Text-Line Extraction from Handwritten Document images using Histogram and Connected Component Analysis

G. G. Rajput
Rani Channamma University
Belagavi, Karnataka

Suryakant B. Ummapure
Dept. of Computer Science
Gulbarga University, Gulbarga

Preethi N. Patil
Dept. of Computer Science
Gulbarga University, Gulbarga

ABSTRACT

Text-line segmentation is an essential part of script identification technique from handwritten and printed document images. In case of handwritten documents, overlapping, touching, skewed and small perforations between the lines makes line extraction difficult task. Presence of such variations leads to errors and wrong identification of script. This paper describes an efficient line extraction technique from handwritten document images using histogram and connected component analysis. Using horizontal histogram profile, a threshold, i.e., 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. Experiments are carried out on document images of Kannada, Telugu, Hindi, English and Malayalam scripts and the results obtained are encouraging.

Keywords

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

1. INTRODUCTION

India is a Multi-Script and Multi-Lingual country having more than 18 regional languages derived from 12 different scripts, namely Kannada, Telugu, Tamil, Malayalam, Punjabi, Gujarati, Marathi, Oriya, Sanskrit, Urdu, Bangla and Hindi. Automatic script identification is prerequisite to Optical Character Recognition (OCR) system in order to feed the document to appropriate OCR for character recognition. Further, accurate text-line segmentation is essential for Script Identification at line level. Text-line segmentation is a very crucial step in optical character recognition. Poor line segmentation leads to wrong results in recognition. In printed text, line segmentation is quite easy but in handwritten text, it is quite difficult due to problems like overlapping, touching of characters and also due to different writing style of a writer. A brief overview of methods and techniques available in the literature for text-line segmentation is described below.

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 bottom-up approach, text-line segmentation algorithm based on Minimal Spanning Tree (MST) clustering with distance metric learning is proposed in [1,2] to extract lines from Chinese handwritten documents. The connected components of document image are grouped into a tree structure and text-lines are extracted by dynamically cutting the edges of the tree using an objective function.

Based on seam carving approach to extract the text-lines from handwritten documents, in [3], energy maps are generated using oriented distance function. The minimum energy path that moves from left to right is computed .This line moves through the midpoint of components that fall on this path. The components belonging single path are used to find the height of the text line. The computed average height is used as threshold to split the touching lines. Experiments are performed on Arabic, Chinese, and English historical documents to separate multi-skew text blocks into lines with high success rates.

A line segment extractor based on the theory of Kalman filtering is proposed in [4]. 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.

A top-down approach based segmentation of touching, overlapping, skewed and short handwritten text-lines is proposed in [6]. The proposed method begins with core detection. To segment the overlapping components, run-length is used for obtaining the structural knowledge which classifies the components into upper and lower text-lines. To segment the short lines and skewed lines, distance metrics and connected component are used recursively. Experiments are performed on IAM Database consisting forms of handwritten English text and also on documents collected from different writers.

Text-line segmentation of handwritten documents in Hindi and English is described in [7]. 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.

A Steerable Directional Local Profile Technique for extraction of Handwritten Arabic text-lines is proposed in [9]. The proposed technique is based on a generalized adaptive local connectivity map (ALCM) using a steerable directional filter. The algorithm is designed to solve the problems such as fluctuating, touching or crossing text-lines. The algorithm consists of three steps. Firstly, a steerable filter is used to probe and determine foreground intensity along multiple directions at each pixel while generating the ALCM. The ALCM is then binarized using an adaptive thresholding algorithm to get a rough estimate of the location of the text-lines. In the second step, connected component analysis is used to classify text and non text patterns in the generated ALCM to refine the location of the text-lines. Finally, the text-lines are separated by superimposing the text-line patterns in the ALCM on the original document image and extracting the connected components covered by the pattern mask. Experiments are performed on the DARPA MADCAT Arabic handwritten document data set. The proposed method is capable of correctly isolating handwritten text-lines.

Line Segmentation of Handwritten Hindi Text to detect header line and base lines accurately for text-line extraction is proposed in [11]. 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 is 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 new sequence of line and word segmentation method to handle some of the deformations usually present in the handwritten document like touching components, overlapping components, skewed lines, words with individual skews etc. and build a proper text image with all these deformations removed is described in [14]. Line segmentation procedure is applied using Hough transform. The proposed method of line segmentation is a sufficiently accurate to extract the text-lines from unconstrained handwritten text documents.

