Separation of Touching or Overlapping Lines from Handwritten Document images using Histogram and Connected Component Analysis

G. G. Rajput
Dept. of Computer Science Rani Channamma University Belagavi
Karnataka, India

Suryakant B. Ummapure
Dept. of Computer Science
Gulbarga University, Kalaburagi
Karnataka, India

Panditkumar Patil
Dept. of Computer Science
Gulbarga University, Kalaburagi
Karnataka, India

ABSTRACT

A generic approach for the separation of overlapping and touching lines within handwritten text document images is proposed in this paper. Presence of touching or skewed that arises due to ascenders or descenders and style of writer makes text line extraction a difficult task. The approach is based on histogram and connected component analysis. The proposed method is a three stage approach wherein non overlapping lines are extracted during the first stage and separation of oriented and touching lines occurs during second and third stages respectively. Average height of a text line computed using histogram profile forms the basis for text line segmentation. The proposed method has been evaluated on 120 handwritten documents written in English, Devanagari, Kannada, Telugu, and Malayalam scripts containing non-overlapping and overlapping or touching occurrences.

Keywords

Handwritten document; Text-line; segmentation; histogram; connected component.

1. INTRODUCTION

Automatic script identification is prerequisite to optical Character Recognition (OCR) system in order to feed the document to appropriate OCR for character recognition. It requires accurate text-line segmentation and word segmentation followed by character segmentation. Text-line segmentation is the first step towards script identification. Poor line segmentation leads to wrong results in recognition. Compared to printed text, line extraction from handwritten documents is a laborious task because of irregular spacing between the text lines. This results in overlapping and touching of characters in adjacent lines. In general, the lines can overlap or touch when their ascenders/descenders regions touch the above or below adjacent text line (Fig.1).

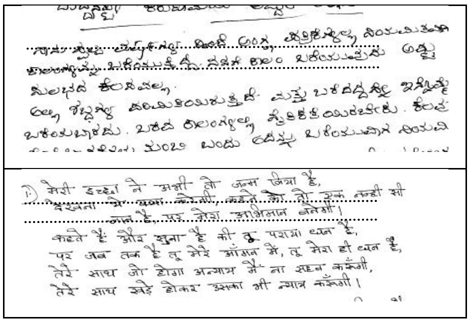

Fig 1: Documents in Kannada and Devanagari Scripts

Very few efforts have been made to the difficult problem of handwritten text-line segmentation. Few authors have roughly categorized the segmentation methods into top-down and bottom-up ones. A line segment extractor based on the theory of Kalmanfiltering is proposed in [1]. According to Kalman theory, text-lines at a certain distance can be viewed as line segments which can be used for effective segmentation of lines from document images. Experiments have been performed on ancient damaged documents of the periods between 18th and 19th century.Several methods have been proposed for dealing with touching adjacent lines. Text-line segmentation of handwritten documents in Hindi and English is described in [3]. The document image is binarized and connected components are identified. Based on Hough lines the text-lines are identified. Skew angle is then determined by calculating the slope of the detected line and the skewness is minimized. Segmentation is then performed and the result is refined by removing the noise which basically comprises components from adjacent lines.

