Method for Line Segmentation in Handwritten Documents with Touching and Broken Parts in Devanagari Script

Shafali Goyal
M.Tech (Final Year)
Yadavindra College of Engineering
Talwandi Sabo, Punjab, India

Ashok Kumar Bathla
Assistant Professor
Yadavindra College of Engineering
Talwandi Sabo, Punjab, India

ABSTRACT
Now days, a vast research is going in Optical Character Recognition (OCR) of handwritten Documents in Indian scripts. A lot of handwritten data is existed in Devanagari script which is still to be recognized. Segmentation is the key step of OCR process. Segmentation is the process of extracting the valuable segments from the text document which are used in the process of recognition of characters. Line segmentation is the process of segmenting the text document into lines. Afterwards, word segmentation and character segmentation is carried out. This paper only deals with the Line segmentation of handwritten documents in Hindi. Devanagari script is the basic script to write Hindi, Marathi, Sanskrit and Nepali languages. In this paper the brief introduction of various existing techniques for segmentation of handwritten text is discussed. Also, develops an algorithm for segmentation of skewed lines, touching lines present in the text document and broken parts in upper modifiers or space present between the upper modifiers. This algorithm is implemented on large database collected from various writers. The proposed algorithm integrated the Projection based method, gap detection between text lines and neighbor pixel analysis method.

General Terms
Optical Character Recognition, Binarization, Segmentation, Documents et. al.

Keywords
Modifiers, Devanagari, OCR, Line Segmentation, Word segmentation, Character Segmentation and Recognition

1. INTRODUCTION
OCR is the process of transforming text of printed or handwritten documents into a computer process able format so that computer can recognize the characters. A good survey on OCR is explained in [1]. A wide research has been done in manly languages like English but a very few research has been done in Indian languages. Hindi is the most popular language in all over India and fourth most trendy in all over world. But a limited report of OCR is generated. In past, research in OCR of printed text in Hindi, Punjabi and other Indian languages has been done. Printed text can be of books, magazines, newspapers and various printed documents. A very good result has been out for printed text but for handwritten text still better results are not achieved. Some of the approaches for printed text are discussed in [2-3]. The Problems occurred in handwritten text are discussed in [4]. In large amount of handwritten data is stored in the offices, books by various writers which are still not recognized. One approach is defined in [5] which are used to recognize handwritten data in Devanagari on Bank cheques. OCR process is executed completely in basic three steps. These three steps are most important steps. The steps are pre-processing, segmentation and recognition. In these steps most prominent step is segmentation. For example- Segmentation of text lines in Indian scripts are explained in [6]. If the false results are generated in this step, then the false values propagated to next and the results obtained are the degraded results. The flowchart of steps of OCR is shown below (see figure 1). After text images are scanned through optical scanner the digitized images are gone through pre- processing in which noise in the form of unwanted pixels is removed and also the thresholding value is set. After that main process segmentation is continued and afterwards recognition of characters is started form segmented parts of text.

