International Journal of Engineering & Technology, 5 (x) (2017) xxx-xxx

SPC

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: xxx
Research paper, Short communication, Review, Technical paper

Hybrid Approach of Handwritten Malayalam Character Recognition

Ajay James[1*], Sonu Varghese K[2] and Chandran Saravanan[3]

[1]*Assistant Professor*
[2]*M.tech student*
[3]*Associate Professor*
[*]*ajayjames80@gmail.com,sonuvk13@gmail.com, dr.cs.1973@gmail.com*

Abstract

Handwritten character recognition of South Indian scripts especially Malayalam is an on-going area of research. Limited works are proposed in this field due to the significant character set with highly complex and similar characters. Here hybrid technique of feature extraction based on geometrical and structural properties of characters is proposed. This method consists of two stages, in the first stage characters are classified into a group based on geometrical features such as the number of ending, bifurcation and loop. And in the second stage characters are recognized based on specific characteristics defined for each group. The proposed method exhibits recognition rate of 96.5% and accuracy of 93.86% on an average.

Keywords: *Chain code, Feature extraction, Handwritten Character Recognition, Malayalam, Water reservoir technique.*

1. Introduction

Handwritten Character Recognition (HCR) is the process of improving and integrating the human-computer interaction by converting an image of handwritten or typewritten documents to automated machine recognizing object. Most of the HCR systems works based on the four general approaches of pattern recognition as template matching, statistical techniques, structural techniques and neural networks [17]. Different stages of HCR system are image acquisition, pre-processing, feature extraction, recognition, and post-processing. Out of these stages, feature extraction and recognition are most important stages which determine the accuracy, recognition rate, and speed. Here fundamental component of characters (features) are extracted, and characters are recognized based on these extracted features. Malayalam official language of Kerala consists of largest character set among all Indian languages. Character set consist of 13 vowels, 36 consonants, five chillu or pure consonants, four consonant signs, 12 vowel signs, anuswaram, compound characters, etc. Vowels and consonants are basic characters. Dependent vowel sign is a diacritic attached to the left, right or both sides of consonant when the consonant is followed by a vowel. The consonant signs are glyph pieces which do not exist on their own and appear either to the left, right or up of consonant. Pure consonants are characters derived from basic consonant units. Compound characters are a special type of characters formed by the vertical or horizontal combination of two or more consonants. The challenges in Malayalam character recognition are due to isolated characters without upper and lower case differences, alpha-syllabic nature of script, use of conjuncts and combinational letters, existence of new script and old script, characters distinct with small variation in appearance, characters with high shape richness, unwanted slants, skew, and curves occurred from change in writing style, absence of standard dataset, etc. The objective of this paper is to introduce a new method of Malayalam HCR, which recognize most of the characters in Malayalam script. Section 2 discusses the related works carried out so far in this domain. The proposed method is detailed in Section 3. Section 4 shows results of experimentation and Section 5 concludes the paper.

2. Related Work

This section discusses the various feature extraction and classification methods developed in the domain of handwritten character recognition. Features are compact and characteristic representation of a character. Feature extraction methods are developed based on statistical distribution of pixels, structural and geometrical properties of the character and global transformation moments [18]. Different methods of feature extraction are discussed below. Zoning or region decomposition is a mechanism for local information analysis on partitions of a given pattern and resembles that of human perception. Zoning is mainly applied to compute the percentage of black pixels in each zone. [1] Suggests the fuzzy zoning method used in Malayalam HCR. Here Image is divided into nine equal zones and each border except outer boundaries to 3 fuzzy zones. Membership value is assigned to each pixel based on position in each fuzzy zone. Then vector distance from the origin to each pixel is calculated and normalized to form the feature vector. A projection profile is a histogram giving the number of black pixels accumulated along parallel lines. Vertical and horizontal projection profiles are well known for their discriminating power. One major limitation of projection profile is the size of the feature vector is relatively large. [2] proposes

