International Journal of Advance Research in Science and Engineering
Vol. No.6, Issue No. 04, April 2017
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

SEGMENTATION APPROACHES USED FOR DEVANAGARI AND GUJARATI OCR SYSTEM – A SURVEY

Sulabh Bhatt[1], Kiran Bidua[2], Prachi Shah[3]

[1]*CSE Department, Institute of Technology, Nirma University (India)*

[2]*MCA Department, Anand Institute of Information Science, GTU (India)*

[3]*IT Department, Birla Vishvakarma Mahavidyalaya, GTU (India)*

ABSTRACT

Optical Character Recognition (OCR) systems are very well developed and providing high accuracy for extracting text from printed or hand written documents and images for English language. But they are still not providing better accuracy for Indian languages due to some language constraints. Hindi language uses Devanagari script. In this paper we have presented a brief survey of various segmentation techniques applied by different researchers for both Devanagari and Gujarati scripts separately.

Keywords: Devanagari script, Gujarati script, OCR, OCR Phases, Segmentation

I. INTRODUCTION

Optical Character Recognition (OCR) is a technology in which a computer can automatically understand the image of a text document, either handwritten or printed, and converts it into editable text. OCR belongs to the Artificial Intelligence (AI) domain. It is a sub field of pattern recognition. This conversion process consists of various phases such as Digitization, Pre-processing, Segmentation, Feature Extraction, Classifications and Recognition. The output of one phase is the input of the next phase [1].

1.1 Phases of OCR

1.1.1 Digitization

Digitization is a process in which a handwritten document is converted into electronic format. In digitization, a document is scanned and its electronic representation as an image file format is produced.

1.1.2 Pre-processing

In Preprocessing, noise and handwriting variations are removed.

1.1.3 Segmentation

In segmentation, a digital image is partitioned into multiple regions.

1.1.4 Feature Extraction

In this phase, features of individual character are extracted. The performance and accuracy of a character recognition system highly depends on the features that are extracted in feature extraction phase.

Out of these phases, Image segmentation is one of the most critical and essential phase. With the help of

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

segmentation, we can understand images and extract information or objects from it.

1.2 Segmentation Techniques

There are two different approaches for image segmentation:

1. Discontinuity based approach

2. Similarity based approach

1.2.1 Discontinuity based approach

In this approach, the partition is carried out based on some abrupt changes in intensity levels in an image or abrupt changes in grey levels of an image.

Under this approach, major interest lies in identification of isolated points, lines or edges.

1.2.1.1 Edge Detection

An edge is a set of connected pixels that lies on the boundary between two regions that differ in grey value. The pixels which are on the edge are called edge points. We can extract an edge by computing the derivative of the image function.

1.2.2 Similarity based approach

In this approach, those pixels of an image which are similar in some sense are grouped.

1.2.2.1 Thresholding

Thresholding is the simplest approach under similarity based technique. In this technique, a threshold level is defined. So all the pixels having less intensity than the threshold value will belong to one region and the pixels having higher intensity than the threshold value will belong to another region.

1.2.2.2 Region growing based approach

In this approach, the adjacent pixels of a particular pixel which are similar in some sense (intensity value is almost same) are grouped together. So, in this technique, starting from a particular pixel we try to grow the region based on similarity.

1.2.2.3 Region splitting and merging

In this technique, an image is split into a number of components and after splitting we try to merge some of those sub components which are similar in some sense to have a larger segment.

II. SEGMENTATION APPROACHES USED FOR DEVANAGARI SCRIPT

There are various approaches which are available for segmentation of printed and handwritten text image. We will start by discussing about the segmentation approach which is both for handwritten and printed text image.

2.1 Histogram Approach

A straightforward histogram based approach bounded box method was adopted for segmentation of available documents into line followed by words and characters respectively [2]. After preprocessing the image, they applied various methods for segmentation of documents. The flow of segmentation was as follows: 1] Extracting line from whole image. 2] Extracting word from line and 3] Extracting character from word. We are reviewing here the segmentation proposed by them step by step.

Step 1: Line Segmentation

This segmentation uses horizontal approach for segmenting document image into rows i.e., identifying line in

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

page. First the image is constructed. Then the white pixel in each row is counted. The rows which are not containing white pixel are replaced by 1. And then such empty rows are replaced by 0. And at the last the original pixels are copied in separate file using bounding box where segmented lines can be extracted as an output.

Step 2: Word Segmentation

Initially the horizontal projection method is used to sum all white pixels in column. Then the vertical histogram for the image is constructed. White pixels in column are counted. Using histogram the column not containing white pixels are replaced by 1 and later on the image is inverted to make the empty rows as 0 and only text words have original pixels. The original pixels are copied into bounding box and saved in separate files where words can be extracted as an output form the input lines.

Step 3: Character Segmentation

Little bit modification is made here to segment a character from a word. The Matlab bwmorph function is used to get thinned image. Now the white pixels are counted in each column and the position containing single white pixel is found and also such single white pixels are replaced by 1. Such image is inverted to make column as 0 and text character contains original pixel.

The original pixels are copied to bounding box using bounding box and saved in separate file. Thus, characters are extracted as an output from word as input.

