

Segmentation of Handwritten Hindi Text

Naresh Kumar Garg

GZS Collage of Engineering & Tech.
Bathinda, Punjab, India

Lakhwinder Kaur

Department of Computer
Engineering, UCOE, Punjabi
University, Patiala, Punjab, India

M. K. Jindal

Panjab University Regional Centre,
Muktsar, Punjab, India

ABSTRACT

The main purpose of this paper is to provide the new segmentation technique based on structure approach for Handwritten Hindi text. Segmentation is one of the major stages of character recognition. The handwritten text is separated into lines, lines into words and words into characters. The errors in segmentation propagate to recognition. The performance is evaluated on handwritten data of 1380 words of 200 lines written by 15 different writers. The overall results of segmentation are very promising.

Keywords

Segmentation, line segmentation, word segmentation, character segmentation, lower modifier, upper modifier, Header line, Base line.

1. INTRODUCTION

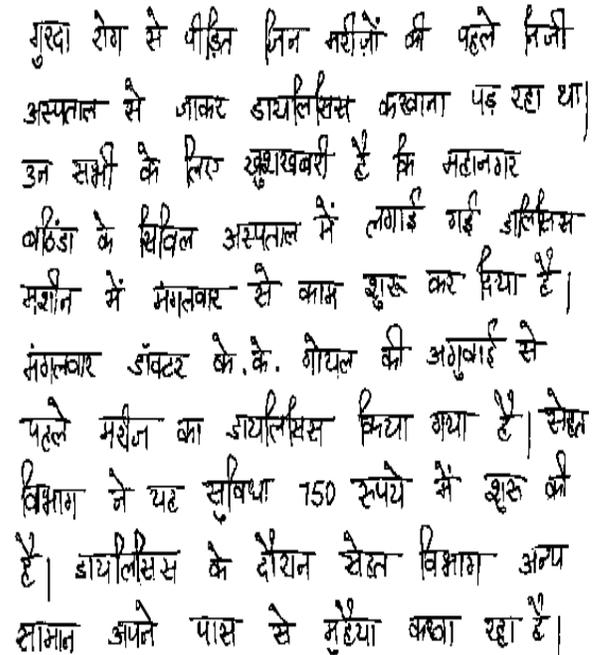
Handwritten character recognition is an important field of Optical Character Recognition (OCR). The recognition of handwritten text in scripts is one of the major areas of research. A good survey about OCR is given in [1]. The recognition of Indian scripts is gaining much attention now days. Hindi being the official language of India yet a few research reports available on it. Devanagari is the most popular script in India. Devanagari is the script for writing Hindi language. Hindi is written from left to right. The first research report on Handwritten Devanagari character was published in 1977 [2], but not much research work was done after that. Researchers worked on isolated handwritten Hindi characters or handwritten Hindi numerals but not on complete handwritten Hindi text. Many approaches have been proposed by researchers for recognition of isolated handwritten Hindi characters or recognition of Hindi numerals. The segmentation is one of the major stages of character recognition. To the best of my knowledge this is the first paper with complete handwritten Hindi text segmentation.

A lot of research is done in the past on line segmentation of handwritten text. A wide variety of line segmentation methods for handwritten documents are reported in the literature. The various existing methods for line segmentation are categorized as projection based[3,4], Hough transform based[5], smearing[6], grouping[7], graph based[8], CTM (Cut text Minimum) approach[9], block covering[10] and linear programming. An overview of OCR research in Indian scripts is given in [11]. Bansal [12] has worked on printed Devanagari text recognition. Among some latest work, Jindal et. al. [13-16] have worked on recognition of degraded printed Gurmukhi script documents and addressed various problems faced during recognition.

The paper is organized as follows. In next Section, we have discussed the creation of database used for the experimental purposes. Section 3 includes the discussion about the characteristics of Hindi language. In Section 4, we have discussed the segmentation technique used for segmenting the handwritten Hindi text. Finally, Section 5 contains results and discussions.

