HandWritten Word Segmentation Upto Character Level

Nidhish Kumar Reddy
Roll No-ICM2016006
icm2016006@iiita.ac.in

Anand Singh
Roll No-IRM2016007
irm2016007@iiita.ac.in

November 27, 2019

Abstract

The objective of this report is to Perform the segmentation up-to character level. Input is handwritten word "drawn".

Keywords: *Character Segmentation, ligatures, Potential Segmentation Columns, Skeletonization,logical space indexes.*

1 Introduction

A character is the smallest unit of any language script and the segmentation of characters is the most crucial step for any **OCR (Optical Character Recognition)** System. This system translates scanned or printed image of the document into a text document that can be edited. The selection of segmentation algorithm being used is the key factor in deciding the accuracy of the OCR system. If there is a good segmentation of characters, the recognition accuracy will also be high. Segmentation of words into characters becomes very difficult due to the cursive and unconstrained nature of the handwritten script.

2 Motivation

Character segmentation has been an active field of research for many years. It still remains an open problem in the field of pattern recognition and image processing. There are mainly two phases of a character recognition system namely preprocessing, segmentation. Preprocessing aims at eliminating the variability that is inherent in printed words. The role of segmentation is to find correct letter boundaries. Segmentation of off-line cursive words into characters is one of the most difficult and important process in handwriting recognition as it directly affects the result of recognition process.

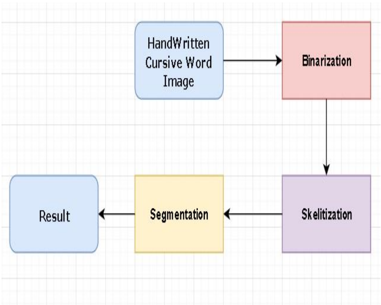

Figure 1: FLOW CHART

3 Algorithm Design

3.1 BackGround

There are two types of characters in English language. First type of characters are called **"Closed Characters"** and contain a loop or a semi-loop such as **'a', 'b', 'c', 'd', 'e', 'g', 'o', 'p', 's'** etc. Second type of characters are termed as **"Open Characters"** and are without a loop or a semi-loop e.g. **'u', 'v', 'w', 'm', 'n', 'i'** etc. In case of open characters, it is very difficult to differentiate between ligatures and characters because of the cursive nature of handwriting. In case of cursive handwritten words, a ligature is a link (small foreground component) which is present between two successive characters to join them. Two consecutive 'i' characters may give an illusion of the presence of a character 'u' and vise versa. Two consecutive characters 'n' and 'i' may look like 'm'. Also, handwritten character 'w' may look like the presence of two consecutive characters 'i' and 'v'.

3.2 Methodology

After the preprocessing of the input handwritten word image, the height and width of the word image is calculated for the analysis of the ligatures in an accurate manner. The word image is scanned vertically, from top to bottom, column wise and the number of foreground pixels in the inverted word image are counted in each column. The positions of all these columns are saved for which the sum of foreground black pixels is either 0 or 1. All these identified columns are termed as PSC (Potential Segmentation Columns).

3.3 Preprocessing

3.3.1 Step:-1

In the first step, the input word image is preprocessed by using various preprocessing techniques such as binarization, thinning, and cropping.This preprocessed word image is taken as input image to be segmented into characters

3.3.2 Step:-2

Thinning is an image morphological operation in which selected foreground pixels are removed by eroding an image until it becomes one-pixel wide. It produces a skeleton of the object present in the image and makes it easier to recognize the object such as character.

3.3.3 Step:-3

To minimize the computation complexity, the input word image is inverted for further processing. By complementing the input binary image, white pixels become the foreground pixels and the black pixels become the background pixels. Hence, it becomes easier to count the foreground white pixels represented by 1, in each column of the word image

3.3.4 Step:-4

This image is now converted from binary format to a RGB format, it becomes computationally easier to display the PSC (Potential Segmentation Columns) in different color other than black and white.