Line Segmentation of Handwritten Devanagari Text to detect header line and base lines accurately for text-line extraction is proposed in [6]. The average line height is estimated, before calculating the header line and base lines. After finding the average height the rows are divided into two equal halves. Rough estimate of the header lines of the text is computed by determining number of black pixels in each row. After finding first header line, a threshold number of rows are skipped to find the next header line. Two consecutive rough header lines are taken; the line is again divided into two equal halves (stripes).The rows with minimum of pixels are taken as base lines separately for each half and then the lines are separated between header lines and base lines separately for each half. A line and word segmentation method to handle the deformations like touching components, overlapping components, skewed lines, words with individual skews To build a proper text image with all these deformations removed is described in [9]. Line segmentation procedure is applied using Hough transform.General Text-line Extraction Approach based on Locally Orientation Estimation for multi-oriented text-line extraction from historical handwritten Arabic documents is reported in [11]. Image paving, using a small window, is done to progressively and locally determine the lines. Snake technique is adopted for line extraction. Once the paving is established, the orientation is determined using the Wigner-Ville distribution on the histogram projection profile. This local orientation is then enlarged to limit the orientation in the neighborhood. Afterwards, the text-lines are extracted locally in each zone basing on the follow-up of the baselines and the proximity of connected components. Finally, the connected components that overlap and touch in adjacent lines are separated with knowledge of the terminal letters of Arabic words. The line segmentation method for touching and broken part in Devanagari script is proposed in [17] by dividing the document into vertical strips. By calculating the midpoint of the gap and if the difference between two midpoint is large, the lines are touching. Average of midpoint is calculated to find the broken part to draw the lines between the calculated midpoints to represent segmentation.The profile technique for line segmentation has been proposed in [18][22] by dividing the entire document into several strips, where minimum(10pixels), maximum(25 pixels) and average(20 to 40 pixels) height of text line is assumed, the text line is merged to previous or next text block if height of text block is less than 10 pixels. A preprocessed image is stored in a two dimensional array to get its height, width and number of black pixels are calculated and stored in array for each divided strip. If the value of an array (height) is less than zero or one or two then red color is assigned to that pixel and if the value of an array (height) is greater than the new pixel value then it is treated as initial point and the process is repeated for until the array value equal to zero. When initial and final values are obtained the steps are repeated to display the entire consecutive text blocks of the document. Average height of the first strip and each text block is calculated. If the average height is less than 10 pixels then merge the text block to previous or next block and large heighted block. Minimum density pixel is traced until the profile array is equal to the point of minimum density pixel to divide the block into two parts.Header and base line detection is proposed in [19] for segmentation of text line. minimum (8pixels), maximum (25 pixels) and average (20 to 40 pixels) height of text line is assumed and finally the number of pixels in row and difference between last and first pixel is calculated, if more characters are combined to form a word then the horizontal lines touch each other .By generating header line which is equal to minimum height, find the row with minimum number of pixels to skip to find next header line. Actual text line is obtained by considering and dividing the two consecutive header lines into four equal parts. A row with minimum pixel are considered as base line by separating the lines between header and base line .Actual text line is obtained by joining four separate lines.

Morphological approach is described in [20] for Persian text line segmentation by constructing complementary image to label all the objects in the image. The objects related to sticking lines via slanted parts, dots and noise are removed. A two stage dilation is performed on remaining objects to label it as lines of handwritten text by assuming that the font and height of the text written , height of all bounding boxes and distance between is same, if not the lines are touching and erosion operation is used to differentiate two lines. And line borders are determined by counters.A method that compute energy map, to determine seams passing across and between lines of an image in [21] with two stage algorithm to work on binary and gray scale image. In first stage the components along the lines are extracted by passing seam middle, along the line to make the components as letter and words assigns the unmarked components to the nearest text line. Distance transform is computed to generate medial and separating seams to determine text lines and upper and lower boundaries of text line respectively in the second stage for segmenting the text lines. In [23] block covering analysis technique is proposed for overlapping and touching line segmentation. The image is divided regularly into vertical halves of width , the obtained shape is projected vertically on the axis to find height and position of covering block and histogram of projection is obtained for empty and non-empty lines with intervals, where non-empty line represents height and position of covering block with the halve. After preprocessing an image Block covering fractal analysis is incorporated to classify tightly spaced and widely spaced documents and the blocks are classified small, average and large .Then these large blocks are segmented, mean while small and average blocks are classified using statistical block analysis and are assigned to neighboring block analysis for text line extraction. Hough transform based method for text line segmentation is proposed in [24] for digitized images. The noise is removed during preprocessing and images are converted to gray scale and binarized for black and white foreground. Connected components in the Hough image are obtained by applying Hough transform. The text lines are extracted using connected component labeling where white pixel searches for its white neighbor and black pixel for black neighbor. 4-connected neighbors are searched and connected components are labeled accordingly.Segmentation of touching, overlapping words in adjacent lines of handwritten text is proposed in [25].Here the entire image is divided into set of connected components for processing. The component which exceeds the average line height denotes the touching or overlapping region. All the pixels included to a single component belong to distinct words, adjacent lines and these locations are marked. Connected component analysis is performed to locate the interconnected boundary area between the components and pixels are labeled belonging to upper and lower line.In [26] a method described for line extraction technique from handwritten document images using histogram and connected component analysis. Using horizontal histogram profile, average height of a line in the given document is computed, using which the non-overlapping lines are extracted. In order to extract overlapping lines, that exceed the given threshold, a rectangular bounding box is imposed over the words of the overlapping lines using connected component analysis. The mid-point of each bounding box is then calculated and compared with the average height of the image to label each component as either belonging to upper line or lower line.All above approaches are applicable to one script or two scripts with specific morphology which are not easy adaptable to other scripts [1-16]. Further, attempts to segment out lines successfully in case of overlapping lines in a handwritten text-line document are not observed in experimental results of the proposed methods. In this paper a script independent text-line approach is described employing histogram profile and connected component technique. The proposed method successfully segments text-lines including overlapping or skewed text-lines from handwritten document images. Five major Indian scripts including English are considered for performing experiments. The method proposed in this paper is extension of the work done in [26].

