Segmentation of Malayalam Handwritten Characters into Pattern Primitives and Recognition using SVM

Baiju.K.B, Sabna.T.S, Lajish.V.L

Abstract*: This paper describes a lexical analysis (segmentation) approach in Pattern Recognition for Online Handwritten Character Recognition (OHCR) in Malayalam. The subunits (Pattern Primitives) in the single stroke vowel characters in Malayalam are identified and marked with pattern primitives to obtain a reference set of characters. Segmentation of the handwritten character samples into pattern primitives is made using a Combined Approach of Ramer Douglas Peucker algorithm and Eight Direction Freeman Code as per reference set. Features that are unique in the primitives of a character are extracted. The discriminating features identified are the direction of first primitive, segment count, cusp in second primitive, crossing in third primitive, and cusp in seventh primitive. The experiments were conducted on 100 samples per character that showed exact segmentation as per the reference set. With a five dimension feature set, the study achieved a recognition rate of 95.77% for five-fold cross-validation using Support Vector Machine with RBF kernel. The study shows that the segmentation of characters into pattern primitives is an effective method to realize accurate Malayalam OHCR systems for real-time applications.

***Keywords*: OHCR, SPR, Pattern Primitives, RDP, EDFC**

I. INTRODUCTION

In the field of human-computer interaction, machine understanding of natural handwriting, termed handwriting recognition, had an in-depth study for centuries. Advances in technology blended more promising results for handwriting recognition methods especially in Online Handwritten Character Recognition (OHCR). OHCR is nowadays a popular technique and used successfully in real-time applications extensively. The studies in OHCR, focused mainly on Latin and CJK scripts, and structurally, the majority of the characters in these scripts are a combination of linear segments [1]. Most of the studies in literature used more number of features and various classifiers. In the Indian context, the geometrical structure of a character is effectively used as a feature in many OHCR studies. The geometrical features of Indic scripts are described in various studies [2]. It is observable that the Indic script contains as many similar shapes when it is divided into segments based on some features like direction change, dominant points, etc. Such generic parts are called primitives, which could be used to distinguish between different classes of characters[3].

An in-depth study on Syntactic Pattern Recognition (SPR) is described by K S Fu and P H Swain, which reports the concept of describing patterns in terms of primitive elements, sub-elements, and their relationships [4]. In the Indian context, a primitive based approach for handwritten Bengali alpha-numeric characters is effectively used in a study by Abhijith Dutta and Santanu Chaudhury [5]. The studies that address recognition of online handwritten character using pattern primitives are not attempted in Malayalam. This paper describes a recognition scheme for online handwritten Malayalam characters based on primitive segments.

Malayalam is the official language of Kerala, a southern state of India. The language is also used in the Indian union territories of Lakshadweep and Pondicherry. The language has 13 vowels (including two diphthongs), 36 consonants, and five chillus. It also consists of vowel modifiers and special symbols including anusvaram, visargam, and chandrakkala.

The detailed description of the studies is given in the following sections that are organized as follows. In section II, the methodology of the study is described. The device and the method of data acquisition is described in section III. Various pre-processing techniques used in the study are discussed in section IV. Descriptions of different pattern primitives of single stroke vowels and the Combined approach for segmentation is mentioned in section V. The distinguishable features of the character samples and extraction of these features are described in section VI. The classification experiments using SVM is detailed in section VII. Results are analyzed in section VIII, followed by the conclusion and future directions described in section IX.

II. METHODOLOGY

The study focuses on the structural aspects of Malayalam characters. Rather than considering the features of the entire character, the study considers subunits (pattern primitives) of the character and the features of the pattern primitives are identified. A character reference set that is manually marked with segmentation points corresponding to pattern primitives is created in the study. The online handwritten characters are segmented into pattern primitives as per the reference set. A simple approach is proposed to extract the features of every character by segmenting them into pattern primitives.

Revised Manuscript Received on February 15, 2020.
***** Correspondence Author
Baiju.K.B*, Department of Computer Science, University of Calicut, Kerala, India. Email:baijukb_dcs@uoc.ac.in
Sabna.T.S, Department of Computer Science, University of Calicut, Kerala, India. Email:sabnats@gmail.com
Lajish.V.L, Department of Computer Science, University of Calicut, Kerala, India. Email:lajishvl@uoc.ac.in

The features of pattern primitives that differentiate eight single stroke vowel characters are analyzed. The features are then extracted from the pattern primitives. The proposed classification technique is Support Vector Machine(SVM). Five-fold cross-validation is also proposed in the study to verify the effectiveness of the method.

