Primitive Segmentation of Online Malayalam Handwritten Strokes using Ramer-Douglas-Peucker Algorithm and Eight Direction Freeman Code

Baiju.K.B[1], Dr. Lajish V.L[2]
[1]*Asst. Professor, Dept. of Computer Science, Govt. College, Kalpetta, Kerala*
[2]*Asst Professor, Dept. of Computer Science, University of Calicut, Kerala*

Abstract— This paper implements the segmentation of Malayalam Online Handwriting strokes using Ramer-Douglas-Peuker algorithm (RDP), Eight Direction Freeman Code (EDFC) and Combined approach of both. The segment points obtained from these techniques were compared against a manually marked reference set to verify segmentation accuracies. Experiments show that combined approach gives promising results compared to the other methods. The segments obtained in the work can be used as sub-primitive set for OHCR in Malayalam.

Keywords—*RDP; EDFC; OHCR; Malayalam;*

I. Introduction

Machine Learning, a branch of Artificial Intelligence has infinite dimensions to coalesce human- machine interaction. It has proven laudable in most of the scientific problems pertaining to training an artificial model to mimic human motor learning model. In machine learning, a computing device is learning from data by continuous training through various algorithms. The model obtained is a true clone of the natural model. For a human, all the five organs are continuously monitoring the environment and trained to act upon situations through certain communication modes. And the medium of the communication are writings, actions, and sound. Hence Machine learning always tries to automate these communications in an algorithmic way. The proposed work exploits the writing aspect of communication and implements a segmentation scheme which is useful in recognition process of Online Handwritings in Malayalam.

The rapid growth of digital devices depicts enormous amount of data in digital form in the storage devices. As humane prefers to communicate in natural languages; the use of Online Digitisers, Smart boards, Touch screen kiosks and mobile devices are increasing for electronic writing. But most of them store the pages in an image format, where the exact information is in a limited region. Online Handwriting is an exception, where the handwritings are stored as a sequence of (x, y) positions. This sequence of points exactly represents the handwritten strokes and can be treated as a non linear time series. The non linearity of the handwriting makes the major challenge to fit the sequence to a general mathematical model. Also the series is progressive in the order of time. By analyzing the time, one can identify the order in which a letter is written in to the paper/device. This time dependency makes the online handwritings to be interpreted as a time series. Time series segmentation is a fundamental component in the process of analysis and research of time series data [1].In the work, the direction and shape features are considered,rather than the time series nature.

The proposed work segments the given handwriting stroke series of a grapheme into N segments based on Douglas Peuker Algorithm [2] and Eight Direction Freeman Code [3]. The entire series is thus broken in to sub-series namely primitive segments based on the reference stroke set. The reference set for vowels is shown in Fig.1.The reference set is created from an online stroke database created using e-write mate digital pen device. The reference set is manually marked with segmentation points. The dataset is segmented and tested against the reference set and the segmentation accuracy is obtained.

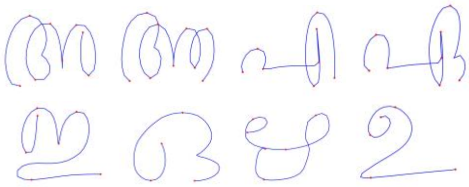


Fig.1 Reference Grapheme set of Malayalam Vowels

The most common feature of Malayalam graphemes is that, it has lot of similarities in shape. These similarities are useful when recognition is performed on them. The simplest way to test the similarity is by divide and conquers method. First the graphemes are segmented and similar segments are identified visually and may be stored to form an online primitive set [4] data base. This primitive set can be used for recognizing various strokes in Malayalam especially in the context of real time recognition based on predictive models.

The rest of the paper is organized as follows. The section II describes data acquisition and pre-processing. Section III describes Ramer-Douglas-Peucker algorithm based segmentation, Direction Code based segmentation, and Combined approach for segmentation. Experimental results are outlined in section IV and Section V presents conclusive remarks with future directions.

II. Data Acquisition and Preprocessing

Data Aquisition

A well structured data set ensures better accuracy in research problems involving data. The proposed work developed a database of 15 writers, each writing 64 Malayalam grapheme in Malayalam 10 times to form 9600 samples. These samples were used as the dataset for segmentation The dataset is designed to be extended as a benchmark database in future works. The writers were provided with A4 sized paper printed with 64 Malayalam graphemes which include 8 vowels, 36 consonants, 5 chillu characters and 15 other symbols.Each row of the paper can capture 10 samples at a time.

Proposed data collection involves the usage of Hi-Tech e-write mate for acquiring online handwriting using pen and paper methods. The device includes a digital pen and sensor and The device ensures simulation traditional script input in a traditional way. 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 hundred A4 sheets in memory. The (x,y) coordinates are available as a text file when the writing process is over.

Pre-Processing

Pre-processing is an essential component in most of the data processing system with natural interfaces.The text file acquired by the device always contain noises like dots, jitters, over writings, skewing ,slant etc. To reduce these imperfections, the device data must be applied against a pre-processor. The acquired data passes through Normalization, Smoothening and Resampling phases through the pre-processor and is cleansed before segmentation [5].