2. METHODOLOGY

Handwritten documents written in Kannada, Telugu, Hindi, English and Malayalam are collected from different writers. The documents are then scanned using HP LaserJet Professional scanner in gray scale with 300dpi resolution and images are saved in jpeg format. Median filter is used to remove noise present in the scanned document images. The gray scale document images are then converted to binary images by using Otsu’s threshold algorithm [2].Objects that are smaller than a threshold size, determined empirically, are eliminated from the documents with the assumption that these object(s) have aroused as a part of binarization. In order to extract text-lines from the scanned document image after pre-processing, a three-stage method is proposed. During the first stage, using horizontal profiles, lines with little or no orientation of text-lines are extracted from the documents. The second stage corresponds to oriented lines (horizontal profiles of text-lines overlap) with certain degree of orientation. The third stage corresponds to those lines that are touching. Using a threshold value overlapping adjacent text lines is detected and connected component technique is employed to extract these lines. The methodology is explained below.

Stage - 1: Horizontal profile of pre-processed document image is computed. The average height of text-line is determined based upon density of pixels in each line of the profile. Using this as a threshold, the document is parsed line by line to determine lines with no on pixels. The text-line lying between such lines is extracted subject to the condition that the height of the text-line meets the required threshold criteria of the average height of the text-line. In case, the height of the text-line exceeds the threshold, it is assumed that there are at least two overlapping lines, and such lines are subjected to second stage for text-line separation. By overlapping lines we mean, the on pixel density in the horizontal profile of neighborhood lines overlap and hence the height of the combined text-lines exceeds the average-height of the line making it difficult to separate the text-lines (Fig.4). Such lines usually arise due to the variation in writing text-lines by the writer.

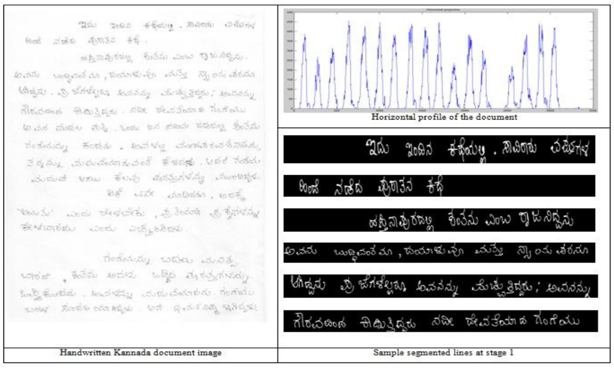
Horizontal profile of the document
Handwritten Kannada document image
Sample segmented lines at stage 1

Fig 2: Sample handwritten document image with little or no orientation of text-lines

Stage – 2: The overlapping lines are subjected to connected component analysis [16] to separate the text-lines. A bounding box is fitted to each of the connected objects (word) of the text-lines. The center of the object is computed. The distance between the object and center of the above and below lines are computed, respectively, and each object is labeled as La (object belongs meaning to above line), or Lb (meaning object belongs to below line) based upon nearest neighbor. Euclidean distance measure is used to compute the distance. Finally, the lines are segmented out as upper line and below lines based on the labels assigned to the objects.

