IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 9, September 2016

Online Recognition of Malayalam Scripts with Minimized Feature Dimensions

Baiju K.B[1], Sabeerath .K[2], Lajish .V.L[3]

Assistant Professor, Department of Computer Science, NMSM Govt. College, Kalpetta, India[1]

Research Scholar, Department of Computer Science, University of Calicut, Calicut, India[2]

Assistant Professor, Department of Computer Science, University of Calicut, Calicut, India[3]

Abstract: The objective of the work is to develop a database for online Malayalam script and identify online handwritten Malayalam characters with the use of Artificial Neural Network (ANN) and K-Nearest Neighbour (KNN) classifiers. The system uses minimum features (ADP, Aspect Ratio, Intersections and Octants)identified as relevant features and obtained 90.40% accuracy which is observed to be the best performance with these features and classifiers. Experiments are done for handwritten basic vowel alphabets (8) and consonants (36) in Malayalam. The database consists of 12800 samples of 64 Malayalam characters with 200 samples per character.

Keywords: Script Database; OHCR; KNN; ANN.

I. INTRODUCTION

The human-machine interaction has greater importance since the industrial age. And machines were hinged together for the progress of mankind till then. With the invention of computers, the realm of the human given its way to the computers as it overridden humane control centre than organs. The rapid progress in technology made it possible to replace the organs with technology clones. All the senses from ears, nose, skin, eyes and tongue have its replicates synonymous to technology. This lead to the need of an area where researches were aligned for achieving a humane machine amalgamation. It was named machine recognition as a fine tuned artifact of computer science.

Machine recognition majorly involves speech recognition, image processing, handwriting recognition (offline and online), video processing etc. Online handwriting recognition involves writing on digital touch pads using stylus, writing in paper using digital pen or writing in touch screen displays using fingers and automatically recognized by an OHCR engine. The strokes corresponding to each character includes coordinate points along the path, pen up, pen down information, time sequence and structural information [1]. The specialty of Online Handwritten Character Recognition (OHCR) is the preservation of traditional styles followed in the past era. Even though the writer is ignorant of technologies, he can enter the scripts in the data processing machine.

Online Handwriting replaces keyboards, which is the commonly used input device for data entry. The complexity of using various Scripts[2] in keyboards can be eliminated through OHCR. Nowadays online handwriting recognition is a popular method in mobile devices, which shows the wide admissibility of the technique. Inputting script in one's own language is a complexity associated with data processing machines. This will be a cumbersome task in India, a multi script-language country. An easy way to input regional languages in a traditional way as inputs for data processing is needed. This lead to the studies for natural interface for data input to the machine. The momentum thus made online handwriting recognition a frontier area of research in computer science. In OHCR, traditional handwriting input is combined with modern technologies to preserve the natural script input style. The OHCR system passes through database creation, pre-processing, feature extraction and classification phases in order to recognize a character.

II. CREATION OF DATABASE

A standard script database is necessary for testing and training the recognition engines. A database of 20 writers, each writing all 64 characters ten times, is collected to form 12800 samples. These samples were used as the database for the recognition engine. The database is designed for extending as a benchmark database in future works. The writers were provided with paper printed with 64 malayalam characters which include eight vowels, thirty six consonants and twenty conjunct consonants and modifiers used in malayalam. Each row of the paper can capture ten samples at a time. The technology used must be compatible enough to simulate traditional script input in a traditional way. Proposed data collection involves the usage of Hi-Tech e-write mate (Fig.1) for acquiring handwriting using pen and paper methods. The device includes a digital pen and sensor. Handwritings on the paper written using the digital pen will be stored in the sensor device as (x,y) coordinates of the neighbouring points. It can store up to hundred A4 sheets. After writing,

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 9, September 2016

the (x,y) coordinatesare available as a text file(Fig.2). The output of the digital pen is a text file of the stroke information. The text file involves pen tip movements x(t),y(t) as well as pen-up/pen-down switching. This kind of data can be regarded as dynamic representation of handwritings and also known as digital ink. The text file consists of three columns: first column indicates the pen-up and pen down information, second and third column indicates the x and y coordinates. Pen-down, pen moving, and pen up information are represented as 1, 2 and 3 in first column. Sixty four basic scripts were acquired and arranged with 200 samples per script in various folders as text files (12800 samples). The proposed system do not consider statistical properties of the stroke in order to minimize features used in the recognition process. The character /a/ plotted in Matlab is shown in Fig.3.