projection profile created by the wavelet transform of an image. Here the image is first re-sized into 32X64. Then count some pixel along each row to form the horizontal projection and apply a 1D transform to approximate to 8 feature values. Similarly, eight feature value along vertical projection is taken as the feature vector. Every character may be identified by its geometric specification such as corners, endings, centroid, centroid angle, phase angle, loops, bifurcations, etc. [3] proposes OCR with the feature vector for each character consists of a number of corners, endings, and bifurcations. Here a crossing number '1, 2, 3' is assigned for each pixel correspond to character ending, corner and bifurcation respectively. It then uses transitions between end points and line tracking and scanning system to resolve ambiguity between the characters having same feature values. [14] proposes feature selection based on corners, loops, edges, and boundaries. Corners are detected by plotting pixel along the Cartesian plane (x, y) such that x remains constant while y keep varying or y remains constant while x keep varying. Edges are point where there is rushed variation in pixel intensity, and boundary extraction technique uses mathematical morphology method. The water reservoir principle uses the features of reservoirs in character. That means if water is poured from a side of a component, the cavity regions of the component where water will be stored are considered as reservoirs. The [4][5] analyze top, bottom, left and the right reservoir formed by the character or component. In [5] water reservoir area, water flow level, reservoir base-line, the height of a reservoir, base-area points, etc. can be further consider as the features for classification. In [4] directional density estimation, maximum profile distances and fill hole density are also taken as features. Chain code captures the directional information which identifies exterior information of any shape or pattern. Chain code encodes successive points in continuous curves that are adjacent to each other. The chain code is defined as four directional or eight directional. In [6] chain code along the edge in the clockwise direction is used for finding the CCH(chain code histogram) and NCCH(normalized chain code histogram). Centroid of image, CCH, and NCCH are used as a feature vector for recognition. Run Length Count (RLC) is the count of a contiguous group of on pixel. Horizontal RLC is a count of on pixels in the top to the bottom scan of a character image, and vertical RLC is on pixels encountered in a left to right scan of an image. In [7] fixed meshes are constructed for binary images by dividing the image into equal parts. The transition from one to zero is considered as edges. A number of transition along a row is considered as horizontal RLC and number of transition in the column is considered as vertical RLC. HLH pattern is an alternative technique to projection profile for identifying foreground and background colors. Here H-L notation is used for representing feature set where H represents valid character path and L the background. In [8][9] uses HLH pattern as a feature vector. Based on HLH pattern in the horizontal, vertical and diagonal direction they are classifying the entire character set into 'ra,' 'pa' and special type characters. Length and breadth of each character can be calculated by manipulating the HLH intensity values of the segregated image. Then characters are reconstructed and identified based on boundary point. [10] discuss about structural features such as angle between the base line and global center of gravity, Distance between local center of gravities in each cell and global center, aspect ratio, number of black pixels in each of the region above and below the diagonals, number of black pixels in the region to the right and left of the vertical line, number of black pixels in the region to the top and bottom of the horizontal line. 17 features are extracted for each character. [11] discuss structural properties such as circularity, regularity, component-based features, etc. Global transformation and series expansion techniques are Fourier transform, Gabor transforms, Fourier descriptor, wavelets, moments, etc. The crossing is used to detect the number of strokes presented in character along a particular path. If this path is along the rows, then it is called horizontal crossings, and if this path is along the columns, then it is called vertical crossings. [12] defines technique of applying wavelet transform over the image and counting the zero crossing in each of 10 sub-bands. Here zero crossing is defined as the occurrence of a positive value after a string of negative and zero values or a negative value after a string of positive and zero values. Here 25 Malayalam character are classified into 11 different groups. In [13][15] discuss two-stage feature extraction. Here the first stage is a grouping, where groups consist of similar characters are formed. In the second stage, a character assigned to a group in the first stage is classified to a particular character class. In [13] the first stage uses gradient features by applying Sobel operator over the normalized image. Then characters are classified into 19 groups. Now features such as loop or near-loop at the beginning, full loop to the left side of the second and fourth character, a loop in the right upper corner, count of end-points, etc., are extracted for character identification.

