A REVIEW PAPER ON TOUCHING CHARACTER IDENTIFICATION AND SEGMENTATION

Himani Gupta
M.tech cse, Galgotias University

Shivani Varshney
M.tech cse, Galgotias University

ABSTRACT
Optical Character Recognition (OCR) is a system which is use for translating scanned documents into text. OCR follows a process that includes six steps. One of them is segmentation. Touching characters will cause errors in the result of segmentation step. This will lead to more errors in next step of OCR. Touching characters affect the performance of OCR. So various approaches are defined for touching character identification and segmentation. Each approach has some advantages and some disadvantages. If one approach give best result for a language then it is not necessary that this approach will also give best for another language. In this paper, touching characters are explained and various approaches which have been developed for touching character identification and segmentation are also explained in brief.

INTRODUCTION
When two or more characters are connected or we can say touched to each other then these characters are called touching characters. Touching characters may be present in any language. Touching characters have an effect on the performance of OCR. There are various applications, where OCR plays an important role. Incorrect segmentation of touching characters causes errors which produces inaccurate result of OCR. There are following some examples of touching characters:

![Figure 1: example of touching characters](image)

Himani, Shivani
Copyright@IJETMAB
Page 25
When the feature vector of these sub images are taken as input for classifier, the classifier compares the feature vector of this sub image with the feature vector of training data, so the character will not be recognize correctly. So the text file, which is produced by OCR will not have same character as in the image, this means that there is no meaning of OCR. So it is very important to identify touching characters and then segment them to get the improved results.

There are various approaches, have been developed to reduce this problem or to identify touching characters and to segment them. But only one approach is not useful for all type of touching characters or for different languages. So different approaches are developed for different languages.

**APPROACHES FOR SEGMENTATION OF TOUCHING CHARACTER**

An OCR system is developed to convert the scanned documents into text form. Various methods are use in each step of OCR. In many papers, various techniques for Touching Character Identification and Segmentation are explained.

1) A character recognition method on ID card using Template Matching Approach is also proposed in which text is detected on Indonesia ID card consist of some phases, which are pre-processing, extract the text regions, segmentation step and recognize the segmented area. In this paper, segmentation is done horizontally and vertically. Horizontally means by dividing line and vertically means by dividing characters in an image. And segmentation is done by using histogram. In this paper, author is using two template matching algorithms which are 3*3 algorithm and pixel-by-pixel algorithm. In the segmentation stage, approximately 93% of character can be cut off correctly [1].

2) A method was proposed as a segmentation method to separate between touching and overlapping connected components in Handwritten Arabic documents. This method can be applied for the connections between text-lines as well as between words. Besides, it can be used for the three classes of connected component connections. It is based on a recognition stage in which they select a similar segmented model and estimate the transformation aligning the two connected components. Then they use
the central points of the Model’s transformed parts to segment the input connected component. This increases performance and robustness of the proposed system [2].

3) An approach is proposed, for machine printed Gurumukhi documents, works for segmentation of pair as well as triplet of touching characters [3]. Design process for this approach is as follows:

![Flowchart of proposed approach](image)

4) A scheme is present towards the segmentation of English multi-oriented touching strings into individual characters. When two or more characters touch, they generate a big cavity region at the background portion. Using Convex Hull information, these background information is used to find some initial points to segment a touching...
string into possible primitive segments (a primitive segment consists of a single character or a part of a character). Next these primitive segments are merged to get optimum segmentation and dynamic programming is applied using total likelihood of characters as the objective function. SVM classifier is used to find the likelihood of a character. To consider multi-oriented touching strings the features used in the SVM are invariant to character orientation. Circular ring and convex hull ring based approach has been used along with angular information of the contour pixels of the character to make the feature rotation invariant [4].

5) A technique for the detection and the segmentation of touching characters in mathematical expressions was presented. In its detection stage, a connected component is selected as a candidate of touched characters if it’s several feature values deviate from the standard feature values of its initially recognized category. In the segmentation stage, two component characters of the candidate are decided by the comparison with touching character images synthesized from two single character images [5].

OUR PROPOSED ALGORITHM TO IDENTIFY AND SEGMENT TWO HINDI TOUCHING CHARACTERS

After reading various papers on touching characters, I have decided to work on Hindi touching characters. And I will use a step by step process to identify and segment Hindi touching characters. This process is:

1. Calculate width of each character of Hindi language.
2. Find maximum and minimum width.
3. Check
   If (size of segmented image > maximum width)
   Then
   This image contain touching characters. /* identification of touching characters.
   Else go to step 9.
4. Scan this image horizontally.
5. Remove the row which contain maximum occurrence of black pixels. (Remove header line.)
6. If size of image = 2*maximum width OR size of image = 2*minimum width
Then image will be segment from half of its size. And go to step 9.

7. Scan image vertically up to size of image is equal to minimum width from starting of image (say x1) and also from end of image (say x2).
8. Find list of top valleys and list of bottom valleys between x1 and x2. And segmentation path is the line joining a pair of top and bottom valleys.
9. Stop.

Assumption in this algorithm:

All Characters have same font size.

CONCLUSION

Optical character recognition (OCR) performs many stages. There may be various problems during execution of stages that can affect the performance of recognition system. One of them is touching character identification and segmentation problem. Performance of system can be improve by minimizing this problem. For minimizing this problem, first we have to identify touching characters and then segment them. So the degradation in inputs for recognition stage will minimize. And we know lesser degraded inputs minimize the error rate. Finally the efficiency of the system will increase.

REFERENCES


[5] Akihiro Nomura, Kazuyuki Michishita, Seiichi Uchida, and Masakazu Suzuki Graduate School of Mathematics, Faculty of Information Science and Electrical