Fig.1 Hi-Tech e-write mate device

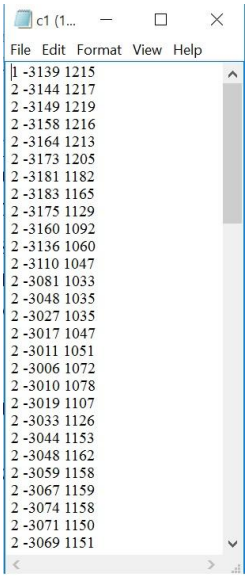

Fig.2 Text file of a stroke

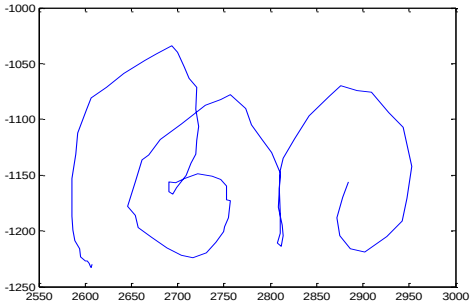

Fig.3 Character /a/ plotted in Matlab

III.PREPROCESSING

The samples were applied for pre-processing to reduce noises and assimilations. The acquired data is cleansed through normalisation, smoothening and resampling phases (Fig. 4) before feature extraction.

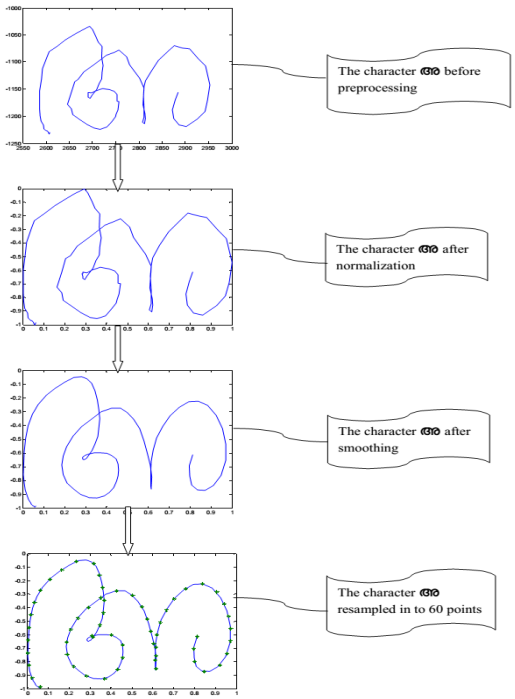

Fig.4 The character /a/ passing through various pre-processing phases

A. Normalisation

Normalisation is the initial phase in pre-processing, where the extracted (x,y) coordinates were standardized to obtain all the points in a range 0 to 1or 0 to -1.This is accomplished using division by the difference of maximum of values and minimum of values for both x and y.The normalized points conserve the shape of the character even if the device coordinates changes. The algorithm is selected to keep the essential structure of the stroke invariant under standard transformations.

Algorithm for normalisation

For all x_i,y_i till x_ny_n perform

$x_i = (x_i - \min(x)) / (\max(x) - \min(x))$

$y_i = (y_i - \min(y)) / (\max(y) - \min(y))$

The new set of x,y will be normalised

B. Smoothening and Filtering

The normalised strokes must be smoothened to remove jitters and used Gaussian function(1) for smoothening.

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{\frac{-(x^2+y^2)}{\sigma^2}} \qquad (1)$$

DOI 10.17148/IJARCCE.2016.59106

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 9, September 2016

where σ is the standard deviation of the points.

Algorithm for filtering

Change stroke values $(x_i,y_i)\ldots..(x_N,y_N)$ as

$$X(i)=\frac{x(i-N)+....+x(i-1)+\infty x(i)+x(i+1)+.....x(i+N)}{N+\infty}$$

$$Y(i)=\frac{y(i-N)+....+y(i-1)+\infty y(i)+y(i+1)+.....y(i+N)}{N+\infty}$$

where *X(i)* and *Y(i)* are new x and y values, the term ∞ denotes the optimised angle subtended by the succeeding and proceeding points to keep sharp edges in the stroke.