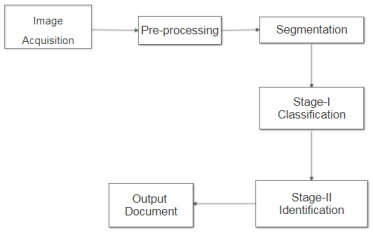

Figure 1: Architecture of the proposed system.

3. Proposed Work

The proposed work exhibit a hybrid technique of Malayalam handwritten recognition. The Figure1 depicts the architecture of given system which consists of different stages of image acquisition, preprocessing, segmentation, Stage-I classification, and stage-II identification. The sections below discuss the system in detail.

3.1. Image Acquisition and Pre-processing

The document is scanned at 600 dpi using a scanner to form the input image. Then in pre-processing step, we remove the noise from the input image by a morphological operation called opening. The image is converted into a binary image by Otsus thresholding method and a morphological operation called thinning is carried to form Skelton of character.

3.2. Segmentation

Segmentation is the process of partitioning an image into homogeneous regions. The different process involved in segmentation is line segmentation, word segmentation, and character segmentation. Different lines in the document are separated by using water flow technique of line segmentation. It is an efficient algorithm for line segmentation since it deals with touching, overlapping and skewed lines in handwritten documents. The segmented lines are passed to the word segmentation phase where words are segmented using spiral run length smearing algorithm. These segmented words are then given to character segmentation phase. Here each character is segmented using bounding box technique of connected component analysis. Each character extracted from each word is assigned in a structure and passed over to stage-I classification.

3.3. Stage-I Classification

In this stage every character is classified into ten different groups based on its geometric specification. Endings (points where the character ends and no more connected pixels can be identified), bifurcations (points where a single line gives rise to two other lines or

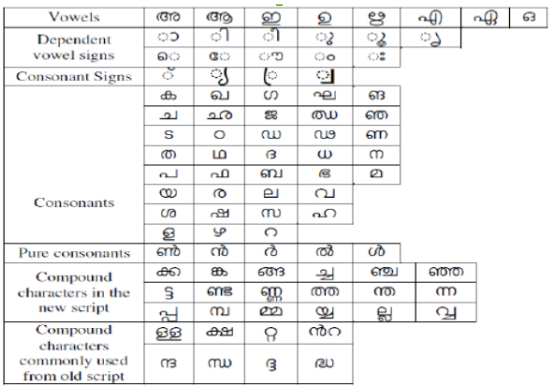

Figure 2: Characters classified into groups

meeting point of two lines) and loops (the circular enclosing which resembles circle without starting and ending point) are different features considered for classification. First, we initialize count of ending, bifurcation and loop as zero. Then modify count of t loop by re-sizing the image to the size of [100 X 100] and applying inbuilt Matlab function bwboundaries. Now we scan each pixel of a thin image and surrounding 8 pixels to identify whether given pixel is an ending or bifurcation. If a pixel is an ending, then there is only one pixel in corresponding 8 pixels surrounding given pixel is 'on'. Different pattern formation of a pixel and surrounding pixel are verified to obtain the count of bifurcation. After analyzing the count of different features for each character, the similarly shaped characters with the same number of features as well as the maximally mis-recognized characters are merged into one group. Different groups formed from the similar characters are shown in Figure2.

3.4. Stage-II Identification

Different feature extraction techniques are adopted by a different group for recognizing characters are given below.

1 Group-I The different features considered are cavity region formed using water reservoir principle as in [4], the position of endpoints concerning the centroid and pixel density.
2 Group-II The end points in each zone when an image is divided into nine fixed zones. Maximum count of the black pixel in horizontal direction and width of the character is the features for recognizing characters.
3 Group-III Characters in this group are identified based on a comparison of position and size of given character with its neighbors.
4 Group-IV Different characters in this group are identified based on the number of cross-points (conversion from black pixel to white pixel) in both vertical direction and horizontal direction. The characters which resemble the same number of cross points are further identified based on the position of their endpoints.
5 Group-V Number of vertical cross points and horizontal cross points along with the region of endpoint when an image is divided into nine equal regions are considered as the feature vector for character identification
6 Group-VI Divide image into four equal sub-image, analyze the structure of sub-image and encode it as chain code. These encoded values along with the vertical and horizontal cross points are considered as the feature vector for this group.
7 Group-VII Character identification process of this group is similar to group-V. Here vertical and horizontal cross points along with the position of endpoints are considered as features.
8 Group-VIII This group uses features similar to group-VI. Here we also consider the position of endpoints and RLC count in both horizontal and vertical direction as additional features for classification.