Text-line extraction from handwritten document pages using Spiral run length Smearing algorithm (SRLSA) is presented in [16]. Firstly, digitized document image is partitioned into a number of vertical fragments of equal width. Then all the text-line segments present in the fragments are identified by applying SRLSA. Finally, the neighboring text-line segments are analyzed and merged (if necessary) to place them inside the same text-line boundary in which they actually belong to. Experiments are performed on document images taken from CMATERdb1.1.1 and CMATERdb1.2.1 databases.

General Text-line Extraction Approach based on Locally Orientation Estimation for multi-oriented text-line extraction from historical handwritten Arabic documents is reported in [17]. Image paving is done to progressively and locally determine the lines. The paving is initialized with a small window and then its size is corrected by extension until enough lines and connected components were found. 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. The morphology analysis of the terminal letters of Arabic words is here considered.

From the available literature, we observe that many of the attempts for line segmentation are proposed for particular type of single script[3-6, 8-11, 13, 16-20] and two scripts [1, 2, 7, 14, and 15]. 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. This issue is attempted in this paper using histogram profile and connected component technique. The proposed method successfully segments text-lines including overlapping/skewed text-lines from handwritten document images. Five major scripts are considered for performing experiments. The rest of the paper is described as follows. Methodology is explained in section-2. Experimental results are described in section-3 and conclusion is given in section-4.

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 [5].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, we propose a two-stage method. During the first stage, using horizontal profiles, lines with little or no orientation of text-lines are extracted from the documents (Fig 1.). The second stage corresponds to those lines that are overlapping (horizontal profiles of text-lines overlap) with certain orientation(Fig. 2).Using a threshold value overlapping text-lines are detected and connected component technique is employed to extract the lines. The methodology is explained in detail below. During stage-1, horizontal profile is computed for the pre-processed document. The average height of the 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

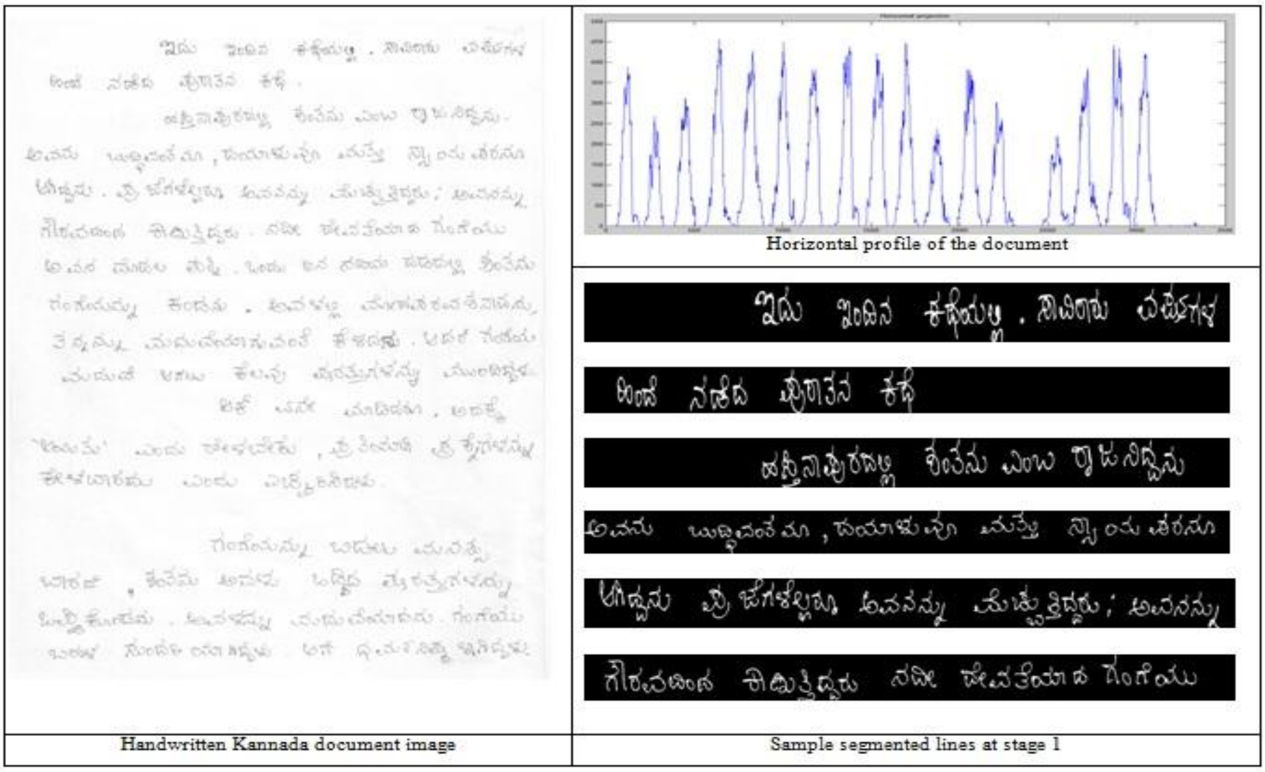

Horizontal profile of the document
Handwritten Kannada document image
Sample segmented lines at stage 1