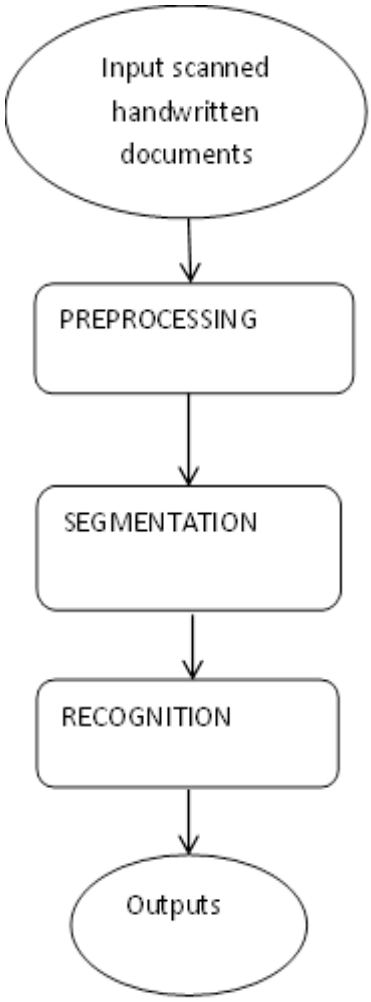


Fig 1: Steps of OCR

2. RELATED WORK

There are many approaches existed for segmentation of handwritten text documents such as detection of header line and base line and following contour technique defined in [7][8], in this approach authors discussed the new method for Line Segmentation of Handwritten Hindi text. This method is suitable for fluctuating lines or variable skew lines of text and also for overlapping of lines. One approach is defined in [9] only to segment the handwritten Devanagari words. Another approaches such as Projection based method [10], Hough transform based method [11], smearing method [12], graph clustering based method [13] are also available for segmentation of handwritten text. Some approaches implemented on Gurumukhi script in [14-16]. One of these approaches used the concept of a window to segment the scanned document image. Initially this considers the whole image as one large window. Then this large window is broken into less large windows giving lines, once the lines are identified then each window consisting of a line is used to find a word present in that line and finally to characters. A robust scheme is discussed in [17] to segment unconstrained handwritten Bangla texts into lines, words and characters. In this approach projection based method and histogram is used. Saiprakash, Renu Dhir and Rajneesh Rani in [18] described an algorithm for finding the header lines and base lines by estimating the average line height and based on it. This algorithm works efficiently on overlapped characters for different text sizes and different resolutions images. Vikas J Dongre and Vijay H Mankar in [19] proposed a simple histogram based approach to segment Devanagari documents. Another approach is implemented on Kannada handwritten text. In this approach morphological operations are integrated with projection based method to segment Kannada documents [20].

3. FEATURES OF DEVANAGARI SCRIPT

Segmentation of handwritten Hindi segmentation is very difficult since handwritten Hindi characters have not fixed shape and size. Basic Devanagari script comprises 14 vowels and 33 consonants. The writing style of Hindi is from left to right and there is no concept of upper or lower case letters as in English. Vowels can be on left, right, top or bottom of the consonant. Vowels are also called as modifiers. Many characters have horizontal line called as header line or Shirorekha at the upper part. Hindi characters have mainly three zones, one above the horizontal line that includes some vowels, matras above the horizontal line, middle zone includes the basic characters and lower zone includes the bottom modifiers below the base line. Hindi language also includes the half characters which increase the complexity of the segmentation of the characters. The half character may join with full consonant called as conjuncts. All these features are represented in figure drawn below (see figure 2).

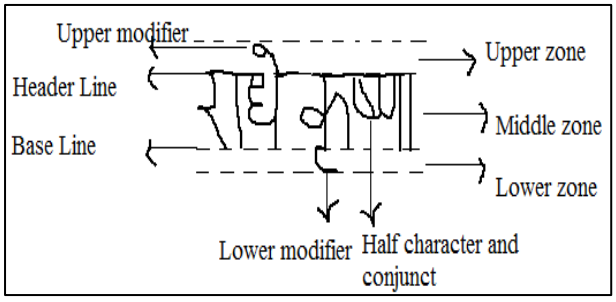


Fig 2: Zones in Hindi language

4. DATABASE

Database will be created by taking handwritten Hindi text data from large number of writers from various backgrounds because handwritten text can greatly depending on the writer skill, disposition and cultural background. The writers may be a student, Professor, Doctor or any other with different background. Database consists more than 100 lines. Database also includes data of different size, resolution and slant. After that written pages are scanned and all work is implemented on these scanned images (see figure 3)