The stroke series were normalized to the range of [0, 1] using min-max normalization. The normalized data is smoothed using moving average filter, which is a low-pass filter with filter coefficients equal to the reciprocals of the span. A low-pass filter ensures preservation of smooth edges ,which is common in most of the Malayalam graphemes.The stroke series has been re-sampled to 60 points for all the samples. The different phases of the Malayalam grapheme ഇ <i> reproduced in MATLAB are given in Fig.2.

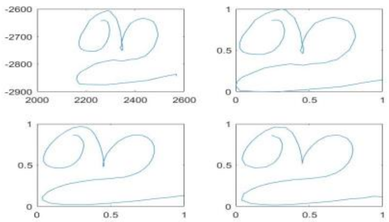


Fig.2 Pre-processing phases of grapheme ‘ഇ’(/i/)

III. Primitive Segmemntation

A. *Ramer-Douglas-Peucker Algoritm based segmentation*

In the segmentation based on Ramer-Douglas-Peucker (RDP) algorithm [6], the input curve to the algorithm is the set of points of a grapheme. A distance dimension parameter ε is set by the system which varies between 0 and 5. In our experiment, the value of ε is set as 0.15.The parameter value is based on the reference set segmentation shapes in our work. For a different reference set, the value must be decided by experimenting with various parameter values. The curve is recursively divided by measuring the perpendicular distance. The algorithm marks the first and last point to be kept. The point that is furthest from the line segment with first and last points as end points is found; this point is obviously furthest on the curve from the approximating line segment between the end points. If the point is closer than ε to the line segment, then any points not currently marked to be kept can be discarded without the simplified curve being worse than ε.

If the point furthest from the line segment is greater than ε from the approximation then that point must be kept. The algorithm recursively calls itself with the first point and the furthest point and then with the furthest point and the last point, which includes the furthest point being marked as kept. When the recursion is completed a new output curve can be generated consisting of all and only those points that have been marked as kept. All the points marked as kept are now segment points. The new set of points for the grahpemes ണ<na>, ഒ<o>, ത<tha> after applying the algorithm is given in Fig.3.

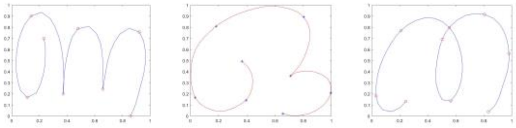


Fig. 3. RDP Segmentation Points

B. *Direction Code based Segmentation*

In this method,the stroke series were transformed into a chain code based on 8-direction freeman code [7]. The difference between neighbouring points corresponding to a grapheme were compared and the difference in x nad y values decide the direction of a point.The conditions which decides the direction are listed in Table 1.

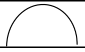
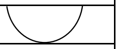
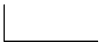
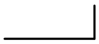
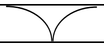
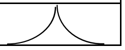
TABLE I. Condition and Direction Code

Condition	Direction Code	Direction
x>0 and y=0	0	→
x>0 and y>0	1	↗
x=0 and y>0	2	↑
x>0 and y>0	3	↖
x<0 and y=0	4	←
x<0 and y<0	5	↙
x=0 and y<0	6	↓
x>0 and y<0	7	↘

Afte me, the chain co ng point to the ned for directio code in between for eg. 1110777. In this sequence 0 must be converted to 1 or 7 for obtaining the correct direction.

The grapheme is segmented by moving direction code window. The code window contains 4 codes with first pair represents starting direction and the second pair represents changing direction. The code windows used in the work is shown in Table II.

TABLE II. CODE WINDOW AND SEGMENTATION SHAPES

Code window	Segmentation shapes
1177 or 3355	
7711 or 5533	
7700 or 6600 or 5500 or 7711 or 6611 or 5511 or 2200	
7744 or 6644 or 5544 or 7733 or 6633 or 5533 or 7711	
7711 or 5533	
1177 or 3355	

If the code window matches with the sequence in the chain code, the position is marked. After all the windows have been passed through the chain code, the algorithm returns a collection of points marked. The marked points are sorted in the order of x-values. There may be same points marked by different windows. These redundant points are eliminated and finally marked as a segmentation point. Direction code and segmentation points by the algorithm for the grapheme ഞ<nja> is shown in Fig.4

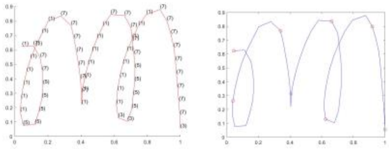


Fig.4 Direction code and Segmentation points with EDFC

C. Combined Approach

It is found that for some graphemes,EDFC performs better for eg. ഹ,ന,ഠ etc..Where as RDP is better for ല,വ,ഥ etc. Hence,we combined both of them for experimenting better results. In the combined approach, the the segmentation points obtained from the RDP algorithm and EDFC algorithm were combined to form new set of segmentation points.The new set points were sorted in the order of x values. After sorting,the points were reduced for redundancy,close points and unrefered directions.

