Segmentation of Touching Charcaters for Printed Hindi Documents

Alok Kumar
Computer Science Department
Center for Development of Advanced Computing
Noida, India
alokhooda@gmail.com

Madhuri Yadav
Computer Science Department
Center for Development of Advanced Computing
Noida, India
madhuri26yadav@gmail.com

Abstract— **In optical character recognition system (OCR) character segmentation remains a challenge due to presence of touching characters. This paper introduces a new technique for identification and segmentation of touching characters in printed Hindi documents which is based on matching of the profiles of the isolated characters and the touching characters. The touching characters are identified from the printed documents based on their aspect ratio as they have higher aspect ratio due to increase in their width, in comparison to isolated characters. After identification of touching characters, the left and lower profiles of the touching characters are matched with the isolated character's profiles and the cut point for segmentation is decided. The experimental results show that the proposed approach is font independent and can segment a pair as well as triplet of touching characters.**

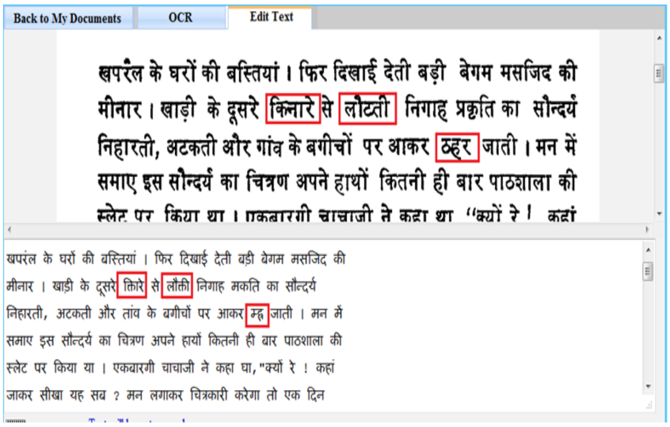
Keywords—OCR; Aspect Ratio;

I. Introduction

Optical Character Recognition (OCR) is the process of converting scanned images of machine printed or handwritten text into a computer process-able format. The traditional scanned documents cannot be processed they exist merely as image files.

In this emerging world, OCR has turned out be an outstanding technology which can provide automation to office works, forms and bank check processing, Document reader systems for the visually impaired, Database and corpus development for language modeling, text-mining and information retrieval [1].The accuracy of OCR system depends upon the number of characters correctly recognized by it. Thus, Character recognition is an important pre-processing step for text recognition. But the characters may not be correctly recognized due touching characters.

This paper focuses on the problem of touching characters, which are mostly found in the historic or old documents of ancient times. There is a need to computerize such documents because they are degrading due to environmental factors and poor paper quality. In such documents, existence of touching characters can lower the recognition rate drastically for any OCR system. Depending upon the scripts for which OCR system is generated classes can be created for each recognized character but when touching characters appear they cannot be recognized and thus arbitrarily classified by OCR resulting in higher error rate and degraded accuracy. The figure 1 explains the above problem and provides the motivation to work in this field:



खपरैल के घरों की बस्तियां। फिर दिखाई देती बड़ी बेगम मसजिद की
मीनार। खाड़ी के दूसरे किनारे से लौटती निगाह प्रकृति का सौन्दर्य
निहारती, अटकती और गांव के बगीचों पर आकर ठहर जाती। मन में
समाए इस सौन्दर्य का चित्रण अपने हाथों कितनी ही बार पाठशाला की
स्लेट पर किया था। एकबारगी चाचाजी ने कहा था "क्यों रे ! कहां

खपरंल के घरों की बस्तियां । फिर दिखाई देती बड़ी बेगम मसजिद की
मीनार । खाड़ी के दूसरे क्तिारे से लौक्ती निगाह मकति का सौन्दर्य
निहारती, अटकती और तांव के बगीचों पर आकर ग्व्ह जाती । मन में
समाए इस सौन्दर्य का चित्रण अपने हायों कितनी ही बार पाठशाला की
स्लेट पर किया या । एकबारगी चाचाजी ने कहा घा,"क्यों रे ! कहां
जाकर सीखा यह सब ? मन लगाकर चित्रकारी करेगा तो एक दिन

Figure 1: The above figure show the wrong recognition of touching characters by web OCR.

II. TOUCHING CHARACTERS

Touching characters are the major problem in character recognition and these patterns emerge when two adjacent characters are written too close; therefore, some parts of character are connected horizontally.

In Hindi language words have three zones namely: upper zone, middle zone, and lower zone. The figure 2 illustrates the zones:

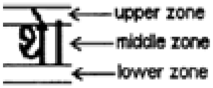


Figure 2: Zones of Hindi words

The figure 3 shows the examples of touching character.

