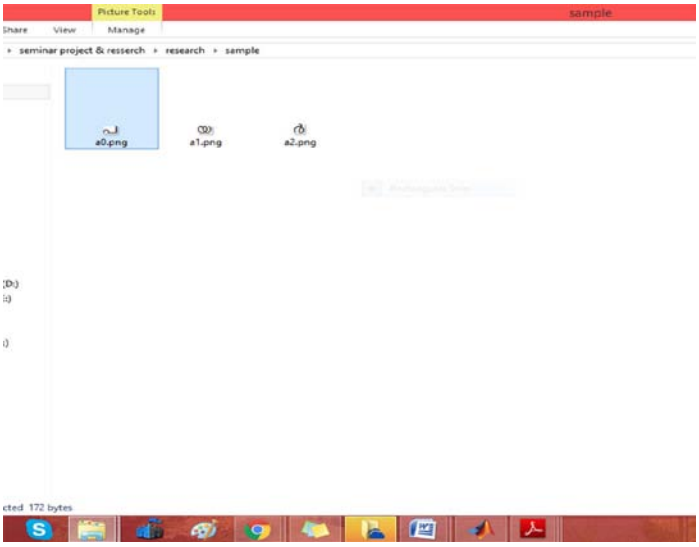


Fig 3: Segmentation Proces

D. Feature Extraction:

Feature extraction is the process of extracting the number of feature from each character. This system used the features like number of end points, number of loops, number of crosses, straight lines, height and width of each character. It also checks for the quadrants. It finds out the 4 quadrants of each letter, it also helps in the easy recognition of similar characters.

For example „vấ and „pa‟ are two Malayalam consonants which looks similar.

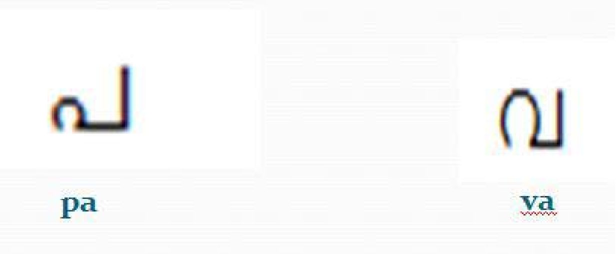


Fig 4:Similar Characters 1

These 2 letters could be differentiated with the help of quadrants. In „va‟ q1 has some value. But in the case of „pa‟ q1 value is zero.

In the case of „fa‟ and „tha‟ also both are differentiated using the number of tips. Here „fa‟ having 2 edges and „tha‟ having

3 edges.

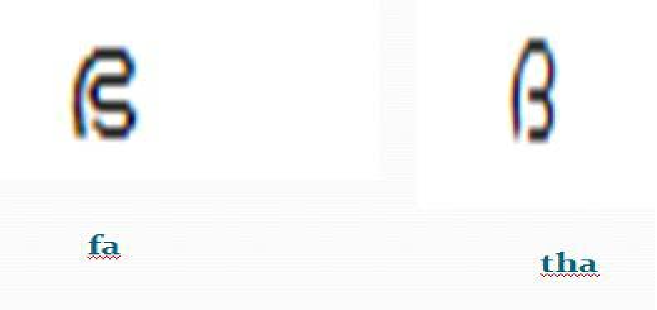


Fig 5:Similar Characters 2

According to the extracted features only the next classification step is happening.

E. Classification:

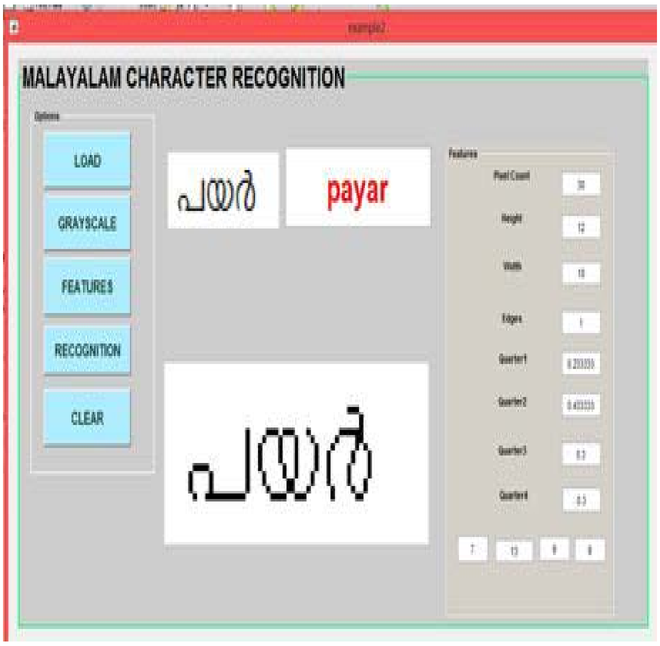
Classification is the next stage after feature extraction in character recognition. Here it compares the extracted output values with the already stored value. Based upon the extracted features it gives label for each character. Minimum distance classifier, SVM classifier, decision tree classification, neural networks are some of the classifiers commonly used for this purpose.

F. Post Processing:

Post processing will give a final output of the recognized characters. It will give the output as a string. Each letter is taken from the folder and recognizes it one by one. The recognized characters are concatenated into a string and finally that string will be displayed.