The following fine tuning were applied for point reduction. If any unrefered direction as per reference set is present,then remove them.If both of the above methods have similar segmentation points,keep only one of them.If the two points are 4 codes apart,then keep both of them else, remove one of them. If we are applying RDP algorithm prior to EDFC algorithm,then the point to be removed is the second one.In the other case i.e, EDFC algorithm prior to RDP algorithm,the point to be removed is first one. In our experiments,we performed the latter scheme. The steps involved in the combined approach is listed in Fig. 5.

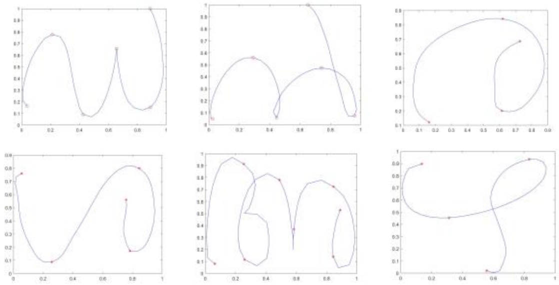


Fig. 5 Segmentation with Combined Approach

The steps involved in the combined approach are given below.

1. Adjust the EDFC algorithm based on the reference set.
2. The acquired points were fed to the above EDFC Algorithm.
3. The output is a set of indices corresponding to the segmentation points.
4. Apply the previous acquired points to the RDP algorithm with ε=0.15.
5. Append the output set of indices with previously generated indices with EDFC.
6. Sort the set of indices in ascending order and remove duplicates if any.
7. If the indices difference is less than 4 for any two nearby points, then remove the first one.
8. The final set of indices corresponds to the segmentation points.

IV. EXPERMENTAL RESULTS

All the algorithms were applied against 1280 samples (20 samples of 64 graphemes).The samples were unique for all the algorithms. The samples with segmentation points were plotted using MATLAB. The output were visually compared with reference set and found the number of graphemes correctly segmented as per reference set. It is found that combined approach gives more accuracy as per reference set compared to the other methods. The accuracy obtained by the three approaches is listed in Table.III.

TABLE III. SEGMENTATION ACCURACY

Method	Segmentation Accuracy
RDP Segmentation	63.13
EDFC Segmentation	67.65
Combined Approach	91.40

The work presented an approach which segments the Malayalam graphemes as per the reference set. The method is efficient for segmenting a grapheme in to sub-segments. As the Malayalam grapheme set contains enormous visually similar sub-segments (primitives), a grouping of primitives can be created using the above method. Also the work may be used to create a database of sub primitives. These primitives can be used in the recognition of Online Handwritten Characters (OHCR) in Malayalam. The method also preserves direction and shape features, which are relevant in most of the OHCR experiments.

The present work considered only 64 graphemes in Malayalam. It can be extended to span all the Malayalam graphemes. Since the work relies on direction and shape features, the OHCR can be tuned to recognize the strokes with negligible time quantum.

References

[1] M. Lovrić, M. Milanović, and M. Stamenković, "Algorithmic Methods for Segmentation of Time Series: an Overview," *Jcebi*, vol. 1, no. 1, pp. 31–53, 2014.

[2] S. Wenzel and W. Förstner, "Finding Poly-Curves of Straight Line and Ellipse Segments in Images<BR>Segmentierung von Pixelketten in Geraden- und Ellipsenelemente," *Photogramm. - Fernerkundung - Geoinf.*, vol. 2013, no. 4, pp. 297–308, 2013.

[3] P. Annapurna, S. Kothuri, and S. Lukka, "Digit Recognition Using Freeman Chain Code," *Int. J. Appl. or Innov. Eng. Manag.*, vol. 2, no. 8, pp. 362–365, 2013.

[4] P. Das, T. Dasgupta, and S. Bhattacharya, "A handwritten Bengali consonants recognition scheme based on the detection of pattern primitives," *Proc. - 2016 2nd IEEE Int. Conf. Res. Comput. Intell. Commun. Networks, ICRCICN 2016*, no. September, pp. 72–77, 2017.

[5] B. Q. Huang, Y. B. Zhang, and M. T. Kechadi, "Preprocessing Techniques for Online Handwriting Recognition," *Seventh Int. Conf. Intell. Syst. Des. Appl. (ISDA 2007)*, pp. 793–800, 2007.

[6] M. Visvalingam and J. D. Whyatt, "The Douglas- Peucker Algorithm for Line Simplification: Re- evaluation through Visualization," *Comput. Graph. Forum*, vol. 9, no. 3, pp. 213–225, 1990.

[7] N. J. Randive and D. G. Thakore, "Static Hand Gesture Recognition using Freeman Chain Code and Neural Network," *Int. J. Adv. Eng. Res. Dev. Sci. J. Impact Factor (SJIF*, vol. 2, no. 5, pp. 3–134, 2015.