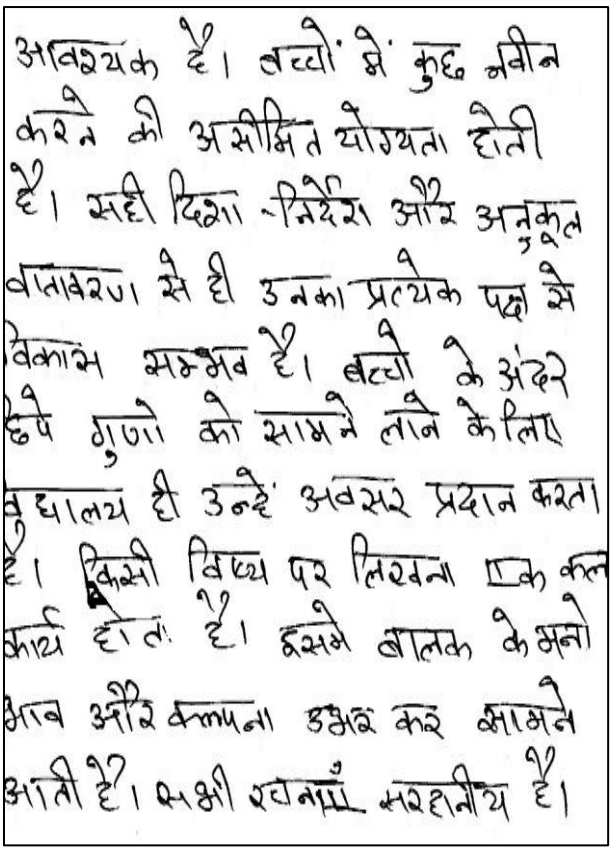


Fig 3: Portion of a database

5. SEGMENTATION

Segmentation is the process of segmenting the handwritten text documents into various segments. The segmentation is categorized as Line Segmentation, Word segmentation and Character Segmentation. Line segmentation is the process of extracting the text lines from document. Word segmentation is the process of extracting the words from segmented lines and character segmentation is the process of extracting the characters from the segmented words.

5.1 Proposed Method of Line Segmentation

This paper defines the method for line segmentation for skewed lines, touching lines and text lines with broken parts in upper modifiers or space between upper modifiers and header line. Flowchart of method for line segmentation is described in figure 5. Some of the assumptions that are assumed in this work are as follows:
The thresholding value is set to 200.
The width of the vertical stripe is 100 pixels.
The height of text line is assumed to 70 pixels.
The neighbour pixels are checked up to 2 pixels.

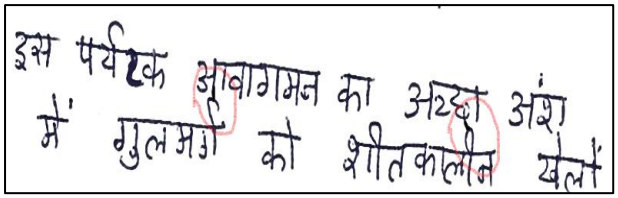


Fig 4: Touching component in text lines shown in red circles

5.1.1 Input Scanned Documents

The Handwritten documents are collected from various writers and are scanned with the help of optical scanner and are stored in the computer memory in the form of images.

5.1.2 Pre- processing

In pre-processing, some processes are executed so that final results are obtained more effective. First, the thresholding value of the images is set to 200. Then, the images are converted into binarized form i.e. in the form of 0's and 1's which is called Binarization. Afterwards, noise is removed from the images. Noise is an unwanted pixel which can degrades the results.

5.1.3 Horizontal and vertical projection

In this process, the whole image of document is divided into vertical stripes. The width of the vertical stripe is of 100 pixels. Then the next processes are executed in each vertical stripe in the horizontal direction.

5.1.4 Calculate gap between lines

In this step, start from the first stripe finds the gap between two corresponding lines. Gaps are calculated by checking the pixels of each line in horizontal direction. If the black pixels exist in line then ignore that line and scan next line. If the whole pixels are white then it is consider as space or gap. Then store these gaps pixels in an array which will be used by next step to calculate the mid-point of these gaps.

5.1.5 Calculate mid- point

Now from the array check the calculated gaps and then find midpoint of it. Now, check the difference between two mid points, if the difference increases from the expected value then calculate the average of these mid points.

5.1.6 Neighbor pixel method for broken parts

To check whether the calculated mid-point is at broken part scan the neighbor pixels values of the midpoint strip, if it contains black pixels then do not segment the line from this mid-point

5.1.7 End

Combine the segmented results and finally, draw the lines on the calculated mid points to represent the segmentation of lines. Afterwards performance analysis of results is performed.

6. RESULTS

The proposed method for line segmentation is implemented on the collected database. The segmentation rate is highly dependent on the type of handwritten text. If there is some space exist between the lines then segmentation is done perfectly. Skewed lines are segmented with high segmentation rate but the documents consisting some touching lines and broken parts in upper modifiers or space between the upper modifiers and header line are not segmented perfectly. Proposed method gives better results but not with 100 % accuracy. Some of the results are shown in figure 6 and 7. In figure 6, (a) represents the segmented text document into lines having total 5 lines without any touching or broken components out of which all lines are segmented correctly, (b) represents the result of segmented document into text lines having 8 text lines with broken parts out of which 7 lines are correctly segmented.