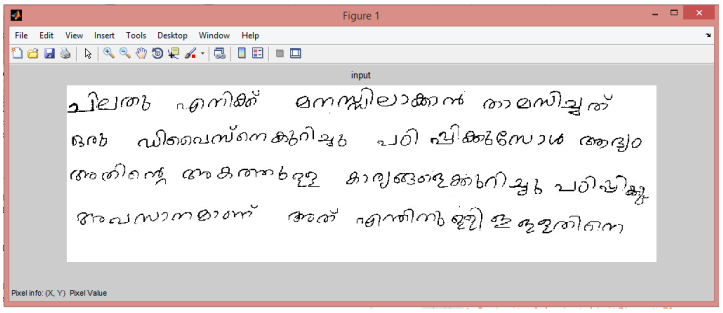

Figure 3: Input Image

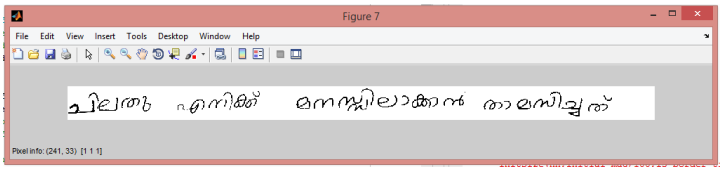

Figure 4: Segmenting individual line.

9 Group-IX This group consists of characters with more number of loops and one or two endings. Thus the position of endpoints along with the vertical and horizontal cross points are good features to classify different characters belong to given group.
10 Group-X HLH intensity pattern of character in the horizontal and vertical direction based on height and width of the character, number of zero crossing and position of endpoints are features considered for recognition. To handle misclassification of characters after the stage-I, the character is checked with a closely related group. For example after stage-I, if the loop is not completed the group IV elements will be recognized as group I . So to avoid misclassification in such cases initially we check for the group I then if the character is not recognized group IV features are considered. When a character is identified then that character is written into a text file by using the Unicode value of character. Thus the test file contains the recognized character. Figure3 shows the input image, Figure4 line segmented output, Figure5 word segmented output, Figure 6 bounding box for character separation and Figure7 the output text obtained for the input image.

4. Experimental Results

This section discusses on the implementation details, the metrics used in the evaluation of the system and results obtained from the ten documents.

4.1. System Implementation Details

The system is implemented using MATLAB 2014b, which is the image processing tool, on windows platform. The hardware used is Intel i7 2.2GHz processor and 4785T with 8 GB RAM.

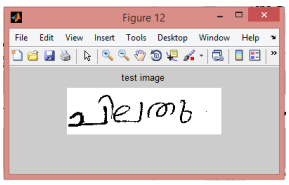

Figure 5: Word Segmentation

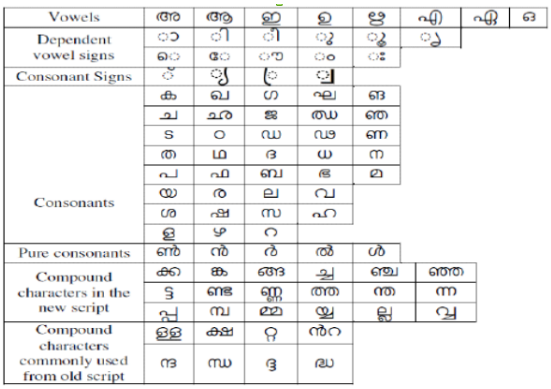

Figure 6: Bounding box over characters.

