Malaya Journal of Matematik, Vol. 9, No. 1, 616-620, 2021

https://doi.org/10.26637/MJM0901/0106

A review on various techniques used to recognize off-line handwritten Malayalam characters

M.P. Ayyoob[1*] and P. Muhamed Ilyas[2]

Abstract

The process of recognizing the script written by different people with different handwriting in different languages with the help of machines is called Off-line handwritten character recognition (OHCR). It is difficult for machines to recognize the script in same language written in different localities and circumstances. It is an important factor that how far machine can achieve the accuracy in recognizing these handwritten records. In this article, it is examined the off-line handwritten records in Malayalam language.

Malayalam is the Dravidian language which is used in the Kerala state of India, Lakshadweep and Mahi. This paper presents a comprehensive survey of various techniques with its accuracy rate on Malayalam off-line handwriting recognition for the past few years. Different feature extraction and classification techniques used in Malayalam off-line handwriting recognition are pointed out. Then databases chosen for these purposes are also discussed.

Keywords

Off-line handwritten character recognition (OHCR), Malayalam off-line handwriting recognition, feature extraction, classification.

[1,2] *Department of Computer Science, Sullamussalam Science College, Areacode-673639, Kerala, India.*

***Corresponding author**: [1] mpayyoobmp@gmail.com

Article History: Received **11** December **2020**; Accepted **10** February **2021**

Contents

1. Introduction

to the introduction of writing, there was only verbal communication. Cultural norms, rituals and stories passed on orally from generation to generation. Gradual evolution of language and cultures forced the need for communication. Handwritten Character Recognition (HCR) is a vital area in the field of pattern recognition. It explores the horizon of many application areas such as speech recognition, image recognition, traffic management, translation of voice into text and vice versa, navigator etc. HCR can be on-line or off-line. In on-line handwriting recognition data is recorded at the time of writing itself. In off-line handwriting recognition prewritten text are scanned. Due to the differences in human handwriting style the automation of handwritten document recognition is a formidable task and it also depends on the situvations of a person. Malayalam is one of the most difficult languages in the world for automatic recognition due to its largest number of letters, variety of letters and similarity between the characters. So the exorbitant accuracy in the field of automatic off-line Malayalam word recognition is a onerous task. Only a few studies are made on the recognition of handwritten Malayalam characters and that are limited to just basic characters too[1].

The main stages in handwritten character recognition system are:

- Data Collection
- Preprocessing
- Feature Extraction
- Segmentation
- Classification
- Post Processing

In data collection, the data is obtained by scanning the image of prewritten text. This is the initial stage in which data are collected for starting the recognition process. After this stage, preprocessing stage is starting. It is an important step in character recognition process. It achieved through binarization, noise removal, skeletonization or thinning, skew correction and normalization. The third stage segmentation is the process of partitioning characters from the handwritten text. It obtained through line segmentation, word segmentation and character segmentation. After segmentation feature extraction and classification processes are performed. The final stage post-processing to be taken after classification process is completed.

This article is arranged as follows, related works explain in Section II , Section III discuss about Malayalam character,Section IV presents various classifiers and feature sets used by authors for the various stages of off-line Malayalam handwriting recognition , Section V is concludes the present article.

2. Related works

There are several works reported in literature for the Malayalam character recognition. Author[2] considered 30 samples of 25 consonants out of 36 Malayalam consonants for data set. An application of Wavelets is used here. Here a member of Daubechie Wavelet family with order 4, db4 used for decomposition. The author claims that, it would have produced better result if the large database (a minimum of 3000 samples per character) and Neural network / Support Vector Machine for classification were used. He also claims that the needs of preprocessing in his work. Authors[3] use isolated characters only. A Multi Level Perceptron (MLP) network and 1D Wavelet transform are used for classification and feature extraction. For training purposes 4950 samples of 33 classes were used. 1650 samples used in testing process. Result showed a recognition accuracy of 73.8%. In [4] author used fuzzy-zoned normalized vector distance for feature extraction and class modular neural network for classification over set of 15,752 samples of the 44 basic Malayalam handwritten characters from 358 writers with 78.87% accuracy.

Author[5] proposes an algorithm in the noiseless environment. In the noiseless environment 4 sets of samples scanned with 661 letters and using HLH intensity Patterns for feature extraction, dynamic matrix for classification and produces an accuracy of 88%.