IV. RESULT

The result of this research is that it recognizes Malayalam words and displays the recognized ones in an understandable language. Here the recognition is performed by the process like pre-processing, feature-extraction, segmentation, classification, post-processing. There are a lot of researches which have already done in the field of character recognition in other languages but character recognition in Malayalam is very rare and very difficult also. Here it is facing some limitation in recognizing vowel signs. When a consonant and vowel sign join together it forms a new letter and it has a different pronunciation. The recognition of these characters are different from the other letters which is another limitation of this research.

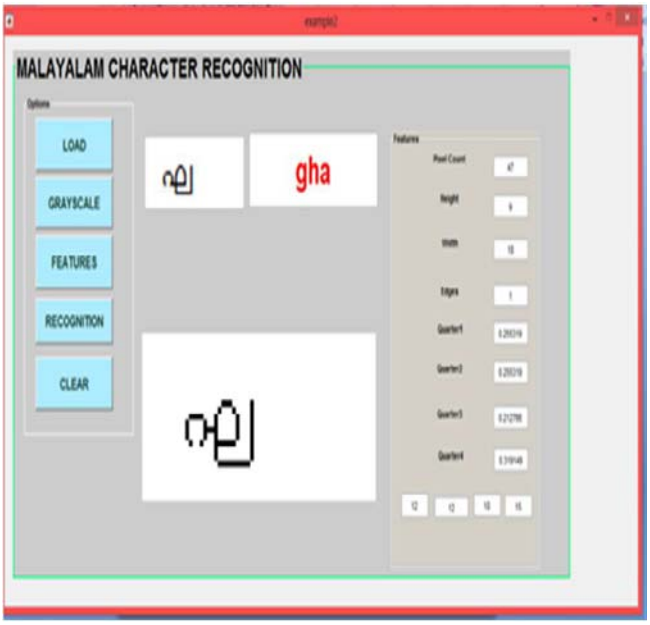


പയർ payar

Pixel Count 30
Height 12
Width 10
Edges 1
Quarter1 0.233333
Quarter2 0.433333
Quarter3 0.3
Quarter4 0.3

7 13 9 9

Fig 6: Word Recognition



ഘ gha

Pixel Count 47
Height 9
Width 10
Edges 1
Quarter1 0.255319
Quarter2 0.255319
Quarter3 0.212766
Quarter4 0.319149

12 12 10 15

Fig 7: Character Recognition

V. CONCLUSION

The character recognition in Malayalam language is difficult compared to other languages. Some of the Malayalam characters look similar. It also increases the difficulty in understanding. This system was developed using Mat Lab. The main steps that are using for this recognition process are image acquisition, segmentation, pre-processing, feature extraction, classification and finally post-processing. The identical characters are differentiated using the extracted features.

This work can be used to develop a system that can help tourists who do not know how to read and write in Malayalam to navigate around Kerala by reading bus boards and other signs and placards. Adding an audio component that can voice the pronunciation of the word can further enhance the system to make it helpful for older people with blurred vision. It actually displaying the pronunciation of a Malayalam word in English therefore it will promote the regional languages to the next generation.

VI. REFERENCES

[1] Anil R ,Arjun Pradeep and Midhun E M. Malayalam character recognition using single value decomposition. International Journal of Computer Applications).2014; (0975 – 8887).

[2] Amritha Sampath ,Tripti and Govindaru . Freeman code based online handwritten character recognition for Malayalam using backpropagation neural networks. An International Journal (ACIJ) .2012.

[3] Te´ofilo E. de Campos, Bodla Rakesh Babu and Manik Varma. Character recognition in natural images. International Journal of Innovative Research in Computer and Communication Engineering .2013.

[4] Adam Coates. Text detection and character recognition in scene image with unsupervised feature learning. International Journal of Innovative Research in Computer and Communication Engineering.2013.

[5] Trupti R.Zalke, Prof.V.N.Bhonge. An optical character recognition system for Indian scripts International Journal of Engineering and Technical Research (IJETR).2015.

[6] Bhindu Philip, R.D Sudhaker Samuel. An efficient OCR for printed Malayalam text using novel segmentation algorithm and SVM classifiers International Journal of Recent Trends in Engineering. May 2009.

[7] Hisham P.M.*, Aneesh C., Sachin Kumar S., and K.P. Soman. Optical Character Recognition for Printed Malayalam Documents Based on SVD and Euclidean Distance Measurement International Conference on Signal and Speech Processing (ICSSP-14) .2014.

[8] Paul Clark, Majid Mirmehdi. Recognizing text in real scene. International Journal on Document Analysis and Recognition (IJDAR) .2002.

[9] Xiangrong Chen, Alan L. Yuille. Detecting and Reading Text in Natural Scenes . IEEE.2004.

[10] Kauleshwar Prasad, Devvrat C. Nigam, Ashmika Lakhotiya and Dheeren Umre. Character Recognition Using Matlab’s Neural Network Toolbox. International Journal of u- and e- Service, Science and Technology, Vol. 6 No. 1. February 2013.

[11] Faisal Mohammad, Jyoti Anarse, Milan Shingote, Pratik Ghanwat. Optical character recognition implementation using pattern matching. (IJCSIT) International Journal of Computer Science and Information Technologies. Vol. 5.2014.

[12] M. Ushman Akram, Zabeel Bashir, Anam Tariq and Shoab A Khan. Geometric Feature Points Based Optical Character Recognition. IEEE .2013.