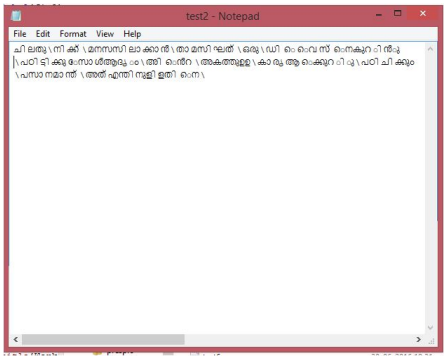

Figure 7: Output text

Table 1: Results obtained from testing on 10 document

Document	Total Characters	Recognized character	Correctly classified (Stage-I)	Correctly Identified (Stage-II)
Doc1	61	61	61	60
Doc2	121	113	114	103
Doc3	68	66	65	60
Doc4	68	67	66	64
Doc5	68	66	67	64
Doc6	85	82	80	75
Doc7	103	100	98	94
Doc8	91	85	84	80
Doc9	78	74	73	68
Doc10	73	71	71	67

4.2. Evaluation Criteria

The performance metric used for evaluation are recognition rate and accuracy. Recognition rate is a measure of total number of characters that are recognized to the total number of characters to be recognized.
Accuracy is the rate of the number of correctly recognized character to the total number of characters. Here we are calculating the accuracy of stage-I classification and stage-II identification. Here we also consider the accuracy of each group separately.

4.3. Results

The results obtained by testing the proposed method on ten documents are shown in Table 1. As shown in Table 2, the proposed work has an average recognition rate of 96.50% and accuracy of stage-I classification as 95.77% and stage-II identification as 93.86%.

Table 2: Recognition rate and accuracy of proposed system

Document	Recognition rate	Stage-I Accuracy (%)	Accuracy Stage-II (%)
Doc1	100	100	98.3
Doc2	93.38	94.21	91.15
Doc3	97.05	95.58	90.90
Doc4	98.52	97.07	95.52
Doc5	97.05	98.52	96.96
Doc6	96.47	94.11	91.46
Doc7	97.08	95.14	94
Doc8	93.40	92.30	94.11
Doc9	94.87	93.58	91.89
Doc10	97.26	97.26	94.36
Average	96.50	95.77	93.86

Table 3: Accuracy of each group

Groups	Accuracy (%)
Doc1	96.96
Doc2	95.52
Doc3	98.3
Doc4	94.36
Doc5	93.86
Doc6	91.89
Doc7	91.46
Doc8	92
Doc9	90.90
Doc10	91.15

Accuracy of each group is shown in Table 3.

5. Conclusion

Most of the works in Malayalam HCR is based on isolated characters. The proposed work deals with the recognition of characters from a document by passing through the line segmentation, word segmentation, and character segmentation phases. The results show that the three-stage recognition scheme based on geometrical and structural properties of character is an efficient method of Malayalam HCR with the recognition rate of 96.50% and accuracy of 95.77% in stage-I and 93.86% in Stage-II.

References

[1] Lajish, V. L. "Handwritten character recognition using perceptual fuzzy-zoning and class modular neural networks." *2007 Innovations in Information Technologies (IIT)*. 2007.
[2] Raju, G. "Wavelet transform and projection profiles in handwritten character recognition-A performance analysis." *Advanced Computing and Communications, 2008. ADCOM 2008. 16th International Conference on. IEEE*, 2008.
[3] Akram, M. Usman, et al. "Geometric feature points based optical character recognition." *Industrial Electronics and Applications (ISIEA), 2013 IEEE Symposium on. IEEE*, 2013.
[4] Dhandra, B. V., R. G. Benne, and Mallikarjun Hangarge. "Handwritten Kannada Numeral recognition based on structural features." *Conference on Computational Intelligence and Multimedia Applications, 2007. International Conference* on. Vol. 2. IEEE, 2007
[5] Pal, Umapada, and Partha Pratim Roy. "Multioriented and curved text lines extraction from Indian documents." *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 34.4 (2004): 1676-1684..
[6] John, Jomy, K. V. Pramod, and Kannan Balakrishnan. "Offline handwritten Malayalam Character Recognition based on chain code histogram." Emerging Trends in Electrical and Computer Technology (ICETECT), 2011 International Conference on. IEEE, 2011
[7] Moni, Bindu S., and G. Raju. "Modified quadratic classifier for handwritten Malayalam character recognition using run length count."