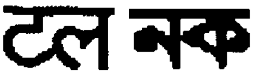


Figure 3: Examples of Touching Characters

Touching characters in the documents may be generated due to any of the following reasons:

- Too little spacing between neighboring characters,
- Excessive ink dispersion,
- Streaks introduced during photocopying that cross over many characters
- Poor printing technology
- Inferior paper quality

III. Proposed Technique

The technique works in two passes: In first pass the sample images are preprocessed to avoid unwanted distortions and noise. These preprocessed images are handled to identify the touching characters. In second pass the touching characters are segmented by using profile matching technique. The flowchart in figure 4 describes the proposed technique.

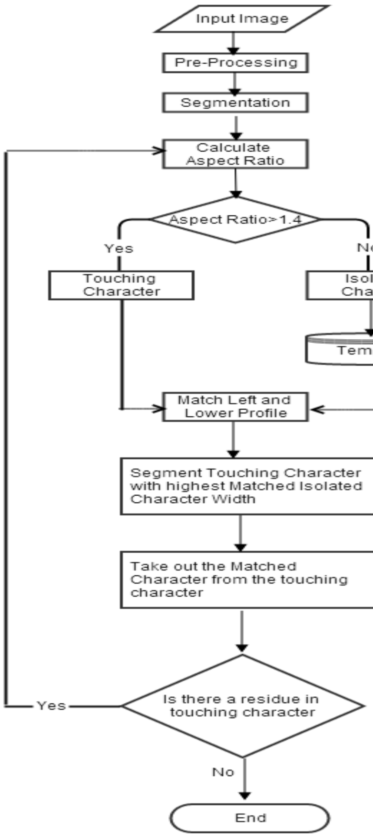

Input Image
Pre-Processing
Segmentation
Calculate Aspect Ratio
Aspect Ratio>1.4
Yes
No
Touching Character
Isolated Character
Templates
Match Left and Lower Profile
Segment Touching Character with highest Matched Isolated Character Width
Take out the Matched Character from the touching character
Is there a residue in touching character
Yes
No
End

Figure 4: Flowchart of Proposed Technique

First Pass: Identifying Touching Characters

A. Pre-processing

This stage is collection of operations that apply successive transformations on an image. It takes in a raw image, reduces noise and distortion, removes skew and performs skeletonization of the image there by simplifying the processing of the rest of the stages.

B. Convert Color Space

Images scanned will be either in raw format or encoded into some multimedia standards. Normally, these images will be in RGB mode, with three channels. Number of channels defines the amount of color information available in the image. There may be some irrelevant information which we do not require so to simplify our work and complexity we reduce this extra information and convert images to grayscale.

C. Binarization

It is the process where each pixel in the image has only two bit values in our case it is either 0 or 1. It is done to convert grayscale image into black and white image. Now the images will have only two intensity values.

D. Line Segmentation

The lines of text block are segmented by finding the valleys of the projection profile computed by counting the number of black pixels in each row. A text line can be found between two consecutive boundary lines as indicated in the figure 5. We have assumed that the text block contain only single column of text.

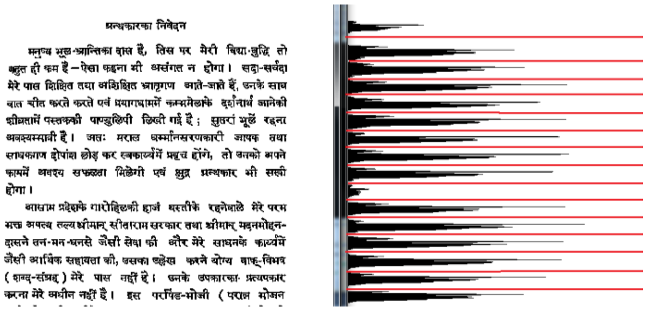
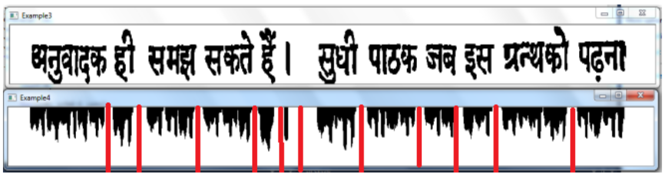


Figure 5: Projection profile of rows in Hindi text (red line shows line boundaries).

E. Word Segmentation



अनुवादक ही समझ सकते हैं। सुधी पाठक जब इस ग्रन्थको पढ़ना

Figure 6: Vertical projection of text line

After line segmentation, we use vertical pixel projection profile to find individual words. Each column is scanned vertically and counts the total number of black pixels. When a word starts there is a transition from white to black and it ends with a transition from black to white. In this way we obtain words of a line. The figure 6 shows the vertical projection of text line.

F. Character Segmentation

After word segmentation we perform character segmentation. Since characters in the word are connected to each other through the headline, they get disconnected once the headline is removed.