C. Resampling

The smoothened points were resampled with N points per stroke. Resampling ensures uniformity of the coordinate size. The resampling is chosen by ensuring the shape distortion problems. Resampled points provide a constant number of points to represent any character. The proposed system resamples the characters with 60 points. Resampling extensively uses interpolation methods to predict the next coordinate automatically from the existing coordinate values. The resampling algorithm uses simple linear interpolation between pairs of points. The resampled strokes are represented as a sequence of points regularly spaced in arc length. The benefit is that the entire sixty points were distributed around the complete arc of the stroke and yields better feature vector values.

IV. FEATURES

Features are the essential components of the character. The system uses four features for the training and testing purpose namely accurate dominant points, aspect ratio, starting and ending octants and intersections. The dimension of feature vector is five. These features are identified as relevant features for malayalam scripts with a detailed review of earlier works in these area.

A. Accurate Dominant Points

The number of dominant points of a stroke cluster provides a well structural description. Dominant points are those points where the values of curvature point (q_i) change noticeably. When there is a noticeable change from one point to next point then mark it as a dominant point. Niranjan et al. [3]define point P to be a dominant point d_i, if the following two conditions (2) are satisfied:

$$(q_{i+1}- q_i + 8 \% 8) \geq CT \text{ and}$$
$$(q_i - q_{i+1} + 8 \% 8) \geq CT \quad (2)$$

where, CT is a curvature threshold for retaining any point as a dominant point, % is the modulo operator and 8 corresponds to the number of levels of quantization of the angle. CT can take any value from the set {0,…4}. By default, the first and the last points of curve are considered as dominant points. Our procedure starts by marking the first pen position as a dominant point. Starting from the current ADP, we calculate the absolute value of the angle between pen directions at successive points and accumulate it along the online trace as long as the cumulative sum is less than a threshold CT. The pen position, at which the accumulated angle exceeds CT, is marked as the next dominant point and the process continues till the end of the trace. The resulting number of ADP extracted is used as a feature for recognition.(Fig. 5)

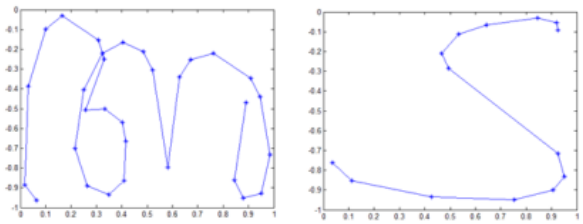

Fig.5 Accurate Dominant Points of /a/ and /ta/

B. Aspect Ratio

The aspect ratio of a 2-d curve/image describes the proportional relationship between its width and its height. Here the width of the character stroke is divided with its height to obtain a value. The value obtained is saved as a feature in the set(Fig. 6).

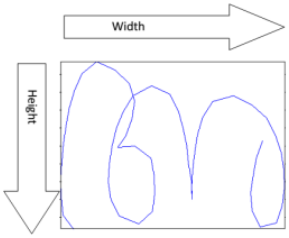

Fig.6Aspect Ratio of the character /a/

C. Starting and Ending Octants

This is a unique feature of most of the Indic scripts. The octant in which the character starts and ends is considered as a feature set. The entire stroke is plotted and divided into eight octants (Fig. 7).

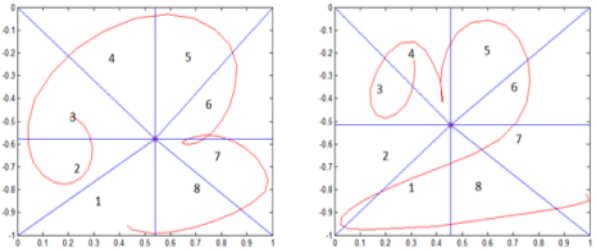

Fig.7Start and End Octants for /o/ and /i/

D. Intersection Points

The point where the character itself makes a crossing is identified.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 9, September 2016

The number of such intersections (crossings) was recorded as a feature (Fig. 8).