Stage – 3: To extract overlapping and touching line,a rectangular bounding box is imposed over each character of overlapping and touching line. A 4 or 8 connected neighbors are searched for component analysis. The mid-point of each bounding box is then calculated and compared with the average height of the image to label each component as either belonging to upper line or lower line.The block diagram of the proposed method is shown in Figure 9.

3. RESULTS AND DISCUSSIONS

The experiments are performed on handwritten document images written in different scripts namely, Kannada, Telugu, Hindi, English and Malayalam, respectively. A total of 120 document images are considered for performing experiments. The preprocessed handwritten document image is subjected to three stage approach for line segmentation and extracted lines are stored in a separate folder. The results for few of the documents are shown in Fig 2. through Fig 5. It is observed that, non-overlapping text-lines, evident from their horizontal profiles, were extracted at stage 1. Further overlapping and touching lines have been successfully segmented from the documents. Fig 6. Through Fig 8. shows results of such document images. Table 1 describes the results for each type of the document considered.

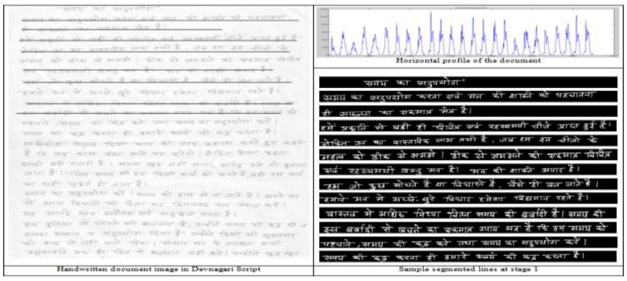
Horizontal profile of the document
Handwritten document image in Devnagari Script
Sample segmented lines at stage 1

Fig 3: Sample handwritten document image with orientated (curved) text-lines (marked by straight lines)

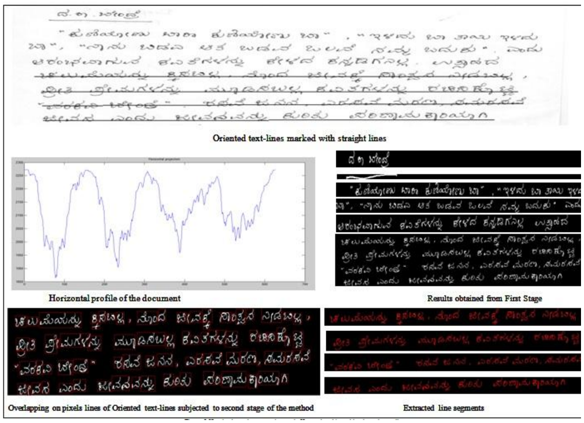
Oriented text-lines marked with straight lines
Horizontal profile of the document
Results obtained from First Stage
Overlapping on pixels lines of Oriented text-lines subjected to second stage of the method
Extracted line segments

Fig 4: Handwritten document image in Kannada with multi-oriented text-lines

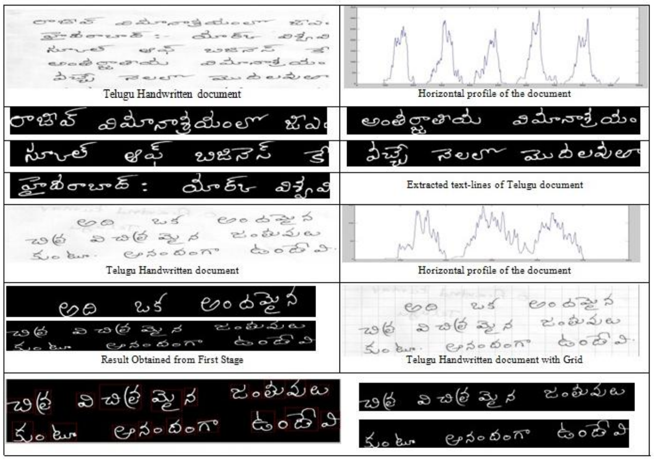
Telugu Handwritten document
Horizontal profile of the document
Extracted text-lines of Telugu document
Telugu Handwritten document
Horizontal profile of the document
Result Obtained from First Stage
Telugu Handwritten document with Grid

