A Brief Analysis on Offline Character Recognition in Malayalam Scripts

Meenu Alex[1] **Smija Das**[2]
[1,2]Department of Computer Science
[1,2]SJCET, Palai, Kottayam, Kerala, India

Abstract— Character Recognition in regional language has become the hotspot of research all around the globe. It is a challenging area of pattern recognition and image processing fields. It has become a vitality in the widespread reach of computational tools. The most significant hazard in this area is the complexity in the strokes and structure of regional languages. Malayalam, being the second most challenging among the languages is under the spotlight for the same. In this paper we aim to study the various techniques employed in character recognition in Malayalam scripts.

Key words: Malayalam Character Recognition, Feature Extraction, Classifier, Neural Networks

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. 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.

അ ആ ഇ ഈ ഉ ഊ ഋ എ
ഏ ഐ ഒ ഓ ഔ അം അ:

Fig. 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

Fig. 2: Malayalam Consonant Set

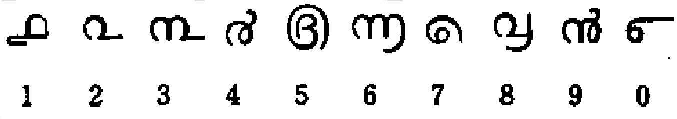

Fig. 3: Malayalam Numerals

III. ARCHITECTURE OF A GENERAL CHARACTER RECOGNITION SYSTEM

As already mentioned, the purpose of a character recognition system is to convert the human readable characters to machine editable form. Initially the character image is scanned, then a sequence of steps take place and finally converted into character codes like ASCII. Through the last step, the system can edit and manipulate the character.

The basic steps in character recognition are listed below:

1) Preprocessing
2) Segmentation
3) Feature Extraction
4) Classification
5) Post processing

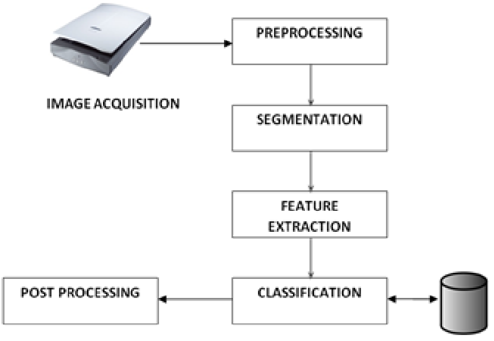

Fig. 4: Character Recognition System

1) *Image Acquisition:*

This is the first step in the character recognition process. The image is acquired by by scanner or digitizer. This is a vital step as any error in acquisition process presents error to the whole effort.

2) *Preprocessing:*

Preprocessing is done on the image to remove as much distortion as possible. Noise may occur due to poor quality of scanned or degraded documents. There are different steps of preprocessing.

A. *Binarization:*

It converts gray scale binary document image to binary images. It is done to extract the object of interest from the image. Here the separation of forward pixels from backward ones take place. Binarization process is done by applying global or local thresholding.

B. *Skeltonization:*

It is also called as thinning. The main purpose of skeltonization is to extract only the essential information from the image. The thickness of the image is reduced in this process.

C. *Noise Removal:*

The image may contain noise, even if we take extra care in the preprocessing stage. It is dependant on a certain number of factors such as the quality of scanner, age of documents, color of ink, type of pen etc. The noise occurred can be salt and pepper noise, Gaussian noise etc. So it is required to filter out these noises by a good filtering technique to get better recognition results.

D. *Skew Correction:*

While an image is fed to a scanner, there is a chance of inclination to occur in the fed document. This inclination is called skew and the angle made by the lines of text with the horizontal direction is called skew angle. Profile Projection is the most commonly employed method in skew removal.

3) *Segmentation:*

Segmentation is a method to isolate individual characters from handwritten text.

1) Line segmentation
2) Character segmentation
3) Word segmentation

Separating lines from image is called line segmentation. This is done after the skew correction. Horizontal projection profile method is the common technique used for line segmentation. Word segmentation follows successful line segmentation. Vertical projection profile can be employed to the segmented lines. In both techniques, the peaks and valleys are identified and corresponding techniques are applied to find the boundaries. It is very hard to separate some characters (eg. letter 'ki'). So a technique called labeling[2] is used in such cases to separate combinatorial characters.

4) *Feature Extraction:*

Feature means the salient characteristics of the image. Here the important features of the image are extracted to form feature vectors. The various characteristics of image are extracted and stored as feature vectors. These feature vectors are used by classifiers for the character recognition process.

