Histogram Based Connected Component Analysis for Character Segmentation

Shilpa C Vijayan*, Jyothy R.L.,Anilkumar A.****

*Department of Computer Science and Engineering, College of engineering Karunagappally.

***Abstract*-** Segmentation is the major part of any character recognition system. The efficiency of recognition can be improved by imposing an effective segmentation method. In Malayalam handwritten character recognition system character segmentation is more complex due to the varying writing styles and conjunction between the characters. There is a vast no of researches have been done in this field of Malayalam handwritten character recognition. In this paper various existing methods are analyzed and a method is proposed for effective segmentation.

***Index Terms*-** Character Recognition, Segmentation, Projection Analysis, Connected Component Analysis.

I. INTRODUCTION

Character recognition has been a most active research area in the field of pattern recognition system. There is a large number of algorithms have been proposed for the same domain. What it matters is how efficient the algorithm and how feasible it is. A very good segmentation technique can improve the efficiency of recognition. There have been reported effective techniques in the field of other languages but it is more complex to find an efficient segmentation technique for Malayalam handwritten documents. Malayalam exhibits no inherent symmetry and thus making the segmentation task very complex.

Character segmentation is an operation that decomposes an image of a sequence characters into sub images of individual symbols. It is one of the decision processes in a system for character recognition. Segmentation can be broadly classified into three types.

1. Explicit segmentation.
2. Implicit Segmentation.
3. Holistic Segmentation.

In explicit segmentation the word image is partitioned into sub images of individual characters. The process of cutting the word images into character sub images is termed as dissection. This technique is used to find all the interconnection between the character images and partition the image through the detected ligatures. The explicit features are likely to be occur within or between the characters in the form of ligatures. The properties of the segments obtained with those expected for valid characters are height, width, separation from neighbouring components, disposition along a baseline, etc.

Implicit segmentation approaches are applied as an alternative to integrate segmentation and recognition process. There is no complex dissection algorithm has to be built. Implicit segmentation based recognition systems searches the image for components that matches the class of alphabets. Hidden Markov Model (HMM) are become evident for this approach. The Markov model represents state-to-state transitions within a character. These transitions provide a sequence of observations on the character. Features are typically measured in the left-to-right direction. This facilitates the representation of a word as a concatenation of character models. In such a system segmentation is (implicitly) done in the course of matching the model against a given sequence of feature values gathered from a word image. That is, it decides where one character model leaves off and the next one begins, in the series of features analyzed.

Holistic approach is also known as segmentation free process, which recognizes the entire word as a unit. The method is usually restricted to predefined lexicon. Since they do not deal directly with the letters but only with words. This point is critical when training on word samples is required: a training stage is thus mandatory to expand the lexicon of possible words. This property makes this kind of method more suitable for applications like check recognition.

Malayalam is one of the south Indian Dravidian languages with about 35 million speakers. It is written in vattezhuthu. Modern Malayalam script is derived from grantha script. The character set consist of fifty three letters called akaras and 13 vowels,2 left vowel signs ,7 right vowel signs, some appear on the both side of conj/constants,30 commonly used conjuncts.

Features of characters

- Type of writing system: A consonants with an inherent vowel is a syllabic alphabet. Diacritics

can appear above, below, before or after a consonant which is used to change the inherent vowel.

- Vowels are written as independent letters appear at the beginning of a syllable.
- When certain consonants occur together, special conjunct symbols are used which combine the essential parts of each letter.

II. RELATED WORKS

There is a lack of efficient segmentation technique for Malayalam characters. For finding an efficient segmentation technique here we come across between certain methods. Dissection means decomposing an image into sequence of sub images using general features. In this method of segmentation which consider the method of whitespace and pitch (the number of characters per unit of horizontal distance), projection analysis (consist of a simple running count of the black pixels in a certain column), connected component processing (in which bounding box analysis and splitting of connected components are performed).

The basic idea of connected component analysis with a simple recognition logic whose role is not to label characters but rather to detect which components are likely to be single, connected or broken characters. Splitting of an image classified as connected is then accomplished by finding characteristic landmarks of the image that are likely to be segmentation points, rejecting those that appear to be situated within a character, and implementing a suitable cutting path.