2. DATABASE

All experiments are conducted on database constructed by taking handwritten data from 15 writers. Ten writers were asked to write paragraph of 10-15 lines of same text. Also five writers were asked to write different text. A healthy mix of people from various backgrounds was taken so as to make such a small database as close as possible to the real database. Data of different sizes and slants is also included in the database. No pre processing is performed on the data. Figures 1 and 2 contain part of handwritten Hindi database.



गुस्सा रोग से पीड़ित जिन मरीजों को पहले नैजी
अस्पताल से जाकर डायलिसिस करवाना पड़ रहा था।
उन सभी के लिए सुझावबरी है कि महानगर
वर्किंग के सिविल अस्पताल में लगाई गई डायलिसिस
मशीन में मंगलवार से काम शुरू कर दिया है।
मंगलवार डॉक्टर के.के. गोयल की अगुवाई से
पहले मरीज का डायलिसिस किया गया है। सेहत
विक्रम ने यह सुविधा 750 रुपये में शुरू की
है। डायलिसिस के दौरान सेहत विक्रम अग्र
शामान अपने पास से मुहैया करा रहा है।

Figure 1. Part of database.

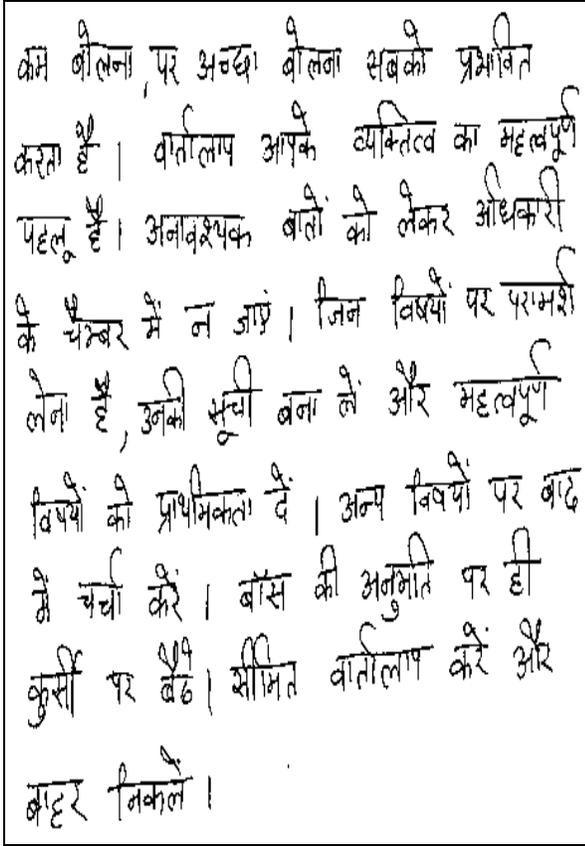


Figure 2. Part of database.

3. CHARACTERISTICS OF HINDI LANGUAGE

Devanagari is the script for writing Hindi, Nepali, Marathi and Sanskrit languages. The alphabets of Devanagari script consists of 33 consonants and 14 vowels. It is written from left to right. There is no concept of lower or upper case in Hindi language.

In Hindi language, most of the characters have a horizontal line at the upper part. In Hindi language characters also have a half form which increases the language complexity for recognition. The half characters may touch with full characters to make the characters called conjuncts. In each conjunct character, the right part is a full consonant, and the left part is always a half consonant. When two or more characters are combined to form a word, the horizontal lines touch each other and generate a header line called *shirorekha*. The vowels (modifiers) can be placed at the left, right (or both), top or bottom of the consonant. The vowels above the header line are called ascenders or upper modifiers and vowels below the consonants are called descenders or lower modifiers. Two consecutive lines touch or overlap each other due to these modifiers. This makes the segmentation of handwritten Hindi text very complex.

4. SEGMENTATION

The text segmentation is divided into three parts:

- Segmentation of lines from the text.

- Segmentation of words from the lines.
- Segmentation of characters from the words.