The handwritten data is acquired from the writers using the device e-Writemate. The acquired data is pre-processed through normalization, smoothing, and re-sampling for removing jitters, noises from writing speed variation, dots *etc*. The handwritten characters are segmented into pattern primitives. Features are extracted from the pattern primitives and the classification experiments are performed on these features. The block diagram of the proposed study is shown in Fig.1.

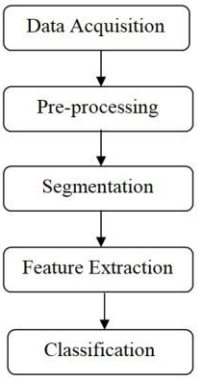


Fig.1. Block diagram of the proposed study

III. DATA ACQUISITION

Data is a highly relevant part of any system which is oriented on training models. In the study of handwriting, the potential of addressing variations in writing styles is an essential factor in the performance of the algorithm. The inclusion of more data samples ensures an accurate fitting model after training. The accuracy of a system for handwriting is measured upon the variability and the features of different calligraphic styles. In the experiments, Hi-Tech e-Writemate (Fig.1) is used for data acquisition, which is compatible enough to capture handwriting input traditionally.

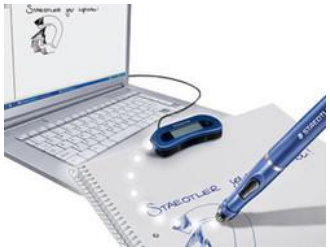


Fig.1. e-Writemate

The device includes a digital pen and a sensor. Handwriting on the paper written using the digital pen will be stored in the sensor device as (x, y) coordinates of the neighboring points. It can store up to a hundred A4 sheets.

After writing, the stroke series are available as a text file for each character. The paper used for recording the handwriting is printed with Malayalam characters in an arranged form. Nineteen samples of eight characters is acquired from a single data sheet(Fig. 2). Ten writers contribute 190 samples per character to form 1520 samples in the dataset. All the writers fall in the age group of 20-40 who are graduates

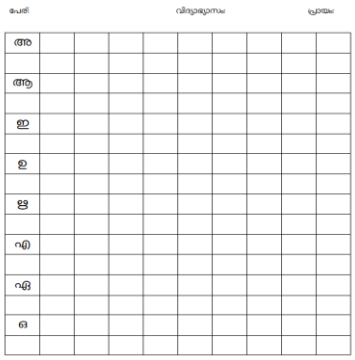


Fig. 2. A sample data capturing sheet

IV. PRE-PROCESSING

Pre-processing is an essential component in most of the data processing systems with natural interfaces. The online handwritten samples acquired through the input devices always contain noises like dots, jitters, over writings, slant, *etc*. The collected handwritten data samples are pre-processed to reduce these imperfections using normalization, smoothing, and re-sampling.

For normalization, the method proposed in the study is min-max normalization. The stroke values of a handwritten character sample are reduced to a scale between 0 and 1 using (1). Here, each of the (x, y) values representing a point in a stroke were divided by the difference of maximum and minimum values of x and y to obtain the new values.

$$X = \frac{x - \min(x)}{\max(x) - \min(x)} \quad Y = \frac{y - \min(y)}{\max(y) - \min(y)} \qquad (1)$$

For smoothing the character samples, a moving average filter is used. A moving average filter where N+1 is the filter length, and the sliding window M is averaged over the entire points *n* is shown in (2)

$$M = \frac{x[n]+x[n-1]+x[n-2]+\cdots\ldots x[n-N]}{N+1} \qquad (2)$$

The primary factor deciding the smoothness of shape in the moving average filter technique is filter length, and it is dependent on the data. As the filter length increases, the smoothness of the shape also increases. A Malayalam character sample അ <a> with filter lengths 5 and 15 is shown in Fig.3(a)(b). The same sample shows a higher shape variation for the two values. In the study, the filter length is fixed as five by visual inspection on various samples for various filter lengths.