2.2 Result of Histogram approach

The result of the discussed approach shows that accuracy decreases from line followed by word to character. Character segmentation accuracy is less in Devanagari due to the two dimensional script as consonants are modified in many ways from top, bottom, left or right to form a meaningful letter.

2.3 Overcoming in Histogram approach

There were various flaws in discussed approach like it is difficult to identify exact connecting points in compound letters for segmentation. The upper and lower modifier requires different approach for segmentation. The full stop and comma appears to be same and thus difficult to distinguish between them. The handwritten unconnected compound letter and handwritten unintentionally connected letter is also critical to distinguish in segmentation.

III. SEGMENTATION FOR HANDWRITTEN TEXT IMAGE IN DEVANAGARI SCRIPT

3.1 Projection profile approach

Now we will discuss about the segmentation for purely unprinted text image of Devanagari script because it is more difficult to segment handwritten script as compared to the printed script due to the overlapping of character.

Here we have discussed the available algorithm for segmentation of text line in handwritten skewed and overlapped Devanagari script [3].

The algorithm proposed in the above paper deals with skewed text and also with overlapping and touching characters. Projection profile technique was used in the proposed algorithm by them. The algorithm uses piecewise projection profile. The segmentation uses the gap between the text lines.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

3.1.1 Overcoming in projection approach for skewed handwritten and overlapped characters

The above algorithm works well for handwritten text even if the text is skewed but it does not work well if there is enough overlapping or touching of characters.

3.2 Non fuzzy linear approach

Another approach for handwritten segmentation in Devanagari script is based on nonlinear fuzzy approach [4]. It is basically focused on improving accuracy in segmentation by overcoming some of the key challenges of handwritten Devanagari word image segmentation technique. The researchers developed a new feature based approach for identifying the matras in word. They designed a non-linear fuzzy membership functions for identifying segmentation points on the matra. The segmentation accuracy achieved by them on 300 word data is 94.8%. The improvement in nonlinear fuzzy approach is about 1.8% over the triangular membership function technique which they referred before using their approach on 300 word data set. They also referred the use of horizontalness and verticalness features of previous technique before further refining. Ultimately reduced high complexity in segmentation by increasing accuracy.

IV. SEGMENTATION FOR HANDWRITTEN TEXT IN GUJARATI SCRIPT

4.1 Difficulties for segmentation in Gujarati script as compared to Devanagari script

The structure of the Gujarati script is quite challenging and the lexicons of the language puts a difficulty during the initial phase of segmentation. “Shirorekha” is not present over the characters to distinguish between words in Gujarati. Characters with diacritics is considered as a whole composite characters. Some characters contains the gap within them. Thus to overcome from the difficulty in Gujarati script segmentation, we need new algorithms for segmentation [5].

4.2 Existing proposed algorithm

Now we will discuss the segmentation in Gujarati script [5]. This research mainly deals with the recognition of Gujarati handwritten characters.

Before proposing a new algorithm the researchers carried out a comparative analysis of different algorithms from different domains [5].

The researchers proposed a new combination of structured and statistical methods to extract feature vectors and achieve good amount of accuracy. And the extracted features produced from proposed algorithm implementation is supplied as input into SVM (support vector machine). SVM gives good performance on data sets with many attributes and it has the capability to handle large number of classes.

The HCR architecture was proposed by them for segmentation. Still the difficulty is more in Gujarati script for segmentation. To tackle from these overcoming they applied morphological operation of dilation to dilate the character for obtaining connected strokes. This resulted into perfectly continuous character which can now be segmented using boundary box. But again this approach was not successful in all cases. Sometimes it changed the original character to a different character. To account from this drawback, they proposed a new algorithm for segmentation. In this proposed algorithm, each scanned sheet is subdivided into equal size of grid boxes where each grid is obtained by its height and width using ‘imtool’ command in Matlab. They proposed two algorithms in which the first algorithm is taking scanned image as input and producing various grid boxes of

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

one image. While in second algorithm the grid image is taken as input and the segmented image is obtained as output.

V. CONCLUSION

In this paper we reviewed various approaches used for segmentation of printed and handwritten text document's image for Devanagari and Gujarati script. Based on the review we conclude that the proposed approaches are increasing accuracy but still a lot of work can be done in this area for especially Gujarati script as the accuracy is reducing marginally in segmentation from line followed by word to character in all approaches.

REFERENCES

[1.] Er. Neetu Bhatia, "Optical Character Recognition Techniques: A Review" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 5, May 2014 ISSN: 2277 128X.

[2.] Vikas J Dongri, Vijay H Mankar, Devanagari document segmentation using histogram approach, International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.1, No.3, August 2011

[3.] Garg, Rahul, and Naresh Kumar Garg. "An algorithm for Text Line Segmentation in Handwritten Skewed and Overlapped Devanagari Script." International Journal of Emerging Trends in Engineering and Development 4.5 (2014): 114-118.

[4.] Sarkar, Ram, et al. "Handwritten Devanagari Script Segmentation: A non-linear Fuzzy Approach." arXiv preprint arXiv:1501.05472 (2015).

[5.] Macwan, Swital J., and Archana N. Vyas. "Classification of offline gujarati handwritten characters." Advances in Computing, Communications and Informatics (ICACCI), 2015 International Conference on. IEEE, 2015.