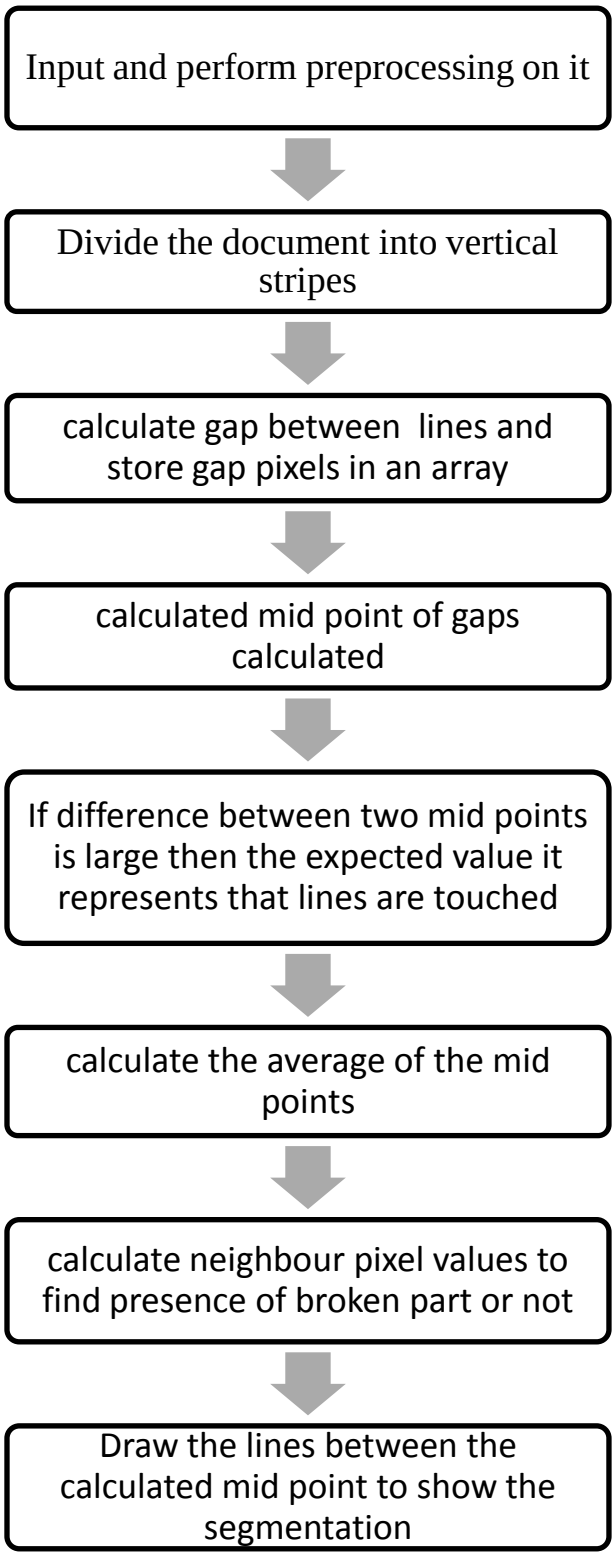
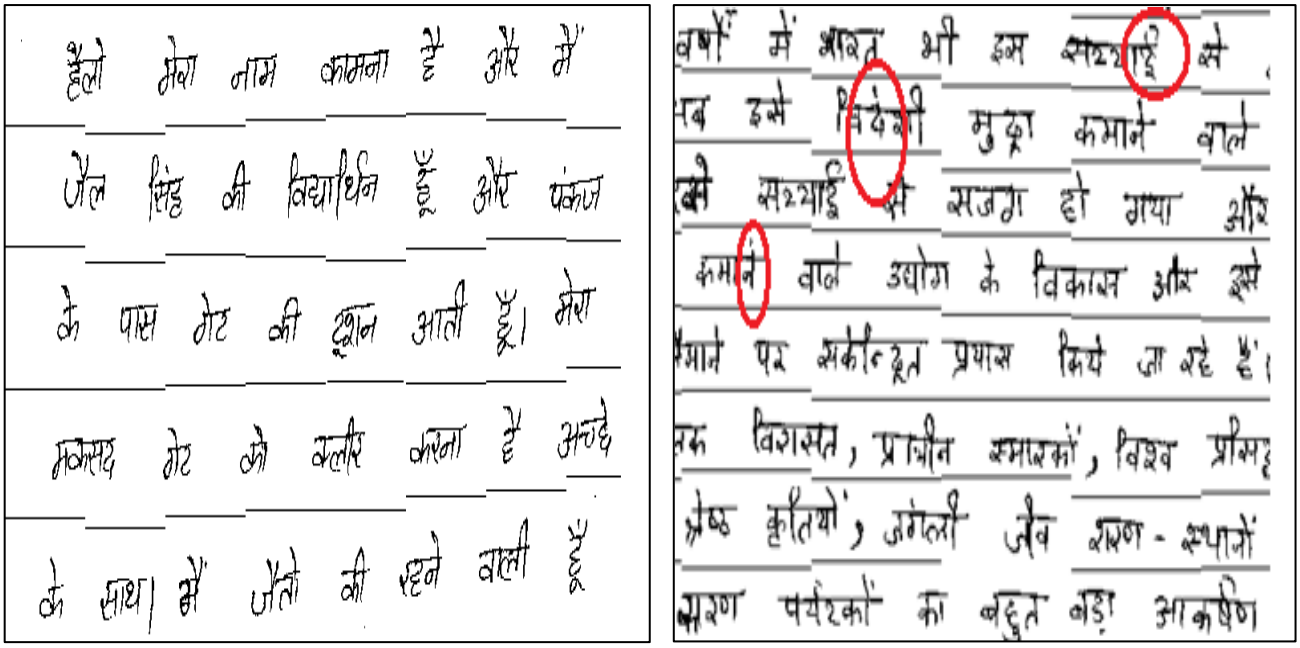