5) *Classification:*

This is the final stage in the character recognition process. Based on the extracted features, unique labels are assigned to the character images. Template matching is extensively used for the classification purpose. Some commonly used classifiers are neural networks, Support Vector machine, Nearest neighbor classifier, Binary tree classifier etc.

6) *Post Processing:*

In the post processing stage, the characters are mapped to their corresponding Unicode values. Also, some disambiguating technique is used in this stage to deal with confusing characters (letters 'pa' & 'va').

IV. STUDIES ON HANDWRITTEN MALAYALAM CHARACTER RECOGNITION

From the past few years, a lot of works have been reported in the field of Malayalam Character Recognition. The first work in Malayalam character recognition was reported by Lajish V.L.[3]. He proposed a novel feature extraction method. It uses fuzzy-zoning method and normalized vector distance measures. The size normalized image is divided into 3x3 uniform sized zones. Perpetual zoning is a method used for local information analysis on partitions of a given pattern. The information is obtained from the most meaningful part of the pattern. The most difficult thing is to identify the meaningful part of the pattern. The Malayalam characters have diverse structural characteristics. Due to the variable length of character character curve, normalized vector distance for each zone is calculated. The obtained features were classified using a class modular neural network. The system showed an accuracy of 78.87% for 44 Malayalam handwritten characters.

He also worked on a method for modeling Malayalam handwritten characters[4]. In State Space Point Distribution (SSPD), parameters obtained from State Space Map (SSM) is used for classification. This method does not require the usual step of binarization. This is a new feature extraction method based on state space representation by trying all possible location in gray scale image. The system showed an accuracy of 73.07%.

Renju John et al [5] came forward with the concept of 1D Wavelet Transform of Projection Profiles for Isolated Handwritten Malayalam Character Recognition. After obtaining the scanned image, preprocessing steps like binarization and thinning are applied on the character image. The charaters are modeled using projection profiles. By applying 1D wavelet transform, feature vector is extracted from the projection profile. A Multi Level Perceptron network is used as the classifier. This system showed an accuracy of 73.8% for 33 characters.

G. Raju[6] had done a remarkable work in handwritten character recognition by applying Daubechie wavelet coefficients. He used a member of Daubechie wavelet family, db4 for decomposition into ten sub images. Preprocessing methods like denoising and thinning are not used in this method. The given image is made into inverted binary images, followed by the application of wavelet transform. A multilayer perceptron network was used as the classifier. It showed an accuracy of 76.8% for 33 characters. They extended their work by using an additional feature-aspect ratio[7], which improve the recognition accuracy to 81.3%.

M. Abdul Rahiman et al [8] proposed a method for recognition of Malayalam characters based on HLH intensity patterns. They focuses on the special inherent patterns in each character and therefore preprocessing is not required. Based on the HLH intensity patterns, characters are grouped into different classes. Before feeding the characters for recognition, all these HLH patterns are separated from the image. This method exhibited an accuracy of 88.6%. They also put forward a vertical and horizontal line positional analyzer algorithm[9] for the recognition of handwritten Malayalam characters. Decision tree was used as the classifier. One major benefit of this method is its capability to identify coloured characters and characters written on a coloured background. This method achieved an accuracy of 91%.

Binu P. Chacko and Babu Anto[10] introduced a method for recognizing Malayalam Characters using discrete features. Skeltonization is applied in the image to extract only the relevant features. The presence of parasitic components will degrade the performance of a character recognition system. So a pruning method was used to remove the spurious branches present in the image. The skeleton pruning was done by contour portioning with discrete curve evolution (DCT). This method showed an accuracy of 90.18% for 33 character classes. Later they proposed a method based on Wavelet Energy Feature (WEF) and Extreme Learning Machine (ELM)[11]. Single Hidden Layer Feedforward Network was used as the classifier. It showed an accuracy of 95.16%.

Jomy John et al [12] proposed a chain code histogram based method for recognizing Malayalam vowels. In this work, Chain Code and Normalized chain code were used for classification. The use of Image Centroid used for extracting features improve the recognition result. For classification, a two layer feed forward network with sigmoid activation function is used. This system showed an accuracy of 72.1% . They also proposed a work for recognizing Malayalam characters using wavelet transform and SVM classifier[13]. Here a Haar wavelet transform is used for feature extraction and SVM is used as the classifier. The recognizer provides an accuracy of 90.25% for 44 characters.