For headline removal, we find the location of headline by using horizontal projection. As headline has the maximum number of black pixels so after horizontal projection the row which has the highest number of black pixels is the headline and then we remove it. Figure 7 shows the segmentation of a character from a word.

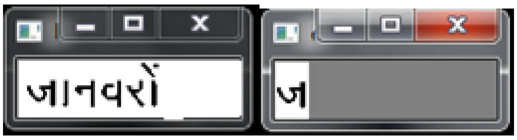


Figure 7: Vertical projection of text line

G. Identication of Touching Character

After segmentation is complete, the next step is the identification of touching characters from the output of character segmentation module. It contains both touching and non-touching characters. To differentiate between them aspect ratio of the Hindi characters was computed.

As shown in figure 8 Aspect ratio is defined as ratio of width of the image to the height of the image.

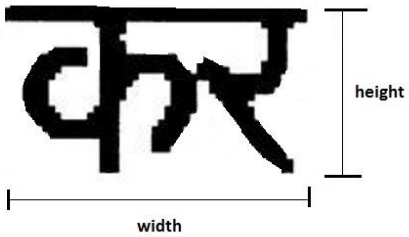


Figure 8: Example of Aspect ratio

It was observed that the width of touching character is more than its height, so the aspect ratio is more for touching characters than single characters. The aspect ratio of each Hindi character was evaluated and it was observed that the neither of the isolated character exceeded the aspect ratio of 1.4 except ख , ज्ञ characters . Thus, a threshold value of 1.4 was used for segmenting isolated characters and touching characters, if aspect ratio of any character is greater than threshold then it is considered as touching character otherwise it is an isolated character.

Second Pass: Segmentation of Identified Touching Characters

A. Profiles

For segmentation of touching characters we computed left and lower profile for every pattern. Only these two profiles are considered because the right profile may change due to presence of touching character.

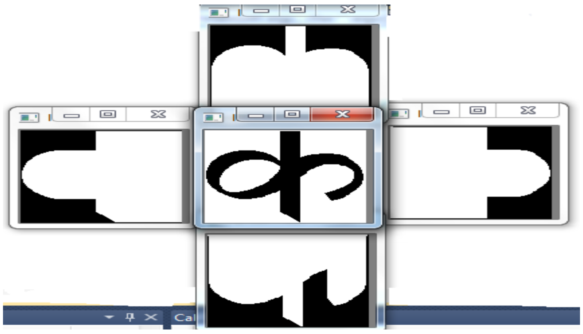


Figure 9: Four side Profiles of Isolated character

The four side profile of a character enclosed within a bounding box are obtained by counting the white pixels in the four directions rightward, leftward, upward, downward respectively until a black pixel is encountered, Fig. 9 shows the four directional profiles of the character.

The merged part of touching character generates different shape of patterns from the primitive character patterns. However, far left side patterns in the touching character will not be affected by touching. Hence we have used left and lower profile for segmenting the touching character. The Fig. 10 shows the left and right profiles of a character which remains unaffected even after the existence of touching character.

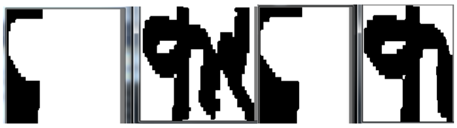


Figure 10: Left profile of touching character and isolated character are same

B. Profiles Matching and Cut Position Detection

In profile matching we match the left and lower profile of touching characters with that of the isolated characters.

The left profile of the isolated and touching characters are stored in different arrays. Then the match score is calculated according to following equation:

For left profile:

ms=ms+1

iff

$$\sum_{1=0}^{k} IC[i] < TC[i] + 2 \quad \&\& \quad \sum_{1=0}^{k} IC[i] > TC[i] - 2$$

Where,

ms= match score

IC= Integer type array for left profile of Isolated characters
TC= Integer type array for left profile of touching characters
k=height of Isolated character

For lower profile:

ms=ms+1
iff

$$\sum_{i=0}^{J} IC[i] < TC[i] + 2 \quad \&\& \quad \sum_{i=0}^{J} IC[i] > TC[i] - 2$$

Where,
ms= match score
IC= Integer type array for lower profile of Isolated characters
TC= Integer type array for lower profile of touching characters
J= width of the isolated character

The touching character at the width of highest match isolated character as shown in figure 11.

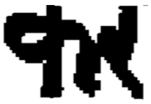
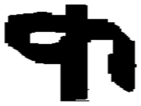
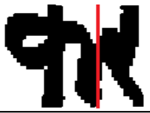
	**Touching character**
	**Matched isolated character**
	**Red line shows the cut position**

Figure 11: Cut Position Detection

IV. Results and Discussions

To evaluate the quantitative performance of the proposed algorithm, it was executed on various documents. It has been found that after applying the proposed approach on degraded documents 92-96% of the touching characters are easily identified.