An algorithm proposed in[6] recognizes the characters with perspective accuracy by making use of the ingrained characteristic features. It is done by utilizing the intensity variations in the original text. It produces the output of editable version of the recognized Malayalam characters for 24-bit bmp enscribed images. The work achieved 92% accuracy for 3 sets of samples ranging 402 letters in noiseless environment. In another attempt [7] produces accuracy of 94 % on a total of 2490 handwritten characters. In [8], for training phase 19,800 handwritten Malayalam character samples was used.The Modified Quadratic Discriminant function (MQDF) taken as classifier. It causes to trim the cost of computation and reduces the number of input variables in dataset to a larger extent when compared to QDF and MQDF. MQDF obtained 95.42% accuracy. Classification performance improves more than 10% also.

Another scheme was proposed by authors[9]. It consists of two stages. The first one is a feature extraction stage and the second one is a classification stage. Haar wavelet transform is used for feature extraction and Support Vector Machine (SVM) is used as classifier. It is an effective method. Using third level decomposition, classification result is 89.64% and 90.25% on second level. They use 228 different writer's 10,000 handwritten isolated Malayalam characters. The feature extraction of the characters is done in[10] by evaluating the position and count of the horizontal and vertical lines. Based on the count and position of the horizontal and vertical lines a classification of the simple and conjunct is devised. An accuracy of 93.83% is achieved through this method. A two layer feed forward neural network as with sigmoid activation function [11] used for feature extraction, chain code for classification. It obtained 72.1% accuracy over 60 handwritten pages of different writers. By using Probabilistic Simplified Fuzzy ARTMAP (PSFAM) authors [12] propose off-line Malayalam handwritten character recognition. They collected isolated samples from 26 informants. In addition to that numbers from 0 to 9, punctuations and 69 compound characters. The result shows 87.81% accuracy.

[13] uses a two stage approach for recognition. In testing stage 36,000 handwritten samples belonging to 90 different character classes were used. In the first stage group consists of similar characters and those that mis-classify among themselves in single stage approach. Second stage contains a character assigned to a specific group is classified to a particular class of that group. Comparing to single stage scheme, the proposed scheme is more error free. The proposed approach is competent under various conditions. 96.21% accuracy was obtained the group classification and 95.01% accuracy was obtained the overall character level. In [14] gradient based featues with three simple features as run length count, aspect ratio and centroid of the image character code are used. The combination of Simplified Quadratic Classifier (SQDF) and Multi Layer Perceptron (MLP) classification is used. It shows a remarkable results of 99.78% accuracy. SURF feature and Curvature feature [15] were used for feature extraction, SVM and Neural Network as classifiers. Using these two dissimilar classifiers the experiment is conducted in 2 phases. 33 isolated Malayalam characters shows 89.2% accuracy in first phase and the second phase shows the accuracy of 81.1% on Malayalam sentences. Authors [16] get more than 90% accuracy over 18000 isolated Malayalam handwritten characters using stacked LSTM. 90 Malayalam character symbols are considered for the recognition. Network consists of three layers , in which two LSTM layers and final output layer for prediction.

Table 1. Comparative Analysis of Off-line Malayalam Handwriting Recognition.

Author(s)	Public -ation Year	Features	Classifiers	Data Set	Accuracy
G. Raju	2006	Count of zero crossing	Feed forward Neural Network	30 samples of each consonants (out of 36)	Not mentioned
Renju John, G.Raju, D. S. Guru	2007	1D Wavelet transform	Multi Layer Perceptron (MLP)	4950 samples of 33 classes.	73.8%
Lajish.V.L	2008	fuzzy-zoned normalized vector distance	Class modular neural network.	Set of 15,752 samples of the 44 basic Malayalam handwritten characters from 358 writers.	78.87%
M. Abdul Rahiman, Aswathy Shajan, Amala Elizabeth, MK Divya, G Manoj Kumar, MS Rajasree	2010	HLH intensity Patterns	Dynamic matrix	4 sets of samples ranging 661 letters.	88%
Abdul Rahiman. M, Aswathy Shajan, Amala Rajasree M S	2010	HLH intensity patterns	Dynamic matrix	629 handwritten characters	92%
Abdul Rahiman M , Rajasree M S	2011	HLH intensity Patterns	Dynamic matrix	total of 2490 handwritten characters.	94%
Bindu S Moni, G Raju	2011	Gradient direction feature.	Modified Quadratic Discriminant function (MQDF).	19,800 handwritten Malayalam character samples	95.42%
Jomy John , Pramod K. V, Kannan Balakrishnan	2011	Haar wavelet transform	Support Vector Machine (SVM).	10,000 handwritten isolated Malayalam characters written by 228 different writers.	90.25%
M. Abdul Rahiman M.S. Rajasree	2011	Counting the number of vertical lines in the characters, calculate the number position of horizontal lines in a character	Dynamic matrix.	handwritten characters of 9 different categories of persons.	93.83%
Jomy John, Pramod K. V Kannan Balakrishnan	2011	chain code	A two layer feed forward neural network as with sigmoid activation function.	60 handwritten pages are collected from different persons	72.1%
V. Vidya, T.R. Indhu, V.K. Bhadran, R. Ravindra Kumar	2013	Fuzzy, geometrical, structural and reconstructive features	Probabilistic simplified fuzzy ARTMAP	isolated samples from 26 informants. In addition to that numbers from 0 to 9, punctuations and 69 compound characters	87.81%