Fig 1: Sample handwritten document image with little or no orientation of text-lines

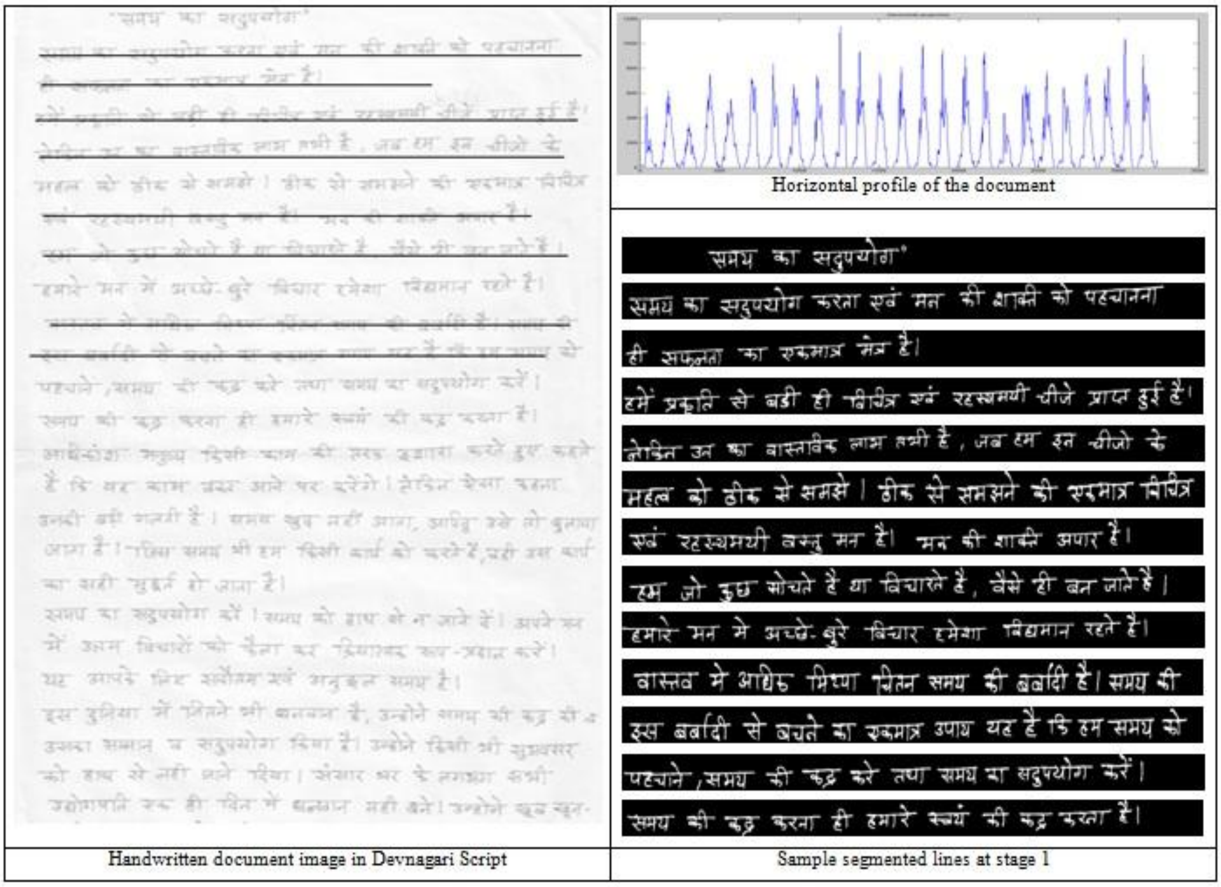

Horizontal profile of the document
Handwritten document image in Devnagari Script
Sample segmented lines at stage 1

Fig 2: Sample handwritten document image with orientated (curved) text-lines (marked by straight lines)

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 exceed the average-height of the line making it difficult to separate the text-lines (Fig.3).

Such lines arise due to the variation in writing text-lines by the writer. In the second stage, the overlapping lines are subjected to connected component analysis [22] to separate the text-lines. Using connected component technique, bounding box is fitted to each of the connected objects (words) of the text-lines. The center of the objects 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. The block diagram of the proposed method is shown in the fig.6.