Fig 5: Flowchart of proposed framework



a) b)

Fig 6: a) result on a document with no broken and touching components b) result on a document with broken parts and space between upper modifiers and header line (shown in red circles)

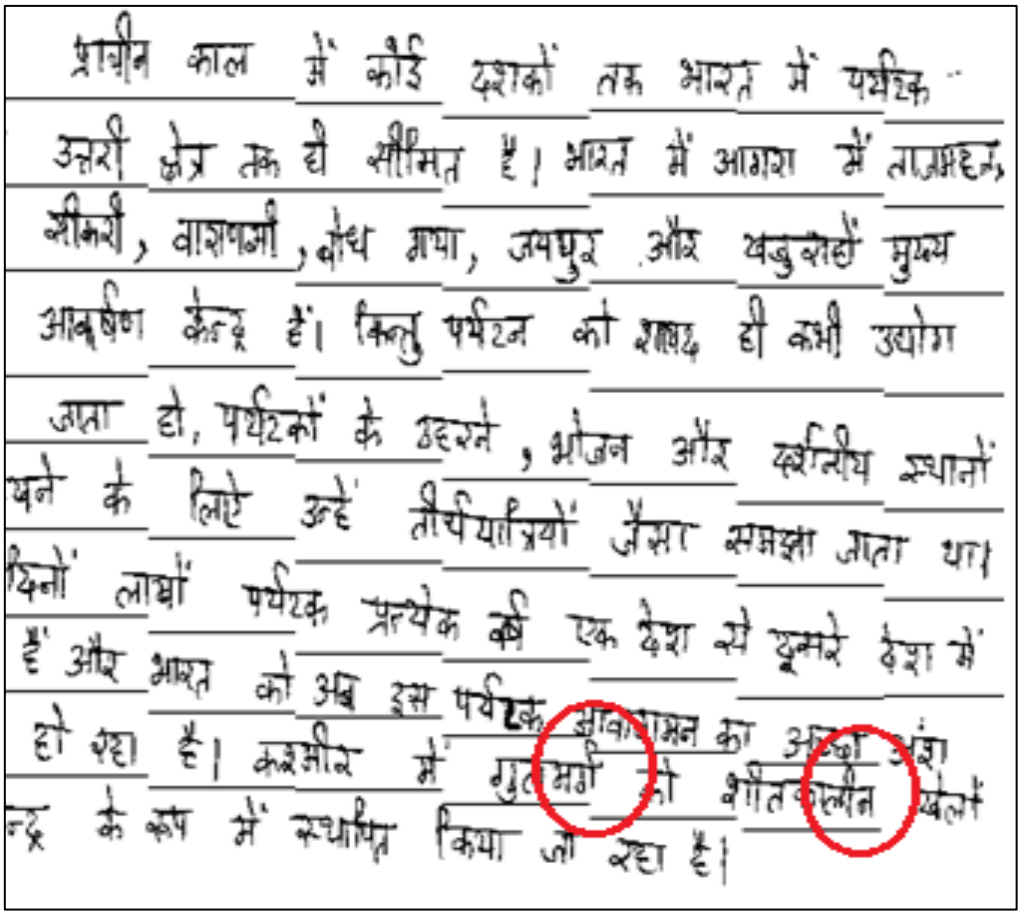


Fig 7: result of a document having touching lines shown in red circle

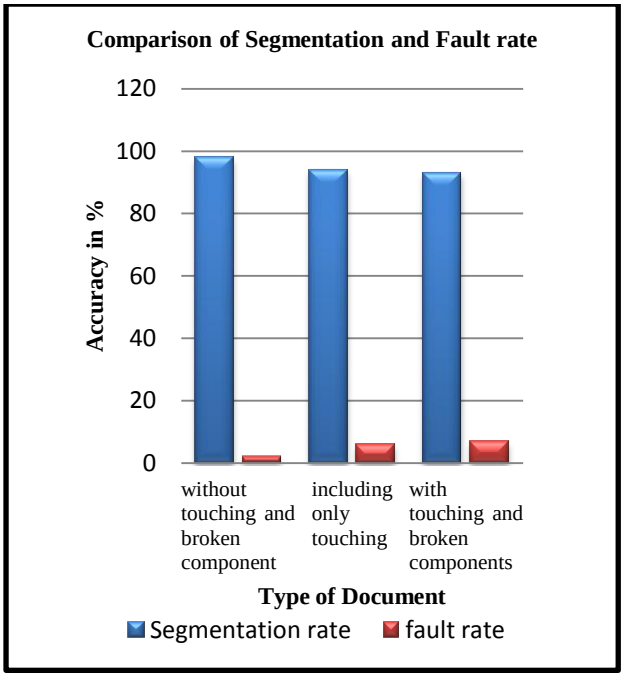


Fig 8: Graph of segmentation rates of different types of documents

Above Graph represents the comparison of accuracy of segmentation rate and fault rate in segmentation of different types of documents without touching components and broken parts, including only touching component and with touching and broken components after implementing the proposed algorithm. It is clear that segmentation rate is high for simple documents and fault rate is less as comparison to different documents.

Table 1: Overall Performance Measures of Results

Total Number of tested documents	20
Total number of lines in documents	200
Total number of lines correctly segmented	190
Accuracy in %	95

The overall results obtained on the complete collected database are shown in table 1. The collected database consist each type of document such as printed, simple handwritten documents, skewed handwritten documents, documents with touching lines and broken parts documents. Total lines in collected database are 200 out of which 190 are correctly segmented and 10 lines are incorrectly segmented. In these incorrect segmented lines few lines are having touching component and few lines having broken parts in upper modifiers. In existing techniques 93-94 % accuracy is achieved. Hence, proposed method achieves better results than existed algorithms.

7. DISCUSSIONS

There are many techniques available for segmentation of text but for handwritten text still there are few techniques. These existing techniques segment the text lines in handwritten text but not with 100 % accuracy due to so much variation in handwriting of writers. In this paper, method for segmentation of handwritten text lines in Hindi is presented. This method can be applied on the skewed lines, touching lines and the broken parts in upper modifiers or space between the upper modifiers and header line. In this method projection based method is combined with the gap analysis and neighbour pixels values analysis method. Results show better performance than existed techniques of the proposed method.

In future this technique can be used on another language and the work can be extended to increase the accuracy. Some other technique can be used for touching lines and broken parts in upper modifiers. Also the assumptions assumed in this paper can be removed. The thresholding value and text line height is automatically detected.