4.1 Line Segmentation

We proposed a line segmentation method which is based on header line detection and base line detection. We have used two-stripe projection for header and base line detection. Header line is the most visible part of the text. Detection of header line is one the most challenging tasks in skew variable or fluctuating line text. Till now most of the researchers are detecting the header line by finding the row with maximum pixel density, but it can not work for skew variable text.

We make following assumptions about the data:

- The minimum height of character consonant in a line is eight pixels. Average line height is 30 pixels.
- The skew in a text is not more than the height of a character consonant.

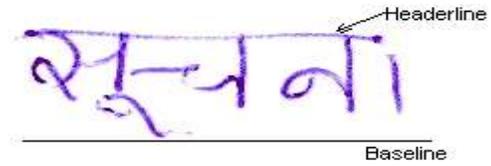


Figure 3. Header line and Base line.

The algorithm for line segmentation has following steps:

Step 1: Initially, rough estimate of the header lines in whole of the text are made by the formula

$$pcol(i) > 15 \ \& \ pcol(i) > pcol(i : i+8) \ \& \ pcol(i : i+8) > 0$$

$$pcol(i) > floor(width(i)/7)$$

where,

$pcol(i)$: No. of pixel in row i

$width(i)$: width of line i i.e difference between last pixel and first pixel position of line i .

Step 2: After finding first header line, we skip 8 rows (equal to minimum height of consonant) to find the next header line.

Step 3: From $(i+8)^{th}$ row to $(i+22)^{th}$ row, we find the m^{th} row with minimum of pixels.

Step 4: We skip the rows upto m^{th} row and goto step 1 to find the next header line.

After finding the header lines, the most challenging task is to find the base line. For finding the base line following procedure is followed:

Step 1: Two consecutive rough header lines are taken.

Step 2: The line is again divided into two equal halves (stripes).

Step 3: The rows with minimum of pixels are taken as base lines separately for each half.

Step 4: Then the lines are separated between header lines and base lines separately for each half.

Step 5: Then two separate lines are joined to get the actual text line.

This method gives good results for uniform and non uniform skewed lines.

4.2 Word Segmentation

After lines are segmented from text, words are segmented from lines by vertical projection profile. For each column of the line the number of black pixels is counted and the columns with zero black pixels are used as delimiters for word separation. To distinguish the character separation from the word separation, we have selected the delimiter as at least three continuous columns with zero black pixels for word separation.

4.3 Character Segmentation

For character separation the vertical projection method is used after header line detection. The algorithm has following steps:

Step 1: The header line is identified using the horizontal projection profile. The line with maximum number of black pixels in upper 10% part of the word is considered as the header line. Let this position be h1.

Step 2: From h1-1(if h1>2 otherwise we assumed no upper modifier present) to top row the vertical projection is made and the columns with zero black pixels is treated as delimiter for separation of ascenders (upper modifiers). This is done for whole of the word starting from first column to last column of the word.

Step 3: To separate lower modifiers, first we find the difference in heights of characters. If difference between maximum height and minimum height is at least 20% of the height, then we assume lower modifier exists otherwise not. Then from the lowest row we find three vertical black pixel crossings or two vertical black pixel crossings with two or more black pixels in second crossing in lower 20% part. Then we separate the lower modifier from the second crossing to the lowest row. We note the position of second crossing say bt1.

Step 4: From h1+1 to bt1 row of the image the vertical projection is made and the column with zero black pixels is treated as delimiter for character separator.

The above method of character segmentation shows good results.

5. RESULTS

The results of text segmentation into lines, lines into words and words into characters are given in the following tables.