Fig. 3(a). Handwritten sample of അ <a> for filter length 5

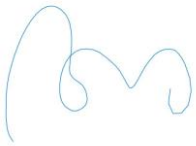
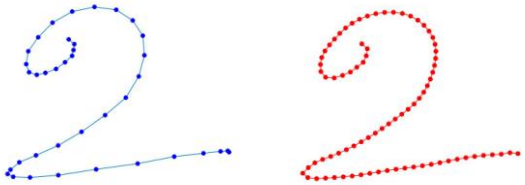


Fig. 3(b). Handwritten sample of ആ <a> for filter length 15

Re-sampling is carried out using the equidistant re-sampling method. The missing points in the character samples are interpolated using parametric spline approximations. Fig.4(a)(b) shows the character ഉ <u> and its re-sampled for 80 points.



(a) Original (b) Re-sampled

Fig. 4(a)(b). character ഉ <u> and the re-sampled form

V. SEGMENTATION OF HANDWRITTEN CHARACTERS INTO PATTERN PRIMITIVES

The lexical analysis phase of Syntactic Pattern Recognition (SPR) describes the segmentation of patterns into subunits that are also suitable for Malayalam handwritten characters. The syntactic and lexical approach to handwriting recognition has been studied in numerous works using segmentation techniques [6]. There are as many visually similar patterns in Malayalam characters. These similar patterns are known as pattern primitives. The pattern primitive of the Malayalam vowel characters in the study is shown in Fig. 5.

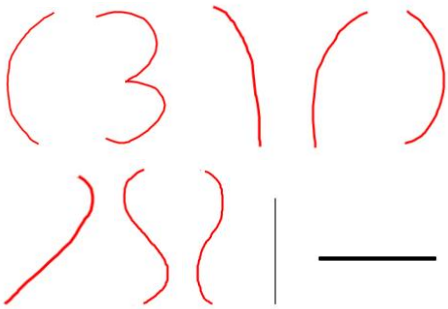


Fig. 5. Pattern Primitives in Malayalam Vowels

These character samples are segmented into the pattern primitives using the segmentation algorithm. Here, a combined approach of the Ramer Douglas Peucker(RDP) algorithm and Eight Direction Freeman Code(EDFC) is used for segmentation purposes[7] [8]. Both of the algorithms are extensively used in the literature of handwriting recognition studies. The algorithm segments the handwritten character samples as pattern primitives based on the manually marked reference set. The character samples marked with segmentation points (reference set) are shown in Fig.6. The maximum segmentation points are seven, *i.e.*, eight segments for □ <a> and minimum segments are obtained for □ <o>.

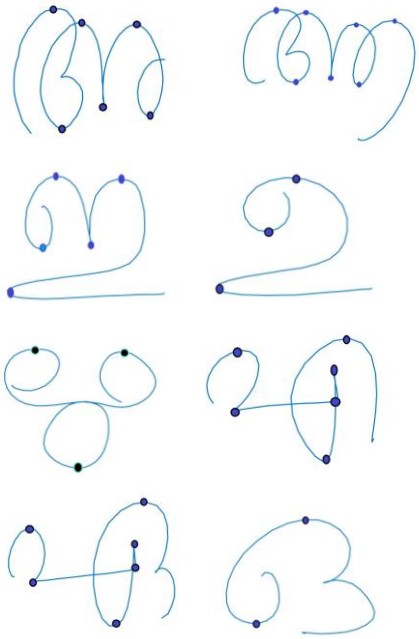


Fig.6. Malayalam vowel characters with marked segmentation points

The Combined approach of segmentation is described in this section. This approach segments the handwritten samples as per the reference set. The segmentation points are marked using the Ramer Douglas Peucker algorithm and Eight Direction Freeman Code. The points obtained from both of the algorithms were combined, and fine-tunings are made to these points for unreferred directions and redundant points as per the reference set. The entire procedure is termed as the Combined approach that is described in algorithm 1.

