A New Technique for Line Segmentation of Handwritten Hindi Text

Saiprakash Palakollu[1] Renu Dhir[1] Rajneesh Rani[1]

[1]NIT Jalandhar, Punjab- India

ABSTRACT
This paper mainly deals with the new method for line segmentation of Handwritten Hindi text. Line segmentation is the major task in segmentation process. If line segmentation fails word segmentation and character segmentation fails automatically. In this paper we have discussed an algorithm to detect header line and base lines accurately so that we can divide the lines correctly. The average line height is estimated, before calculating the header line and base lines. Performance of this method is checked on the 500 lines of different handwritten documents.

Key Words
Text line segmentation, projection, average line height, header line, base line detection.

1. INTRODUCTION
Handwritten character recognition is difficult task compared to machine printed character recognition in the area of Optical Character Recognition. Devanagari is a script for Hindi text. A lot of research is done on the printed Hindi text, but less work has been done on the handwritten Hindi text recognition. It has three major steps such as preprocessing, segmentation and recognition. Segmentation is the important step. Segmentation also contains three major steps such as line segmentation, word segmentation and character segmentation. If we fail in doing line segmentation then entire segmentation process goes wrong. A lot of research has been 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. Some of the methods for line segmentation are projection based method [1] [2], Hough transform based method [3], smearing method [4], grouping method [5], graph based method [6], CTM (Cut text Minimum) approach [7], Block covering method [8] and linear programming method [9].

The projection based methods are successful in the case of straight and easily separable lines only. A lot of research is going on how to detect overlapped lines and characters. The method which is based on header line and base lines detection and average line height is assumed as 30 pixels gives good results in case of fixed resolution images. The main purpose of this paper is estimating the average line height and based on it, finding the header lines and base lines.

In next Section, we have discussed the creation of database used for the experimental purposes. Section 2 includes the discussion about the characteristics of Hindi language. In Section 3, we have discussed the segmentation technique used for segmenting the handwritten Hindi text. Finally, Section 4-5 contains experimental results and discus-sions.

2. CHARACTERISTICS OF HINDI LANGUAGE
Devnagari is most popular script to write Hindi as well as Sanskrit, Marathi, Sindhi, and Nepali language with minor modifications. The alphabets of Devanagari script consists of 33 consonants and 14 vowels. There is no concept of lower or upper case in Hindi language. In Devnagari script, a text word may be partitioned into three zones. The upper zone denotes the portion above the headline, the middle zone covers the portion of basic and compound characters below the headline, and the lower zone may contain where some vowel and consonant modifiers can reside. For a long number of characters (basic as well as compound) there exists a horizontal line at the upper part called "shirorekha" or headline in Hindi. All characteristics with detailed explanation and images given by Raghuraj Singh [10].

The imaginary line separating the middle and lower zone may be called the base line. 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 Two consecutive lines touch or overlap each other due to these modifiers. This makes the segmentation of handwritten Hindi text very complex.

3. LINE SEGMENTATION
First we need to calculate header lines and base lines for the Line segmentation. As shown in the figure 1. header line has maximum number of black pixels and base line has minimum number of black pixels. Finding header line is a challenge because of skew in headline. Till now most of the researchers are detecting the header line by finding the row with maximum pixel density, but it cannot work for skew variable text. The algorithm proposed by M. K. Jindal helps to find header line and base line [11][12].

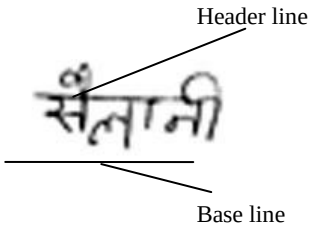

Figure 1. Header line and Base line

This method gives good results for uniform and non-uniform skewed lines. But average line height is assumed to be 30 pixels.

But different users handwritings are different, so average line height of users may be greater than 30 pixel or less than 30 pixels. Average line height of high resolution images greater than low resolution images. If 450 X 540 image has average line height as 30 pixels, 900 X 1080 image average line height will be 60 pixels, 225 X 270 image average line height will be 15 pixels. In the above two situations this algorithm fails to detect the lines accurately.

The new algorithm can work for the images having different average line height. In this new algorithm before finding the header lines and baselines we have to estimate average line height roughly and give it to the algorithm which finds baselines and header lines as an Input.

Algorithm to find out average height of a text line.

Divide the Image into three equal halves (stripes). Because skew in entire line may be high compared to skew in the first half. Perform the following steps in first half.

i. Find out the rows with minimum number of pixels and replace that row pixels with black pixels .So that we can separate text lines with black rows.

ii. Find out the height of the text lines using those black rows.

iii. Store the heights of lines in array and use sorting technique to sort the elements and take median as average height of the lines.

iv. By using average line height calculate minimum height of the consonant in a line.