Projection histogram [6] counts the number of pixels in specified direction. This approach can applied in three directions of horizontal, vertical and diagonal. The histograms are computed by counting the number of foreground pixels. In horizontal histogram the pixels are counted in row wise. In vertical the counting is done column wise.

The bounding box analysis is a method of connected component labelling. In this by testing their adjacency relationship to perform merging, or their size and aspect ratios to trigger splitting mechanism, segmentation can be performed.

III. PROPOSED WORK

Segmentation is an important part of any recognition system. The above methods alone will not provide an effective segmentation method for Malayalam characters. For an effective segmentation we can combine two methods (the connected components with profile method) can provide effective segmentation for Malayalam characters.

Algorithm

Step 1. Find the size of the image document [p, q].

Step 2. Calculate the column sum of pixel value, called Vertical Projection (VP) value for each column from 1 to q.

for i=1 to q

$$VP(i) = \sum_{j=1}^{p} Image(j)$$

end

Step 3. If the number of column with VP value zero, first coming column with the non-zero VP value is chosen as the segmentation point. The first column with VP zero value is chosen as last segmentation point.

Step 4. Apply connected component analysis.

IV. EXPWRIMENTAL RESULTS

Here in this paper the different dissection methods are implemented. The result of each method is shown below.

Fig 1. Result of connected component method.

Fig 2. Result of profile method.

From the result we came to the solution as the conjucted characters cannot be segmented efficiently. The proposed method is implemented and the result is shown in the fig 3.

Fig 3. Result of proposed method.

V. CONCLUSION

In this paper different dissection methods are implemented and based on the experiment none of the method alone can provide an efficient segmentation technique. The projection analysis combined with connected component analysis can provide efficient method for segmentation.

REFERENCES

[1] Casey, Richard G., and Eric Lecolinet. "A survey of methods and strategies in character segmentation." Pattern Analysis and Machine Intelligence, IEEE Transactions on 18.7 (1996): 690-706.

[2] Tyagi, Karishma, and Vedant Rastogi. "Implementation of Character Recognition using Hidden Markov Model." International Journal of Engineering Research and Technology. Vol. 3. No. 2 (February-2014). ESRSA Publications, 2014.

[3] Shanjana, C., and Ajay James. "Online Recognition of Malayalam Handwritten Text." Procedia Technology 19 (2015): 772-779.
[4] John, Jomy, K. V. Pramod, and Kannan Balakrishnan. "Online handwritten Malayalam Character Recognition based on chain code histogram." Emerging Trends in Electrical and Computer Technology (ICETECT), 2011 International Conference on. IEEE, 2011.
[5] John, Jomy, K. V. 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, pp. 199-204. IEEE, 2014.
[6] Kazi, M. M., and Y. S. Rode. "Handwritten and Printed Devanagari Compound using Multiclass SVM Classier with Orthogonal moment Feature."(2013).
[7] Papavassiliou, Vassilis, Themos Stafylakis, Vassilis Katsouros, and GeorgeCarayannis. "Handwritten document image segmentation into text lines and words." Pattern Recognition 43, no. 1 (2010): 369-377.
[8] P. Gader, M. Magdi and J-H. Chiang, Segmentation-Based Handwritten Word Recognition, Proc.USPS 5th Advanced Technology Conference, Nov/Dec 1992.
[9] S. Bercu and G. Lorette, On-line Handwritten Word Recognition: An Approach Based on Hidden Markov Models, Pre-Proceedings IWFHR III, Buffalo, page 385, May 1993.

AUTHORS

First Author – Shilpa C Vijayan, PG Scholar, Department of Computer Science and Engineering, College of Engineering Karunagappally, CUSAT, Kerala, shilpacvijayan@gmail.com.
Second Author – Jyothi R L, Assistant Professor, Department of Computer Science and Engineering, College of Engineering Karunagappally, CUSAT, Kerala, jyothianil@gmail.com.
Third Author – Anilkumar A, System Analyst Department of Computer Science and Engineering, College of Engineering Karunagappally, CUSAT, Kerala, anilanirudh@gmail.com