Algorithm 1

……

Input: PointList
Output: FinalList
CombinedList1=**DouglasPeucker** (PointList)
CominedList2=**Freeman_Code**(PointList)
CombinedList=CombinedList1+CombinedList2
For all points in the list **do**
Make the point list unique
Remove two nearby points closer than 5 points
Delete points with unreferred direction
End
FinalList // The segmentation points as per the reference set

……

VI. FEATURE EXTRACTION FROM PATTERN PRIMITIVES

After the segmentation process, the entire samples are segmented into pattern primitives. The features specific to pattern primitives need only to be considered to distinguish various character classes, Visual inspection of various characters identified that the following features are enough to distinguish the characters.

A description of the features is listed in table I.

Table-1: Features of Pattern Primitives

Features	Description
Direction of first primitive	The direction of first primitive is upwards for the characters അ <a>, ആ<a:>, ഋ<r>, ഏ<e>, ഐ<e:> and downwards for ഒ<o> ഇ<i>, ഉ<u>
Cusp in the second and seventh segment	A sharp cusp in some primitives is unique. In second segment, a cusp is present for അ <a>, ആ<a:>.Similarly a cusp in the seventh segment of ഐ<e:>
Crossing in primitives	Some primitives intersect other primitives in the same character. In the third segment of അ <a>, ആ<a:> and ഋ<r> a crossing is observed.
Segment Count	The number of segments in most of the characters are different

A. The direction of Pattern Primitives

The direction is an important feature of online handwriting. The characters in the study are either starting in an upward direction or downward direction. The direction code of the pattern primitive is obtained through the Eight Direction Freeman Code, and a reduced direction code is obtained from the Eight Direction Freeman Code. In reduced representation, an Eight Direction Freeman Code of '3333331111' is represented as '31'.Here, the reduced directions are obtained for the first primitive and compared against all the combinations of directions in the upward direction ('1','2','3') and downward direction ('5','6','7'). If the count for the directions is more towards upward directions, the pattern primitive is treated as upward else a downward pattern primitive. The direction code of a pattern primitive is shown in Fig.7.

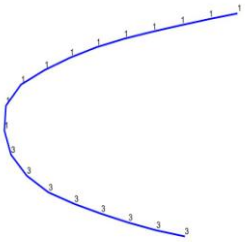


Fig. 7. Direction Code of a Primitive Segment

B. Cusp in Pattern Primitives

A cusp is a point on a curve where a moving point on the curve must start to move backward. Sharp cusps are observable in characters like അ<a>, ആ<a:>, ഐ<e:> and ഒ<o> (Fig.8).

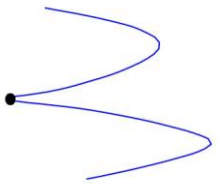


Fig. 8. Cusp in a Primitive Segment

A sharp turn in a curve is identifiable by measuring the curvature between three consecutive points. Let (x_1,y_1), (x_2,y_2), and (x_3,y_3) be three such adjacent points. The curvature of a circle drawn through them is four times the area of the triangle formed by the points divided by the product of its three sides, as shown in (3). The cusps observed in second and seventh pattern primitives are considered as features in the experiment.

$$K=\frac{2*((x2-x1)*(y3-y1)-(x3-x1)*(y2-y1))}{\sqrt{((sqr(x2-x1)+sqr(y2-y1))*(sqr(x3-x1)+sqr(y3-y1))*(sqr(x3-x2)+sqr(y3-y2))}} \quad (3)$$

C. *Crossings by Pattern Primitives*

A common point shared by two segments is known as the point of intersection, and the following method is used in the studies to determine those points. Given two line segments, *S1* and *S2*, with ending points *(x1(1),y1(1))* and *(x1(2),y1(2))* for *S1* and *(x2(1),y2(1))* and *(x2(2),y2(2))* for *S2*. Four equations with four unknowns are to be solved, namely *t1*, *t2*, *x0* and *y0*, where *(x0,y0)* is the intersection of S*1* and *S2*, *t1* is the distance from the starting point of S*1* to the intersection relative to the length of *S1* and *t2* is the distance from the starting point of *S2* to the intersection relative to the length of S*2*. If $0 \le t1 < 1$ and $0 \le t2 < 1$, then S1 and S2 segments cross, and the point *(x0, y0)* is marked as the crossing point. The matrix form of the four equations is displayed below. The crossings observed in the third pattern primitive is treated as a feature in the experiment.