3.4 Segmentation of Handwritten Cursive Word

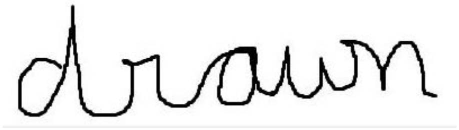

Figure 2: Input

Algorithm 1 Segmentation

Input = Output from PreProcessing

Output = Segmented upto Character level Handwritten Word Image

```
procedure SEGMENTATION(Input)
Sum all rows of each column of the input
if Sum of a particular column is greater than 1 then,
    assign 0 to it.
end if
if Sum is less than or equal 1 then,
    assign 1 to it.
end if
Obtained New Sum is assigned as LogicalSpaceIndexes(LSI)
Obtain Number of Rows and columns of Input
Since Image is in RGB,$columns \leftarrow columns/3$
for each row of Input do
    $prev \leftarrow 0$
    $prevmid \leftarrow 0$
    $flag \leftarrow 0$
    for each column of Input do
        if LSI of that column is 1 and flag==0 then
            $flag \leftarrow 1$
            $prev \leftarrow j$
        end if
        if LSI of that column is 0 and flag==1 then
            $flag \leftarrow 0$
            Get the mid of ligature
            if Gap b/w prevPSC  currPSC > 60 then
                if currPSC is mid of ligrature > 8 then
                    Draw the currPSC
                end if
            end if
        end if
    end for
end for=0
```

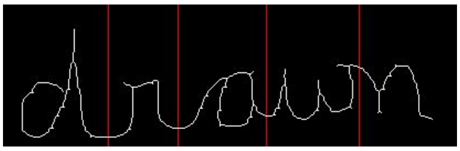

Figure 3: Output

4 Analysis

4.1 Time Complexity Analysis

In the segmentation algorithm since we are traversing each row and column of the input image .So the time complexity is

$$F(n) \propto x * y/3$$

.Since x is the number of rows in the image and y is the number of columns in the image.Since it is RGB image we divide 3 to the number of columns.

5 Discussion

To solve this problem we have divided it into 2 Stages, which performs pre-processing and segmentation of given handwritten cursive word.

6 Applications

Off-Line handwriting segmentation and recognition has been a challenging and exciting area of research for many years. The popularity of this field of research is mainly due to the unconstrained and cursive nature of human handwriting. The segmentation and recognition of such type of handwritten script is still an open problem and is an active area of research these days. The character recognition accuracy of an OCR system can be improved remarkably if the characters within a word are correctly isolated. Hence, segmentation is the most crucial step in the off-line cursive handwritten script recognition process

7 Conclusion

A new vertical segmentation technique is developed to enhance the oversegmentation of the handwritten word image by thinning the word image to a single pixel width. The objective of the proposed approach is to oversegment the handwritten word image sufficient number of times to ensure that all possible character boundaries have been dissected.

References

[1] . N. Nain and S. Panwar, Handwritten text recognition system based on neural network, Academy Publish Journal of Computer and Information Technology, Vol.2, No. 2, Pages 88-97, 2012.

[2] G. Louloudis, B. Gatos, I. Pratikakis, and C. Halatsis , Text Line and Word Segmentation of Handwritten Documents, Pattern Recognition, Vol.41, No. 12, Pages 3758-3772, 2008.

[3] . S. Mathur, V. Aggarwal, H. Joshi, A. Ahlawat, Off-Line Handwriting Recognition using Genetic Algorithm, Proc. Of Sixth International Conference on Information Research and Applications – i.Tech, Varna, Bulgaria, 2008.

[4] A. Dutta, J. Llados, and U. Pal, Bag-of-GraphPaths Descriptors for Symbol Recognition and Spotting in Line Drawings, Proc. Of 9th international conference on Graphics Recognition: new trends and challenges, 2011.