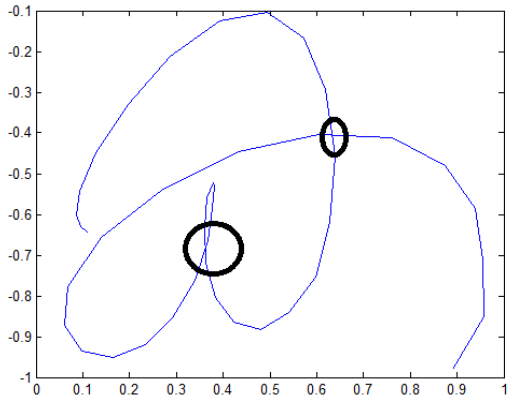

Fig.8. Intersections of the character /ka/

V. CLASSIFICATION TECHNIQUES

Classification is the most important phase in recognition process. Classifiers map the feature vector that represents an input character into one of the possible character classes. For most practical purposes, this is the final step and although the choice of features has a larger bearing on the overall accuracy and feasibility of the HCR algorithm. The selection of an appropriate classifier ensures low misclassification and rejection rates. We have used two types of classifiers for recognition of Malayalam script namely Artificial Neural Network (ANN) with MLP and K Nearest Neighbour (KNN).

A. Artificial Neural Network and Multi Layer Perceptron(MLP)

MLP is a feed forward ANN model that maps sets of input data onto a set of appropriate outputs. It consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. MLP utilizes a supervised learning technique called back propagation for training the network[4]. MLP is a modification of the standard linear perceptron and can distinguish data that are not linearly separable.

B. K-Nearest Neighbour classifier(KNN)

K-Nearest Neighbour classifier is a special type of nearest neighbour classifier. It uses the instance based learning by relating unknown pattern to the known according to some distance or some other similarity function. K is the number of nearest neighbours to be considered and the class of majority of these neighbours is determined as the class of unknown pattern [5] [6].

VI. RESULTS AND DISCUSSIONS

The work focused on two major objectives, for developing a bench mark database for online Malayalam handwritten scripts and recognizing Malayalam characters with minimum features using the above database. The database includes 12800 samples of 64 characters. The system is trained and tested for 100 samples of 44 characters which include 8 vowels and 36 consonants. Four features identified as relevant to the work has been chosen, namely accurate dominant points, starting and ending octants, aspect ratio and intersections. The test is carried out in a core i3 (2.7 GHz) machine with 3 GB RAM. The recognition rate obtained is 90.40% with ANN and 82.18% with KNN as listed in Table.1.

The work reported an accuracy which is similar to other reported works in Malayalam having more number of features with KNN and MLP classifiers. This shows selection of relevant features play a key role in recognition accuracy. The system can be trained and tested with more samples in future works to test the effect of increased samples than increased number of features. The future study may also be focused on recognizing all Malayalam characters including numerals using new classification strategies by considering recognition time.

TABLE I RECOGNITION RATE

Classifier	Feature	Recognition Rate (%)
MLP	ADP, Aspect Ratio, Octants, Intersection	90.40
KNN	ADP, Aspect Ratio, Octants, Intersection	82.18

REFERENCES

[1] Van der Geer J, Hanraads JAJ, Lupton RA. The art of writing a scientific article. J Sci Commun 2000;**163**:51-9
[2] Strunk Jr W, White EB. The elements of style. 3rd ed. New York: Macmillan; 1979.
[3] N Joshi, G Sita, A G Ramakrishnan, S Madhvanath, Comparison of Elastic Matching Algorithms for Online Tamil Handwritten Character Recognition, Proc. IWFHR (2004) 444-449.
[4] Verma B.K, "Handwritten Hindi Character Recognition Using Multilayer Perceptron and Radial Basis Function Neural Network", IEEE International Conference on Neural Network, 4, pp. 2111-2115, 1995.
[5] K-NNclassifier,[Online]. Available: http://www.fon.hum.uva.nl/praat/manual/ kNN_classifiers_1__What_is_a_kNN_classifier_.html.
[6] Wikipedia website- K-Nearest Neighbour Algorithm, [Online]. Available:http://en.wikipedia.org/wiki/K-arest_neighbor_algorithm.

BIOGRAPHY

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 a PhD scholar under the supervision of **Dr. Lajish. V.L**. His area of interest includes Pattern Recognition studies in Computational Linguistics, Machine Recognition and Data Analytics.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 9, September 2016

Sabeerath K has completed M.Sc. Computer Science from Calicut University in 2009. She is currently an MPhil Research Scholar at Department of Computer Science of Calicut University, Kerala working in the area of Pattern recognition under the guidance of **Dr. Lajish V. L.**

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.

DOI 10.17148/IJARCCE.2016.59106