Fig 5: Handwritten document image in Telugu with multi-oriented text-lines and segmented lines

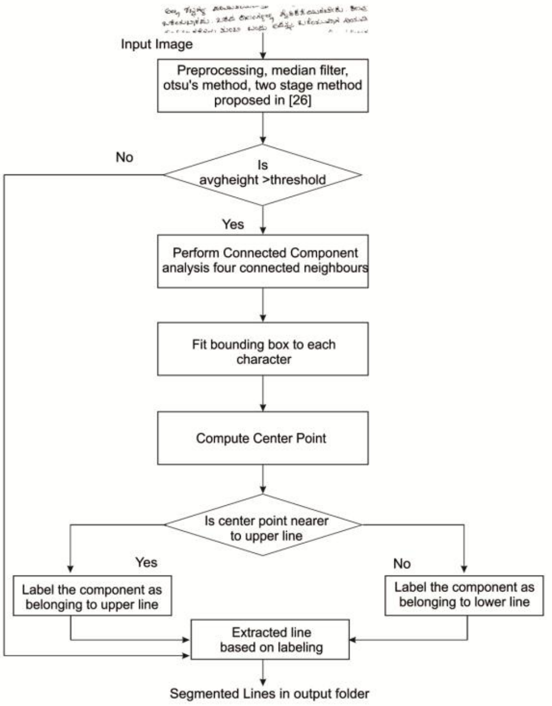
Input Image
Preprocessing, median filter, otsu's method, two stage method proposed in [26]
Is avgheight >threshold
No
Yes
Perform Connected Component analysis four connected neighbours
Fit bounding box to each character
Compute Center Point
Is center point nearer to upper line
Yes
No
Label the component as belonging to upper line
Label the component as belonging to lower line
Extracted line based on labeling
Segmented Lines in output folder

Fig 6: Block diagram of proposed method

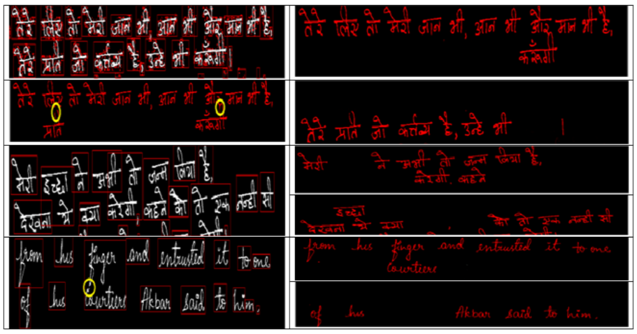

Fig 7: Sample images with touching characters and results after seconds stage of the proposed method

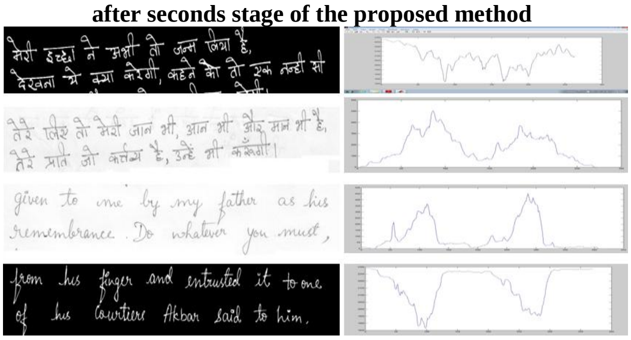

Fig 8: Touching lines with histogram

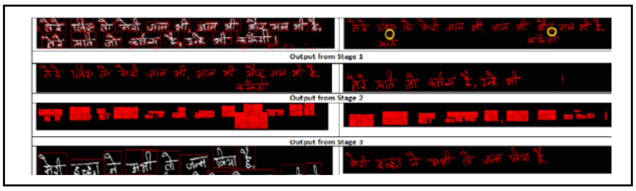

Fig 9: Sample Images with touching characters and their Outputs after third stage of proposed method