Table 1. Accuracy of Text line segmentation

Total Lines	Lines correctly segmented	% of accuracy
200	183	91.5

Table 2. Accuracy of Word segmentation

Total Words	Words correctly segmented	% of accuracy
1380	1354	98.1

Table 3. Accuracy of Consonants

Total	Consonants correctly	% of accuracy
3870	3062	79.12

consonants	segmented	
3870	3062	79.12

Table 4. Accuracy of Ascenders (Upper modifiers)

Total ascenders	Ascenders correctly segmented	% of accuracy
1366	1305	95.5

Table 5. Accuracy of Descenders (lower modifiers)

Total lower modifiers	Lower modifiers correctly segmented	% of accuracy
132	109	82.6

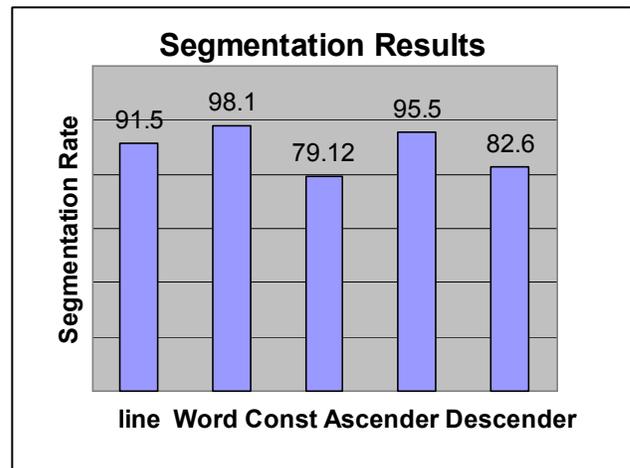


Figure 4. Segmentation results



Figure 5. Correctly Segmented Lines (Result of figure 2)

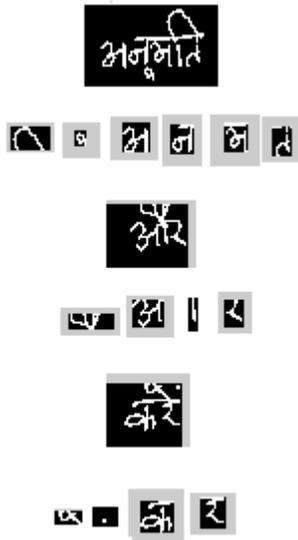


Figure 6. Correctly Segmented Word Samples

The character images, which are not correctly segmented, are the unsegmented images. Some of the error figures are shown in the figure 7.

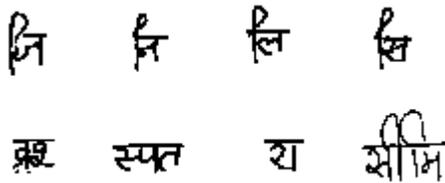


Figure 7. Unsegmented image samples.

6. DISCUSSIONS

From the above tables, it is clear that the segmentation techniques developed for line segmentation, word segmentation and character segmentation are giving good results. The segmentation problem occurs where characters touch each other (some examples shown in figure 4). The segmentation problem occurs in ascenders when two ascenders touch each other. The ascenders which touch each other are segmented as one unit instead of two separate units. The segmentation of lower modifier from consonants is done correctly. But in some cases where lower modifier is very small or not forming the loop are not correctly segmented. The maximum problem of lower modifier separation from consonants occurs in

character , due to presence of lower modifier like loop in lower part of this character.

The identification of header line affects the results. If header line and base lines are accurately identified the segmentation errors can be further reduced. The maximum accuracy occurs in word segmentation due to clear separation of words in a line or large gaps between the words. Some errors in word separation occur due to incorrect line segmentation. The errors which occurs in text line segmentation also creates problem in word segmentation and character segmentation. We have confirmed that the errors in line

segmentation propagate to character segmentation. Most of the half characters are also segmented correctly but work of proper segmentation if half characters is still in progress.

The study may be carried out in future with following direction:

- The text line segmentation technique given above does not work for large skewed lines and touching lines. So text line segmentation can be changed to improve the segmentation results.
- The segmentation of half characters is not done yet. It may be carried out in the future.
- The character separation technique explained above can be applied on other Indian scripts.