Emerging Trends in Electrical and Computer Technology (ICETECT), 2011 International Conference on. IEEE, 2011.

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

[9] Rahiman, M. Abdul, and M. S. Rajasree. "An HCR System for Combinational Malayalam Handwritten Characters based on HLH Patterns." *International Journal of Computer Applications,(IJCA), USA* 8.11 (2010): 19-23.

[10] Gayathri, P., and Sonal Ayyappan. "Off-line handwritten character recognition using Hidden Markov Model." *Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on. IEEE*, 2014

[11] Obaidullah, Sk Md, Anamika Mondal, and Kaushik Roy. "Structural feature based approach for script identification from printed Indian document." *Signal Processing and Integrated Networks (SPIN), 2014 International Conference on. IEEE*, 2014.

[12] Raju, G. "Recognition of unconstrained handwritten Malayalam characters using zero-crossing of wavelet coefficients." *2006 International Conference on Advanced Computing and Communications. IEEE*, 2006.

[13] Gatos, Basilis, Georgios Louloudis, and Nikolaos Stamatopoulos. "Segmentation of historical handwritten documents into text zones and text lines." *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on. IEEE*, 2014

[14] Fernández-Mota, David, Josep Lladós, and Alicia Fornés. "A graph-based approach for segmenting touching lines in historical handwritten documents." *International Journal on Document Analysis and Recognition (IJDAR)* 17.3 (2014): 293-312.

[15] Bhattacharya, Ujjwal, S. K. Ghosh, and S. Parui. "A two stage recognition scheme for handwritten Tamil characters." *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*. Vol. 1. IEEE, 2007

[16] Gaikwad, Ms VA, and Dr DS Bormane. "An Overview of Character Recognition Focused On Offline Handwriting." *International Journal Of Computer Science And Applications* 1.3 (2008): 0974-1003

[17] Tokas, Rajbala, and Aruna Bhadu. "A comparative analysis of feature extraction techniques for handwritten character recognition." *International Journal of Advanced Technology and Engineering Research* 2.4 (2012): 215-219

[18] Chacko, Anitha Mary MO, and P. M. Dhanya. "A Comparative Study of Different Feature Extraction Techniques for Offline Malayalam Character Recognition." *Computational Intelligence in Data Mining*-Volume 2. Springer India, 2015. 9-18.

[19] T. R. Vijaya Lakshmi, P. Narahari Sastry, T. V. Rajinikanth, "Hybrid Approach for Telugu Handwritten Character Recognition Using k-NN and SVM Classifiers", *International Review on Computers and Software (IRECOS)*, Vol 10, No 9 (2015)

[20] Jaspreet Kaur, B. S. Dhaliwal, S. S. Gill,"Offline Handwritten Gurmukhi Character Recognition using Particle Swarm Optimized Neural Network", *International Journal of Computer Applications (0975 – 8887) International Conference on Advances in Emerging Technology (ICAET* 2016)

[21] T. R. Vijaya LakshmiEmail authorPanyam Narahari SastryT. V. Rajinikanth,"Feature Optimization to Recognize Telugu Handwritten Characters by Implementing DE and PSO Techniques",*Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications* pp 397-405, 03 March 2017)

[22] Youssef Boulid, Abdelghani Souhar, Mohamed Youssfi Elkettani,"Handwritten Character Recognition Based on the Specificity and the Singularity of the Arabic Language",*International Journal of Interactive Multimedia and Artificial Intelligence* · June 2017

[23] Abhisek SethyEmail authorPrashanta Kumar PatraDeepak Ranjan Nayak,"Off-line Odia Handwritten Character Recognition: A Hybrid Approach",*Computational Signal Processing and Analysis*, pp 247-257, 03 April 2018