Table 1. Result Analysis of Separation of Touching or

Result Analysis of Separation of Touching or Overlapping Lines			
Type of Handwritten document	**Occurrence of Overlapping / Touching line**	**Separations failed**	**Successful Separation Rate %**
Kannada	92	2	97.82 %
Hindi	99	6	93.93 %
English	81	3	96.29 %
Telugu	245	11	95.51 %
Malayalam	187	5	97.32%

4. CONCLUSION

An efficient three stage method for text line extraction proposed in this paper is extension of the proposed method in [26] to extract text-lines from handwritten document images. Non overlapping lines are extracted during the first stage and separation of oriented and touching lines occurs during second and third stages respectively. Average height of a text line computed using histogram profile forms the basis for text line segmentation. Extraction of text-lines from document images with lines appearing curved (oriented) and characters of adjacent lines touching poses difficulty in segmentation. Using Connected Component Analysis bounding box is inserted on the components (words/characters) of adjacent lines and each component is labeled as to belong to a specific text-line based on a threshold. The resulting lines are then extracted using these labels. Experiments are carried out on handwritten document images written in different scripts and the results obtained are encouraging. The proposed method also extracts the lines with touching characters successfully. However, in certain cases for eight connected component the resulting segmentation is not accurate and will be improved in future work.

5. REFERENCES

[1] Lemaitre, Aurélie, and Jean Camillerapp. “Text-line extraction in handwritten document with Kalman filter applied on low resolution image”. Document Image Analysis for Libraries, 2006. DIAL'06. Second International Conference on. IEEE, 2006.

[2] Anusree.M and Dhanya.M.Dhanalakshmy.“Text-line Segmentation of Curved Document Images”.Anusree.M et al Int. Journal of Engineering Research and Applications ISSN : 2248-9622, Vol. 4, Issue 5(Version 5), May 2014, pp.32-36

[3] Sunanda Dixit, Sneha, Nilotpal Utkalit and Suresh .H.N. “Text-line Segmentation of Handwritten Documents in Hindi and English”. International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 2 Issue: 4 733 – 739.

[4] Vikas J Dongre and Vijay H Mankar. “DEVNAGARI DOCUMENT SEGMENTATION USING HISTOGRAM APPROACH”.International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.1, No.3, August 2011. 4

[5] Neha Sahu. “DEVANAGIRI DOCUMENT SEGMENTATION USING HISTOGRAM BASED APPROACH”.International Journal of Electronics, Electrical and Computational System IJEECS ISSN 2348-117X Volume 3, Issue 3 May 2014.

[6] Saiprakash Palakollu, RenuDhir and Rajneesh Rani. “A New Technique for Line Segmentation of Handwritten

Hindi Text". Special Issue of International Journal of Computer Applications (0975 – 8887) on Electronics, Information and Communication Engineering - ICEICE No.5, Dec 2011.

[7] Saiprakash Palakollu, RenuDhir and Rajneesh Rani. "Segmentation of Handwritten Devanagari Script". SaiprakashPalakollu et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 (3), 2011, 1244-1247. ISSN: 0975-9646.

[8] Rahul Garg and Naresh Kumar Garg. "An algorithm for Text-line Segmentation in Handwritten Skewed and Overlapped Devanagari Script". International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 5, May 2014).

[9] Varsha Hole, LeenaRagha and Pravin Hole. "Text-line and Word Segmentation of Indian Script Handwritten Document". International Conference & Workshop on Recent Trends in Technology,(TCET) 2012 Proceedings published in International Journal of Computer Applications®(IJCA).

[10] M.Ravi Kumar, B.P.Pragathi and Nayana N Shetty. " Text-line Segmentation of Handwritten Documents using Clustering Method based on thresholding Approach". International Journal of Computer Applications (0975 – 8878),on National Conference on Advanced Computing and Communications - NCACC, April 2012

[11] Nazih Ouwayed, Abdel Belaid and Francois Auger. "General Text-line Extraction Approach based on Locally Orientation Estimation". Author manuscript, published in "Document Recognition and Retrieval XVII - DRR 2010, 17th Document Recognition and Retrieval Conference, San Jose, CA : United States (2010)".