8. REFERENCES

[1] U.Pal and B. Chaudhuri, "Indian Script Character Recognition: a survey", Computer Vision and Pattern Recognition Unit, Vol. 37, pp. 1887-1899, 2004.

[2] Vijay Kumar and Pankaj K. Sengar, "Segmentation of Printed Text in Devanagari Script and Gurumukhi Script" International Journal of Computer Applications, Vol. 3, No. 8, June 2010.

[3] 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, Vol.3, No.4, pp. 277–286,2007.

[4] Naresh Kumar Garg, Lakhwinder Kaur and M.k Jindal "The Hazards in Segmentation of Handwritten Hindi Text" International Journal of Computer Applications, Vol. 29, No. 2, Sept. 2011.

[5] Ashwin S Ramteke and Milind E Rane, "Offline Handwritten Devanagari Script Segmentation" International Journal of Scientific & Technology Research Volume 1, Issue 4, MAY 2012.

[6] Bidyut B. Chaudhuri and Sumedha Bera "Handwritten Text Line Identification In Indian Scripts" 10th International Conference on Document Analysis and Recognition, IEEE, 2009

[7] Naresh Kumar Garg, Lakhwinder Kaur and M.k Jindal "Segmentation of Handwritten Hindi Text" International Journal of Computer Applications, Vol. 1, No. 4, 2010.

[8] Naresh Kumar Garg, Lakhwinder Kaur and M.k Jindal "A New Method for Line Segmentation of Handwritten Hindi Text", IEEE, 2010.

[9] Miss Vandana M. Ladwani and Mrs. Latesh Malik, "Novel Approach to Segmentation of Handwritten Devnagari Word", Third International Conference on Emerging Trends in Engineering and Technology, IEEE, 2010.

[10] N. Tripathy and U. pal, "Handwritten Segmentation of Unconstrained Oriya text", International Workshop on Fronteirs in Handwriting Recognition, pp. 306-311, 2004.

[11] Satadal Saha, Subhadip Basu, Mita Nasipuri and Dipak Kr. Basu, "A Hough Transform based Technique for Text Segmentation", Journal of Computing, Vol. 2, ISSUE 2, Feb 2010.

[12] Partha Pritam Roy, Umapada Pal and Josep Llados, "Morphology Based Handwritten Line Segmentation Using Foreground and Background Information", ICFHR, 2008.

[13] Vasant Manohar, Shiv N. Vitaladevuni, Huaigu Cao, Rohit Prasad and Prem Natarajan, "Graph Clustering-based Ensemble Method for Handwritten Text Line Segmentation" International Conference on Document Analysis and Recognition, IEEE 2011.

[14] Dharam Veer Sharma and Gurpreet Singh Lehal "An Iterative Algorithm for Segmentation of Isolated Handwritten Words in Gurmukhi Script" The 18th International Conference on Pattern Recognition, IEEE, 2006.

[15] Rajiv Kumar and Amardeep Singh "Detection and Segmentation of Lines and Words in Gurmukhi Handwritten Text", IEEE, 2010.

[16] Namisha Modi and Khushneet Jindal, "Text Line Detection and Segmentation in Handwritten Gurumukhi Scripts", IJARCSSE, vol. 3, issue-5, May 2013.

[17] U.Pal and Sagarika Datta, "Segmentation of Bangla Unconstrained Handwritten Text", Proceedings of the Seventh International Conference on Document Analysis and Recognition, 2003.

[18] Saiprakash Palakollu, Renu Dhir and Rajneesh Rani, **"**Handwritten Hindi Text Segmentation Techniques for Lines and Characters**"** Proceedings of the World Congress on Engineering and Computer Science, San Francisco, USA, Vol. I Oct. 2012.

[19] Vikas J Dongre and Vijay H Mankar, "Devanagari Document Segmentation Using Histogram Approach" International Journal of Computer Science, Engineering and Information Technology, Vol.1, No.3, August 2011.

[20] Mamatha H R and Srikantamurthy K, "Morphological Operations and Projection Profiles based Segmentation of Handwritten Kannada Document" International Journal of Applied Information Systems, Vol. 4, No.5, Oct. 2012.