References

[1] Mr.Nithya.E Dr. Ramesh Babu D R Volume 3, Issue 6, June 2013 ISSN: 2277 International Journal of Advanced Research in Computer Science and Software Engineering page 102-105.

[2] G. Congedo, G. Dimauro, S. Impedovo, G. Pirlo, "Segmentation of Numeric Strings", 1995, IEEE, p.p- 1038-1041

[3] Utpal Garain and B. B. Chaudhuri , "Segmentation of Touching Symbols for OCR of Printed Mathematical Expressions: An Approach based on Multifactorial Analysis", Proceedings of the 2005 Eight International Conference on Document Analysis and Recognition (ICDAR'05), IEEE , 2005.

[4] Dong-Yu Zhang, Xue-Dong Tian, and Xin-Fu Li "An Improved Method for Segmentation of Touching Symbols in Printed Mathematical Expressions". 2nd International Conference on Advanced Computer Control (ICACC), vol.2, pp. 251 - 253, 2010.

[5] U.K.S. Jayarathna G.E.M.D.C. Bandara, "A Junction Based Segmentation Algorithm for Offline Handwritten Connected Character Segmentation", International Conference on Computational Intelligence for Modelling Control and Automation, IEEE, 2006.

[6] K. B. M. R. Batuwita, G.E.M.D.C. Bandara, "An online Adaptable fuzzy system for offline handwritten Character recognition", Proceedings of 11th World Congress of International Fuzzy Systems Association (IFSA 2005), Beijing ,China, Springer-Tsinghua, 2005, Vol. II, p.1185-1190.

[7] Salman Amin Khan "Character Segmentation Heuristics for Check Amount Verification".

[8] U. Garain and B.B. Chaudhuri. "Segmentation of Touching Characters in Printed Devnagari and Bangla Scripts Using Fuzzy Multifactorial Analysis." IEEE Transaction on Systems, Man and Cybernetics,Vol. 32(4), 449-459, 2002.

[9] U. Pa1, A. Belai and C. Choisy, "Water Reservoir Based Approach for Touching Numeral Segmentation".2001

[10] George Nagy, Thomas A. Nartker, Stephen V. Rice, Optical Character Recognition: "An illustrated guide to the frontier", Procs. Document Recognition and Retrieval VII, SPIE Vol. 3967, 58-69.

[11] Y. Lu, "Machine Printed Character Segmentation – an Overview", Pattern Recognition, vol. 29, no. 1, pp. 67-80, 1995

[12] S.Kahan, T.Pavlidis, and H.S.Baird, " on the recognition of printed characters of any fonts and sizes", IEEE Trans. Pattern Analysis andMachine Intelligence, vol. 9, no. 2, pp. 274-288, Mar. 1987

[13] S. Tsujimoto and H. Asada, " Resolving Ambiguity in Segmenting Touching Characters" Ist Int. Conf. on Document Analysis and Recognition ,pp. 701-709, Saint-Malo, France, Oct 1991. R.G.Casey and G. Nagy, "Recursive Segmentation and Classification of Composite character Patterns", Proc. 6th Int. Conf. on Pattern Recognition, pp. 1023-1026, Munich, germany, 1982.

[14] Veena Bansal and R.M.K. Sinha , "Segmentation of touching and Fused Devnagari characters, ", Pattern recognition, vol. 35, pp. 875-893, 2002.

[15] U. Garain, B.B. Chaudhuri, "Segmentation of touching characters in printed Devnagari and Bangla scripts using fuzzy multifactorial analysis", IEEE Trans. Systems Man Cybern. Part C-32 (2002) 449–459.

[16] B.B. Chaudhuri, U. Pal and M. Mitra, "Automatic Recognition of Printed Oriya Script", *ICDAR*, pp.795-799,2001.

[17] U. Garain, B.B. Chaudhuri, "On recognition of touching characters in printed Bangla Documents", Proceedings of the Fourth International Conference on Document Analysis and Recognition, 1997, pp. 1011–1016.

[18] M. K. Jindal, G.S. Lehal and R.K. Sharma," A Study of Touching Characters in degraded Gurmukhi Script", in Int. Conf. on Pattern Recognition and Computer Vision, PRCV 2005, pp. ?, 25-27 February 2005, Istanbul, Turkey.

[19] G. S .Lehal and Chandan Singh, "Text segmentation of machine printed Gurmukhi script", Document Recognition and Retrieval VIII, Proceedings SPIE, USA, vol. 4307, pp. 223-231, 2001.

[20] G. S. Lehal and Chandan Singh, "A technique for segmentation of Gurmukhi script", Computer Analysis of Images and Patterns, Proceedings CAIP 2001, W. Skarbek (Ed.), Lecture Notes in Computer Science, vol. 2124, Springer-Verlag, Germany, pp. 191-200, 2001