[12] Saiprakash Palakollu, RenuDhir and Rajneesh Rani. "Handwritten Hindi Text Segmentation Techniques for Lines and Characters". Proceedings of the World Congress on Engineering and Computer Science 2012 Vol IWCECS 2012, October 24-26, 2012, San Francisco, USA. 12

[13] Jayant Kumar, Le Kang David, Doermann Wael ,Abd-Almageed. "Segmentation of handwritten text lines in presence of touching components." Document Analysis and Recognition (ICDAR), 2011 International Conference on. IEEE, 2011.

[14] NazihOuwayed, Abdel Belaid. "Separation of Overlapping and Touching Lines within Handwritten Arabic Documents". Xiaoyi Jiang and Nicolai Petkov. The 13th International Conference on Computer Analysis of Images and Patterns - CAIP 2009, Sep 2009, Munster, Germany. Springer Berlin / Heidelberg, 5702, pp.237-244.

[15] Ram Sarkar,Nibaran Das,Subhadip Basu,Mahantapas Kundu,Mita Nasipuri and Dipak Kumar Basu. "CMATERdb1:a database of unconstrained handwritten Bangla and Bangla-English mixed script document image".IJDAR DOI 10.1007/s 10032-011-0148-6 Published online:24 February 2011.

[16] Rafael C. Gonzalez and Richard E. Woods " Digital Image Processing", Third Edition, Published by Pearson Education,Inc. and Dorling Kindersley Publishing,Inc. ISBN 978-81-317-1934-3.

[17] Shafali Goyal, Ashok Kumar Bathla " Method for Line Segmentation in Handwritten Documents with Touching and Broken Parts in Devanagari Script" .International Journal of Computer Applications (0975 – 8887) Volume 102– No.12, September 2014.

[18] Rahul Garg, Naresh Kumar Garg " An algorithm for Text Line Segmentation in Handwritten Skewed and Overlapped Devanagari Script". International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 5, May 2014).

[19] PRAMOD S. MALGI & SHAILJA GAYAKWAD "LINE SEGMENTATION OF DEVNAGRI HANDWRITTEN DOCUMENTS". International Journal of Electronics, Communication& Instrumentation Engineering Research and Development (IJECIERD) ISSN (P): 2249-684X; ISSN (E): 2249-7951Vol. 4, Issue 2, Apr 2014, 25-32© TJPRC Pvt.Ltd.

[20] Abdollah Amirkhani-Shahraki, Amir Ebrahimi Ghahnavieh and Seyyed Abdollah Mirmahdavi. "A Morphological Approach to Persian Handwritten Text Line Segmentation". 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation.

[21] Raid Saabni , Abedelkadir Asi , Jihad El-Sana . "Text line extraction for historical document images" Pattern Recognition Letters 35 (2014) 23–33 0167-8655/$ - see front matter _ 2013 Elsevier.

[22] Naresh Kumar Garg, Lakhwinder Kaur and M. K. Jindal. "Segmentation of Handwritten Hindi Text". ©2010 International Journal of Computer Applications (0975 – 8887)Volume 1 – No. 4

[23] Abderrazak Zahour, Brunco Taconet, Laurence Likforman-Sulem and Wafa Boussellaa. "Overlapping and multi-touching text-line segmentation by Block Covering analysis". Pattern Anal Applic (2009) 12:335–351 DOI 10.1007/s10044-008-0127-9 Springer.

[24] Satadal Saha, Subhadip Basu, Mita Nasipuri and Dipak Kr. Basu ."A Hough Transform based Technique for Text Segmentation".JOURNAL OF COMPUTING, VOLUME 2, ISSUE 2, FEBRUARY 2010, ISSN 2151-9617 .

[25] KALYAN TAKRU and GRAHAM LEEDHAM "SEPARATION OF TOUCHING AND OVERLAPPING WORDS IN ADJACENT LINES OF HANDWRITTEN TEXT". Proceedings of the Eighth International Workshop on Frontiers in Handwriting Recognition (IWFHR'02) 0-7695-1692-0/02 $17.00 © 2002 IEEE.

[26] G. G. Rajput, Suryakant B. Ummapure and Preeti N Patil. "Text-Line Extraction from Handwritten Document images using Histogram and Connected Component Analysis". International Journal of Computer Applications (0975 – 8887) National conference on Digital Image and Signal Processing, DISP 2015.