Minimum height of the consonant= Average line height/4.

The algorithm for line segmentation has following steps:

1) The rows will be divided into two equal halves. For each half use the following steps to find out header lines. We are calculating header line for each half so even though there is skew in header line up to some level we can calculate header line.

2) Roughly estimate of the header lines of the text using this formula. Find number of black pixels (nopr) in each row.

   nopr(i) > 7 & nopr(i) > nopr(i :i+hc) & nopr(i:i+hc)>0
   nopr(i) > floor(wdth(i/14))

   where,

   hc : Minimum Height of consonant
   nopr(i): number of black pixels in row i
   width(i): width of line i , i.e difference between last pixel and first pixel position of line i.

3) After finding first header line, we skip number of rows equal to hc to find the next header line.

4) From (i+hc)th row to (i+rl)th row, we find the mth row with minimum of pixels.
   Where,
   rl= average line height – hc

5) We skip the rows up to mth row and go to step 2 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 lines following procedure is followed:
1) Two consecutive rough header lines are taken.
2) The line is again divided into two equal halves (stripes).
3) The rows with minimum of pixels are taken as base lines separately for each half.
4) Then the lines are separated between header lines and base lines separately for each half.

4. EXPERIMENTAL RESULTS AND ANALYSIS

Experiments for line segmentation were carried out with images which are scanned from the handwritten Dataset. Handwritten data is collected from the 30 different writers from various backgrounds such as students, teachers, doctors, lawyers and business men. Data of different sizes, i.e. writers handwritting is different so the size of characters varies from 15-60 pixels and different resolutions. Data collected from the people from various backgrounds, so the database works like real database. This database contains 29 pages and 500 lines. Number of lines per page varies from page to page. Figure 2 contain part of handwritten Hindi database.

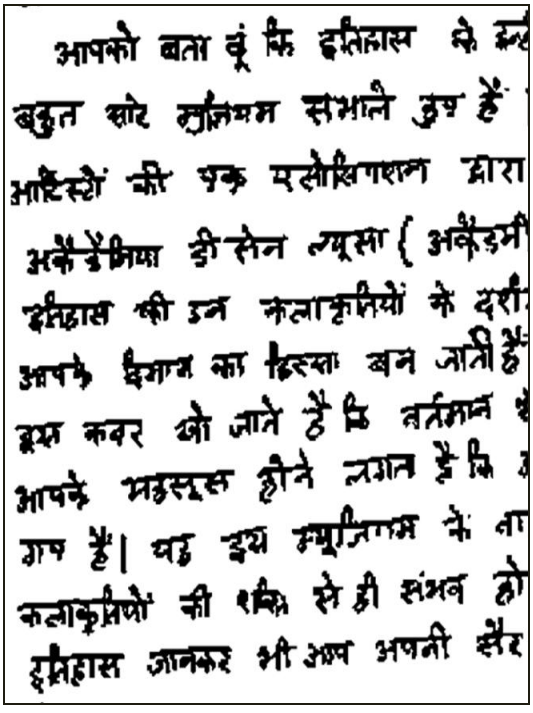

Figure 2. Part of Database

The accuracy of line segmentation depends upon the accuracy of header line and base line detection. If we assume average line height as 30 pixels in high resolution images lines are not

detected accurately as show in figure 3, because in high resolution images number of pixels used per line is high.

To solve this problem we are using the new technique for estimating average line height and lines are detected correctly as shown in figure 4. It works for high resolution images and different size characters.

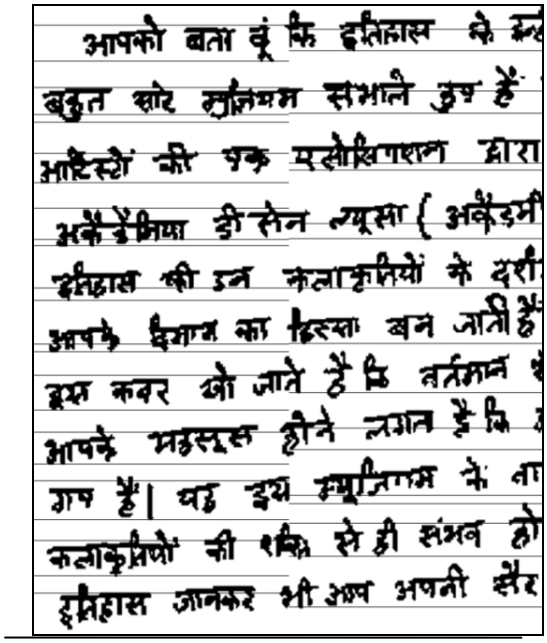

Figure 3. Result of segmentation (Average line height is assumed as 30 pixels).