$$\begin{bmatrix} x1(2)-x1(1) & 0 & -1 & 0; \\ 0 & x2(2)-x2(1) & -1 & 0 \\ y1(2)-y1(1) & 0 & 0 & -1 \\ 0 & y2(2)-y2(1) & 0 & -1 \end{bmatrix} \begin{bmatrix} t1 \\ t2 \\ x0 \\ yo \end{bmatrix} = \begin{bmatrix} -x1(1) \\ -x2(1) \\ -y1(1) \\ -y2(1) \end{bmatrix}$$

D. Segment Count of the Character

The number of segments, which is also the number of pattern primitives, is an important feature to differentiate various characters and considered as a feature here. The segmentation algorithm in the experiment segments the characters into N number of pattern primitives, as listed in table II.

Table-II: Segment Count of various characters

Character	Segment Count (N)
അ <a>	7
ആ <a:>	8
ഇ <i>	6
ഉ <u>	4
ഋ <r>	4
ഏ <e>	7
ഐ <e:>	7
ഒ <o>	3

The above-listed features are extracted from the pattern primitives obtained from the segmentation phase and used for classification experiments detailed in the following section.

VII. CLASSIFICATION EXPERIMENTS USING SUPPORT VECTOR MACHINES (SVM)

Classification is an essential phase in the recognition process. Classifiers map the feature vector that represents a character in one of the possible classes.SVM is a popular classification technique which is found to be effective in OHCR for Malayalam[9][10[11].

SVM is a supervised learning technique where training information with an accurate specification of different classes pertains to train a novel model. This novel model is utilized to test knowledge for appropriate categorization. In the SVM model, example points are represented in space and are mapped to divide the individual category examples by a clear gap that is as wide as possible. New instances are then assigned into the same location, predicting that they belong to a category based on the side of the class they fall. SVMs will attempt to build a model based on the given set of training samples. Each training data example is labeled as one of two classes. The SVM will try to attempt to separate the data instances into those two categories with a p-1 dimensional hyperplane, where p is the size of each data instance. This model can then be used on a new data instance to predict which category it would fall.

For every class, 100 samples were taken, which gave expected segmentation as per the reference set described in section V. Among these 100 samples, 60 were given for training and 40 for testing, *i.e.*, 6:4. The dimension of the feature set is five, as detailed above. The experiments are conducted in a *python* environment using *jupyter* notebook and necessary packages. Five-fold cross-validation is performed in the experiment to measure the average performance of the technique.

VIII. RESULT ANALYSIS

Three basic numeric scores, true positive, false positive, and false negative, are computed to evaluate the performance of the system. The performance scores *viz*. precision, recall, and F1-score of the system are computed based on the above parameters. The performance of the technique for a single fold validation is shown in table III, and the confusion matrix in Fig.9. Table IV shows the accuracy of the technique and the overall accuracy from five-fold cross-validation is obtained as 95.77% for RBF kernel. The experimental results show that pattern primitive segmentation is a better choice for OHCR in Malayalam.There is a drop in F1 score values for ആ <a:> and ഏ <e:>. This is due to the higher visual similarity observed in these group of characters especially അ <a>, ആ <a:>, എ <e>, ഏ <e:>.

Table-III: Performance score of a single fold

Character	Precision	Recall	F1-Score
അ <a>	1	1	1
ആ <a:>	1	0.5	0.67
ഇ <i>	1	1	1
ഉ <u>	1	1	1
ഋ <r>	1	1	1
എ <e>	1	1	1
ഏ <e:>	0.67	1	0.80
ഒ <o>	1	1	1

Table-IV: Accuracy of the proposed technique

Samples/ Character	Feature Dimension	No of Folds	SVM Kernel	Accuracy (%)
100	5	5	RBF	95.77

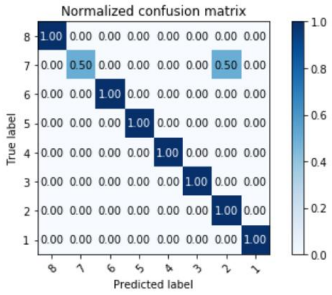


Fig. 9. Confusion Matrix corresponding to the single fold

IX. CONCLUSION AND FUTURE DIRECTIONS

The study demonstrated an approach to recognize single stroke online handwritten Malayalam vowel characters through the segmentation of characters into pattern primitives. It is found that the characters are easily recognizable when they are divided into pattern primitives. The average accuracy of 95.77% is promising compared to other studies in Malayalam OHCR that majorly focused on the entire character. The number of features in the study is well enough to classify the eight character classes. This also depicts the fact that feature selection must be carried out by analyzing the constituent character pattern primitives. The study is suitable for developing real-time applications in HCR where the pattern primitives play a major role.