7. REFERENCES

- [1] S. Mori, C. Y. Suen and K. Yamamoto, "Historical review of OCR Research and development", *Proceedings of the IEEE*, Vol. 80(7), pp. 1029-1058, 1992.
- [2] K. Sethi, "Machine recognition of Constrained hand printed Devanagari", *Pattern Recognition*, pp.69-75, 1977.
- [3] A. Zahour, B. Taconet, P. Mercy, and S. Ramdane, "Arabic Hand-written Text-line Extraction", *Proceedings of the Sixth International Conference on Document Analysis and Recognition, ICDAR, Seattle, USA*, pp. 281–285, 2001.
- [4] N. Tripathy and U. Pal, "Handwriting Segmentation of unconstrained Oriya Text", in the proceedings of International Workshop on Frontiers in Handwriting Recognition, pp. 306–311, 2004.
- [5] G. Louloudis, B. Gatos, I. Pratikakis and K. Halatsis, "A Block Based Hough Transform Mapping for Text Line Detection in Handwritten Documents", in the proceedings of Tenth International Workshop on Frontiers in Handwriting Recognition, La Baule, pp. 515-520, 2006.
- [6] Y. Li, Y. Zheng, D. Doermann, and S. Jaeger, "A new algorithm for detecting text line in handwritten document", in the proceedings of International Workshop on Frontiers in Handwriting Recognition, pp. 35–40, 2006.
- [7] L. Likforman-Sulem and C. Faure, "Extracting text lines in handwritten documents by perceptual grouping", *Advances in handwriting and drawing : a multidisciplinary approach*, C. Faure, P. Keuss, G. Lorette and A. Winter Eds, Europia, Paris, pp. 117-135, 1994.
- [8] I.S.I. Abuhaiba, S. Datta and M. J. J. Holt, "Line Extraction and Stroke Ordering of Text Pages", in the Proceedings of Third International Conference on Document Analysis and Recognition, Montreal, Canada, pp. 390-393, 1995.
- [9] C. Welwitage, A. L. Harvey and A. B. Jennings, "Handwritten Document Offline Text Line Segmentation", in the Proceedings of Digital Imaging Computing: Techniques and Applications, pp. 184-187, 2005.
- [10] A. Zahour, B. Taconet, L. Likforman-Sulem and Wafa Bousellaa, "Overlapping and multi-touching text-line segmentation by Block Covering analysis", *Pattern analysis and applications*, 2008.
- [11] B. A. Srinivas, A. Agarwal, and C. R. Rao, "An overview of OCR research in Indian Scripts", *International Journal*

- of *Computer Science and Engineering Systems*, pp.141-153, 2008.
- [12] Veena Bansal, “Integrating knowledge sources in Devanagari text recognition”, Ph.D. thesis, IIT Kanpur, INDIA, 1999.
- [13] M. K. Jindal, G. S. Lehal and R. K. Sharma, “On Segmentation of touching characters and overlapping lines in degraded printed Gurmukhi script”, *International Journal of Image and Graphics (IJIG)*, World Scientific Publishing Company, Vol. 9, No. 3, pp. 321-353, 2009.
- [14] M. K. Jindal, R. K. Sharma and G. S. Lehal, “Segmentation of Horizontally Overlapping Lines in Printed Indian Scripts”, *International Journal of Computational Intelligence Research (IJCIR)*, Research India Publications, Vol. 3, No. 4, pp. 277-286, 2007.
- [15] M. K. Jindal, R. K. Sharma and G. S. Lehal, “Segmentation of Touching Characters in Upper Zone in printed Gurmukhi Script”, in Proceedings of 2nd Bangalore Annual Compute Conference on 2nd Bangalore Annual Compute Conference, (Bangalore, India, January 09 - 10, 2009). COMPUTE '09. ACM, New York, NY, 1-6.
- [16] M. K. Jindal, R. K. Sharma and G. S. Lehal, “Structural Features for Recognizing Degraded Printed Gurmukhi Script”, in Proceedings of the IEEE 5th International Conference on Information Technology: New Generations (ITNG 2008), pp. 668-673, April 2008.