Jomy John, Pramod K. V. Kannan Balakrishnan, Bidyut B. Chaudhuri	2014	Gradient Features	Multi-Layer Perception (MLP), K-Nearest Neighbor, Support Vector Machine (SVM), Extreme Learning Machine (ELM).	6,000 handwritten samples belonging to 90 different character classes.	95.01%.
G Raju, Bindu S Moni, Madhu S Nair	2014	gradient-based features and run length count (GBF–RLC)	Multi Layer Perceptron (MLP) and Simplified Quadratic Classifier (SQDF)	44 classes of 19,800 isolated handwritten characters	99.78%
Meenu Alex, Smija Das	2016	SURF feature and Curvature feature	SVM and Neural Network	Phase I: 33 Malayalam isolated character classes. Phase II: Malayalam Sentences.	Phase I: 89.2% Phase II: 81.1%
Jino P.J, Jomy John, Kannan Balakrishnan	2017	It will extract features or it can directly learn from the raw data.	Stacked Long Short Term Memory (LSTM)	Isolated Malayalam handwritten character (Ninty symbols) total samples are 18000.	more than 90 %
P.J. Jino, Kannan Balakrishnan, Ujjwal Bhattacharya	2018	Convolutional Neural Network (CNN)	Support Vector Machine (SVM)	10,676 handwritten samples corresponding to 314 classes	96.90%

In [17] offline handwritten Malayalam word recognition(6360 samples for training and 2120 samples for testing) achived 96.90% accuracy by using deep hybrid neural network architecture with CNN and SVM. CNN alone provided 95.74% accuracy. A better result provided when the same architecture used on Hindi and Marathi word databases. 94.15% accuracy obtained on Hindi databases. Two databases are used for Marathi in which one contains 10260 and another contains 7980 samples. It provides 92.60% and 92.19% accuracy respectively. CNN and SVM are used for feature extraction and classification. Table 1 gives a comparative analysis of the off-line Malayalam handwriting recognition.

3. Malayalam script

Malayalam is the language which is spoken in the Kerala state of India, Lakshadweep and Mahi which is a part of Pondicherry (or Puducherry). It is included in the Dravidian language family. Malayalam is one among the twenty-two official languages of India which are included in the 8th schedule of Indian constitution.

The oldest records concerned with the evolution of Malayalam as an independent language are seen in the order of 9th century. Malayalam, the one among the Dravidian language family, has clear relation with the other Indian languages such as Sanskrit and Tamil.

There are different language variants in modern Malayalam. It can be said that Malayalam is a entirety of different language variations based on locality, caste, labour and manners. Newspapers, radio, education through lakhs of printed text books help to spread Malayalam as a communicative language. Except the geographical, social and cultural factors, caste and religion also cause language variations in Kerala. As a result , the automatic recognition of them becomes a more difficult task.

4. Feature extraction and classification techniques

A) feature extraction approaches

Feature extraction deals with extracting interested attributes for differentiating one class of objects from another. For classification and for building models, this information is passed to the recognizer. Various types of features are proposed here like 1D Wavelet transform, HLH intensity Patterns, Gradient direction feature, Haar wavelet transform, Chain code and Convolutional Neural Network (CNN). The characteristics of fuzzy, geometrical, structural and reconstructive features, surf and curvature or it will extract features or it can directly learn from the raw data, counting the number of vertical lines in the characters, calculate the number and position of horizontal lines in a character are also causes for feature selection.