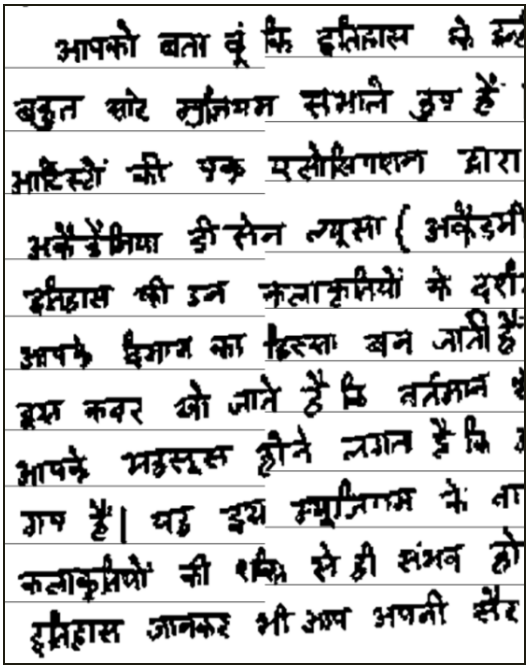

Figure 4. Result of segmentation (Average line height is estimated using algorithm).

Table 1. Accuracy of Line segmentation

	Number of lines	**Lines correctly segmented**	**% of Accuracy**
Doctors	53	50	94.33
Students	147	141	95.91
Teachers	67	61	91.04
Business men	122	113	92.62
High resolution images	111	103	92.79
Total	**500**	**468**	**93.6**

5. DISCUSSIONS

From the above table we can say that the new method is working efficiently in the cases of different text sizes and different resolution images. This method is also applicable for printed Hindi text. The lines which have broken parts in upper modifiers are not correctly recognized as shown in figure 5. The lines with thick parts in upper modifiers also not correctly recognized. Touching lines are not correctly recognized. For applying contour technique we need estimate header lines and base lines correctly, this method gives good results in finding those.

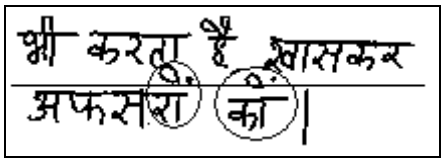

Figure 5. Wrongly segmented lines with broken parts in upper modifiers.

In future the study may be carried out on following direction:

- The above method can be applied on other languages.
- Some other technique may be tried for segmentation of overlapped or touching lines within a stripe.
- Segmentation of lines with broken parts is not done yet. It may be carried out in the future.
- Line segmentation technique mentioned above does not work for large skewed lines. So line segmentation method can be changed to improve accuracy.

6. REFERENCES

[1] B. M. Sagar, G. Shobha, P. Ramakanth kumar, "Character Segmentation Algorithms for Kannada optical character recognition", Proceedings of the International Conference on Wavelet Analysis and Pattern Recognition, 2008.

[2] Seong-whan Lee, Dong-June Lee, Hee-Seon Park, "A new Methodology for Gray-Scale Character Segmentation and Recognition", IEEE transactions on pattern analysis and machine intelligence, 1996.

[3] Satadal Saha, Subhadip Basu, Mita Nasipuri and Dipak Kr. Basu, "A Hough Transform based Technique for Text Segmentation", journal of computing, 2010.

[4] 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, 2006.

[5] 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, 1994.

[7] 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, 1995.

[8] C. Weliwitage, A. L. Harvey and A. B. Jennings, "Handwritten Document Offline Text Line Segmentation", in the Proceedings of Digital Imaging Computing: Techniques and Applications, 2005.

[9] A. Zahour, B. Taconet, L. Likforman-Sulem and Wafa Boussellaa, "Overlapping and multi-touching text-line segmentation by Block Covering analysis", Pattern analysis and applications, 2008.

[10] Raghuraj Singh, C. S. Yadav, Prabhat Verma, "Optical Character Recognition (OCR) for Printed Devnagari Script Using Artificial Neural Network", International Journal of Computer Science & Communication, 2010.

[11] Naresh Kumar Garg, Lakhwinder Kaur, M. K. Jindal, "Segmentation of Handwritten Hindi Text", International Journal of Computer Applications (0975 – 8887), 2010.

[12] Naresh Kumar Garg, Lakhwinder Kaur, MK. Jindal, "A New Method for Line Segmentation of Handwritten Hindi Text", Seventh International Conference on Information Technology, 2010.