The study only considered eight vowel characters and may be extended to the whole Malayalam character set in future extension studies. Implementation of a real-time application is also possible using the above methods. Other alternatives for classification and segmentation are also possible for better comparisons in further studies.

REFERENCE

1. Sriganesh Madhvanath Bharath A, "Online Handwriting Recognition for Indic Scripts," in OCR for Indic Scripts: Document Recognition and Retrieval.: Springer, 2008.
2. Pijush K. Ghosh, "An Approach to Type Design and Text Composition in Indian Scripts," 1983.
3. Tanmoy Dasgupta , Samar Bhattacharya Priyanka Das, "A Handwritten Bengali Consonants Recognition Scheme based on the detection of Pattern Primitives," in Second International Conference on Research in Computational Intelligence and Communication Networks., 2016.
4. K S Fu and P H Swain, "On Syntactic Pattern Recognition," in Proceedings of the Third Symposium on Computer and Communications, Florida, 1969, p. 164.
5. Santanu Chaudhury Abhijith Dutta, "Bengali alpha-numeric character recognition using Curvature features," Pattern Recognition, vol. 26, no. 12, pp. 1757-1770, 1993.
6. M. Borahan Tumer, Tunga Gungor Aleksei Ustimov, "A Low-Complexity Constructive Learning Automaton Approach to Handwritten Character Recognition," in ALC, 2010.
7. URS Ramer, "An Iterative Procedure for the Polygonal Approximation of Plane Curves," Computer Graphics and Image Processing, pp. 244-256, 1972.
8. D. G. Thakore N. J. Randive, "Static Hand Gesture Recognition using Freeman Chain Code and Neural Network," Int. J. Adv. Eng. Res. Dev. Sci. J., pp. 3–134, 2015.
9. Abdul Hameed Steffy Maria Joseph, "Online handwritten Malayalam character recognition using LIBSVM in MatLab," in 2014 IEEE National Conference on Communication, Signal Processing and Networking (NCCSN), 2014.
10. Baiju.K.B, Sabeerath.K, "Online recognition of Malayalam handwritten scripts — A comparison using KNN, MLP and SVM," in International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2016.

11. K. G. Sulochana, T. R. Indhu R. Ravindra Kumar, "Online Handwriting Recognition for Malayalam Script," Information Systems for Indian Languages, pp. 199-203, 2011.

AUTHORS PROFILE

Baiju.K.B, is Assistant Professor and Head of the Department of Computer Science at NMSM Govt. College Kalpetta, Kerala. He has completed his Masters in Computer Science from University of Calicut, MPhil degree in Computer Science from Bharathidasan University, Trichy. He is doing PhD at Department of Computer Science, University of Calicut under the supervision of Dr. Lajish. V.L. His area of interest includes Pattern Recognition studies in Computational Linguistics, Machine Recognition and Data Analytics.

Sabna.T.S, has completed MCA from University of Calicut in 2006. She was working as Assistant Professor at KMCT College of Engineering, Kozhikode. She is currently an MPhil Research Scholar at Department of Computer Science of Calicut University, Kerala working in the area of Advanced Pattern Recognition under the guidance of Dr.Lajish V. L.

Dr.Lajish.V.L, has been associated with University of Calicut, Kerala, as Head of the Department of Computer Science. He has worked as Scientist R&D in TCS Innovation Labs, Tata Consultancy Services Ltd. Mumbai, prior to joining the University. His prime areas of research include Digital Speech and Image Processing, Pattern Recognition algorithms, Indian language script technology solutions for masses. After his Masters in Computer Applications from Vellore Institute of Technology, he earned his Ph.D in Computer Science from University of Calicut, Kerala in 2007. He is a senior life member of International Association of Computer Science and Information Technology.