B) classification approaches

A number of classifiers have been utilized for recognition of off-line Malayalam handwritten characters and words. These include Feed forward Neural Network, Multi Layer Perceptron (MLP), Class modular neural network, Dynamic matrix, Modified Quadratic Discriminant function (MQDF), Support Vector Machine (SVM) and Stacked Long Short Term Memory (LSTM).

Developing highly reliable feature extraction and classification techniques in off-line Malayalam handwritten character recognition is a challenging task and much work is needed.

5. Conclusion

In this article, survey of published research works in the field of Malayalam handwriting recognition is presented. Using the general framework for handwritten text recognition, we have compared the feature extraction and classification phases of Malayalam text recognition system. We hope that this survey encourages the off-line handwritten recognition research of Malayalam. Most of the recent work on off-line Malayalam text recognition has focused on isolated characters. There are very few attempts on Malayalam page of text (or lines of text) recognition. Therefore, more research effort is needed for unconstrained Malayalam handwritten text recognition.

References

[1] Umapada Pal, Ramachandran Jayadevan, and Nabin Sharma. Handwriting recognition in indian regional scripts: a survey of offline techniques. *ACM Transactions on Asian Language Information Processing* (TALIP), 11(1):1, 2012.

[2] G Raju Recognition of unconstrained handwritten Malayalam characters using zero crossings of wavelet coefficients, *Proc. Of International Conference on Advanced Computing and Communications*, ADCOM, 217-221, 2006.

[3] 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, 13-15K.

[4] Lajish V L, Handwritten Character Recognition using perpetual Fuzzy zoning and Class modular Neural Networks, *Proc. of fourth International Conf on Innovations in IT*, 2007.

[5] M Abdul Rahiman et. al., Isolated handwritten Malayalam character recognition using HLH intensity patterns, *2010 Second International Conference on Machine Learning and Computing.*

[6] M Abdul Rahiman et. al., An HCR System for Combinational Malayalam Handwritten Characters based on HLH Patterns, *International Journal of Computer Applications* (0975 – 8887) 8(11)(2010).

[7] M Abdul Rahiman and Rajasree M S, An Efficient Character Recognition System for Handwritten Malayalam Characters Based on Intensity Variations, *International Journal of Computer Theory and Engineering*, 3(3)(2011).

[8] 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.

[9] 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.

[10] A. Abraham et al. *Recognition of Simple and Conjunct Handwritten Malayalam Characters Using LCPA Algorithm ACC* 2011, Part III, CCIS 192, 304–314, 2011. Springer-Verlag Berlin Heidelberg 2011.

[11] Jomy John, Pramod K. V, Kannan Balakrishnan *Offline Handwritten Malayalam Character Recognition Based on Chain Code Histogram*, Proceedings Of ICETECT 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,* 182(2013), 273-283.

[13] Jomy John, KV Pramod, Kannan Balakrishnan, and Bidyut B Chaudhuri. *A two stage approach for handwritten malayalam character recognition. In Frontiers in Handwriting Recognition* (ICFHR), 2014 14th International Conference on, 199–204. IEEE, 2014.

[14] G Raju, Bindu S Moni, and Madhu S Nair. A novel handwritten character recognition system using gradient based features and run length count. *Sadhana*, 39(6)(2014), 1333–1355.

[15] Meenu Alex and Smija Das. *An Approach towards Malayalam Handwriting Recognition Using Dissimilar Classifiers.*

[16] Jino P J., Jomy John., Kannan Balakrishnan.: Offline Handwritten Malayalam character Recognition using stacked LSTM. In: 2017 *International Conference on Intelligent Computing,Instrumentation and Control Technologies* (ICICICT).

[17] P. J. Jino, Kannan Balakrishnan and Ujjwal Bhattacharya.: *Offline Handwritten Malayalam Word Recognition Using a Deep Architecture.* In: SocPros2017, 1.10.1007/978-981-13-1592-3_73.

⋆⋆⋆⋆⋆⋆⋆⋆⋆
ISSN(P):2319 – 3786
Malaya Journal of Matematik
ISSN(O):2321 – 5666
⋆⋆⋆⋆⋆⋆⋆⋆⋆