Bindu S. Moni and G. Raju proposed a handwriting recognition system based on run length count (RLC). When the process of scan is carried out from top to bottom/left to right, a continuouous group of 1's may be obtained. This is termed as RLC. Characters are blocked using Meshing technique. RLC of different block form the feature vector. Modified Quadratic Discriminant function (MQDF) is used as the classifier. This feature extraction method provides an accuracy of 94.18% for 30 selected characters. They have also proposed a work for recognizing handwritten Malayalam characters. Gradient feature is extracted from the character images. Here also, Modified Quadratic Discriminant function was used as the classifier. This approach attained an accuracy of 95.42% for 44 Malayalam characters.

Vidya V. et al[13] proposed a method for handwritten character recognition based on Probabilistic Simplified Fuzzy ARTMAP (PSFAM). It is a combination of Simplified Fuzzy ARTMAP and Probabilistic Neural Network. Features like Zernike moment features , distance feature, cross feature and fuzzy depth are extracted from the character glyph. The classifier used is PSFAM. This approach obtained an accuracy of 87.81% for 142 Malayalam characters.

Recently, Anitha Mary M.O Chacko and Dhanya P.M.[14] introduced a method for Malayalam character by introducing the concept of multiple classifiers. Instead of using a single classifier for classification, this method make use of multiple classifiers. This overcomes the limitations of individual classifiers. Two feedforward neural networks are used as the classifiers. In feature extraction stage two features, chain code histogram and fourier descriptors are extracted. This system proved an accuracy of 92.84% for 33 character classes.

V. CONCLUSION

In this paper various mechanisms for character recognition in Malayalam scripts is presented. An overview of a general character recognition system and the stages involved in it is also discussed. Malayalam character recognition is a very tedious process. This is mainly due to the presence of large character set, similarity between the characters etc. The development of a complete character recognition system in Malayalam is still lacking in this scenerio. Character recognition is matured for foreign languages, but complicated for Indic scripts especially South Indian languages. The Malayalam characters are of different dimension. So it is difficult to map them to a square grid like English languages. Another difficulty is that there is no standard database in Malayalam character that we can use for validating the test results. We hope that our study will help researchers working in this area.

WORK	FEATURES EXTRACTED	CLASSIFIER USED	ACCURACY (%)
Lajish V.L.	Fuzzy Zoning, NVD	CMNN	78.87
Lajish V.L.	SSPD	CMNN	73.03
John et al	Wavelet	MLP	73.8
G. Raju	Wavelet, Aspect Ratio	MLP	81.3
M. Abdul Rahiman et al	HLH Intensity features	---	88.6
M. Abdul Rahiman et	Vertical & Horizontal line count and position	Decision Tree	91

al			
Binu P. Chacko et al	WEF, ELM	SLFN	95.16
Jomy John et al	CCH, NCCH, Image Centroid	Feedforward NN	72.1
Jomy John et al	Haar Wavelet Transform	SVM	90.25
Bindhu S. Moni et al	RLC	MQDF	94.18
Bindhu S. Moni et al	Gradient	MQDF	95.42
Vidhya V. et al	Fuzzy depth, Zernike & distance moment, cross feature	PSFAM	87.81
Anitha Mary et al	Chain code Histogram, Fourier Descriptors	Feedforward NN	92.84

Table 1: Comparison of Different Malayalam Character Recognition Systems

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] B. Anuradha and B. Koteswarra; “An efficient Binarization technique for old documents”, Proc.of International c onference on Systemics,Cybernetics and Inforrmatics,Hyderabad, pp771-775,2006

[3] 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

[4] 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.

[5] 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 13-15.

[6] 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.

[7] G. Raju, “Wavelet transform and projection profiles in handwritten character recognition- A performance analysis”, Proc. Of 16th International Conference on Advanced Computing and Communications, Chennai 2008, pp 309-314.

[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] 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.

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

[11] Binu P. Chacko, V. R. Vimal Krishnan, G. Raju, P Babu Anto, “Handwritten character recognition using wavelet energy and extreme learning machine”, International Journal of Machine Learning and Cybernetics June 2012, pp 149-161.

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

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

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

[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] 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.

[17] Anitha Mary M.O. Chacko, Dhanya P.M, “Combining Classifiers for Offline Malayalam Character Recognition”, Emerging ICT for Bridging the Future-Vol 2, Advances in Intelligent Systems and Computing 338, Springer International Publishing Switzerland 2015.