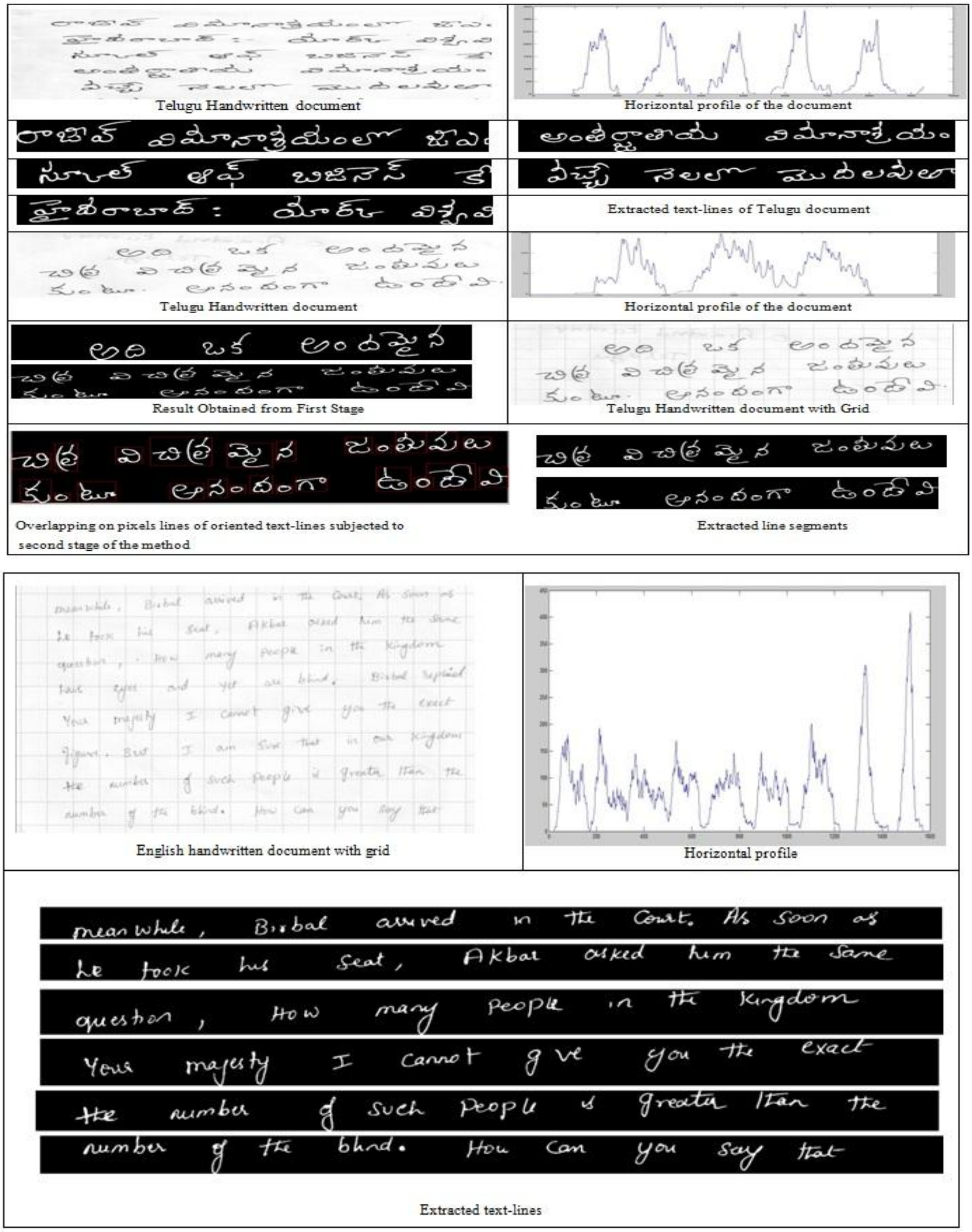
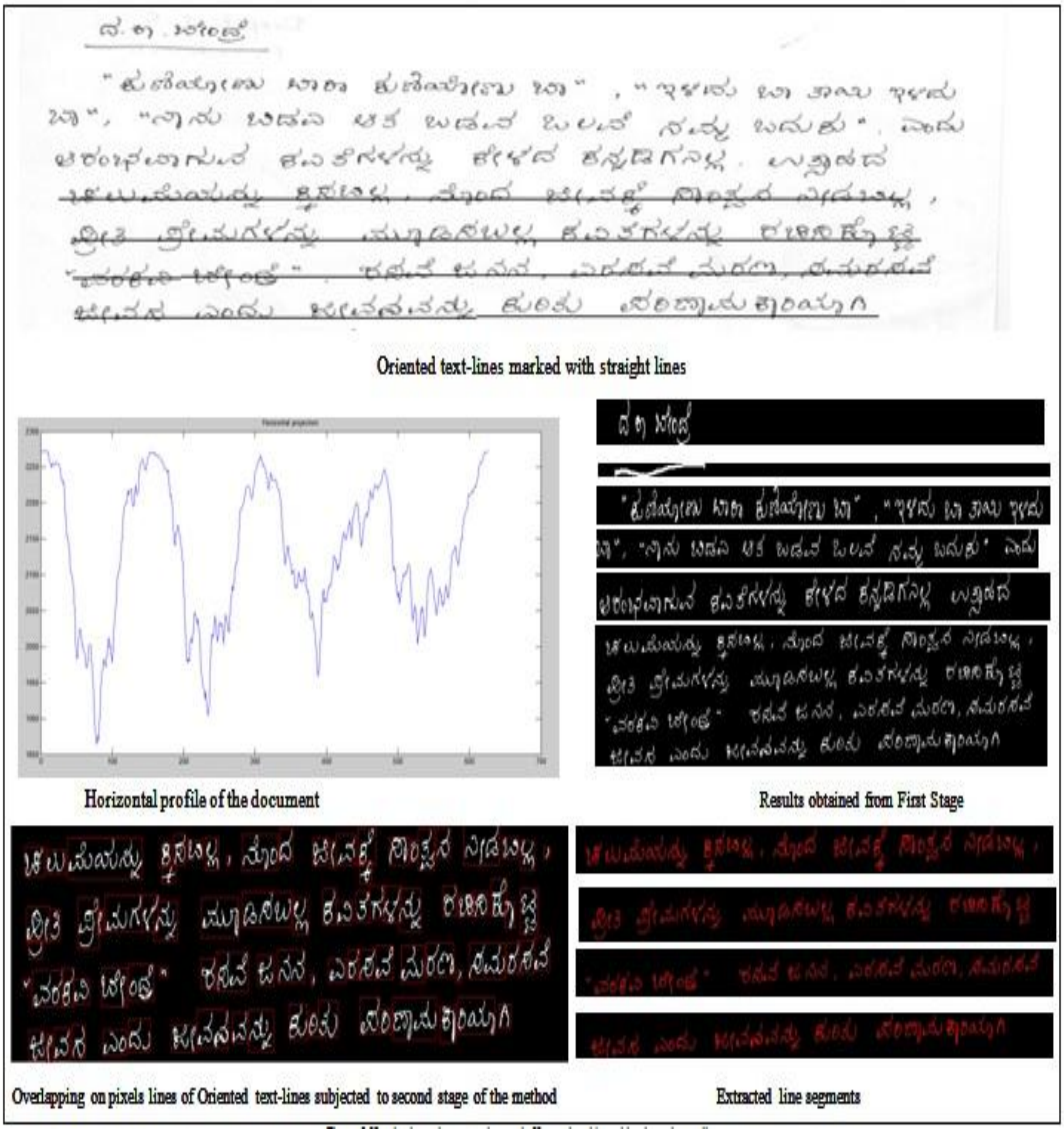

Telugu Handwritten document
Horizontal profile of the document
Extracted text-lines of Telugu document
Telugu Handwritten document
Horizontal profile of the document
Result Obtained from First Stage
Telugu Handwritten document with Grid
Overlapping on pixels lines of oriented text-lines subjected to second stage of the method
Extracted line segments
English handwritten document with grid
Horizontal profile
Extracted text-lines

Fig3: Handwritten document image in English with oriented text-lines with Grid

3. RESULTS AND DISCUSSIONS

The experiments are performed on handwritten scanned documents written in scripts namely, Kannada, Telugu, Hindi, English and Malayalam and these documents were considered for performing experiments. The preprocessed handwritten document image was subjected to two stage approach for line segmentation and extracted lines were stored in a separate folder. The results for few of the documents are shown in Fig. 1 through Fig4. It is observed that, the handwritten documents having text-lines separated from each other, evident from their horizontal profiles, were extracted at stage 1. Failure cases were observed from the resulting segments. The proposed algorithm failed to segment out lines from the documents wherein the adjacent lines had character(s) of above line touch the character(s) of below line. Figure 3 shows results of such document images.


Oriented text-lines marked with straight lines

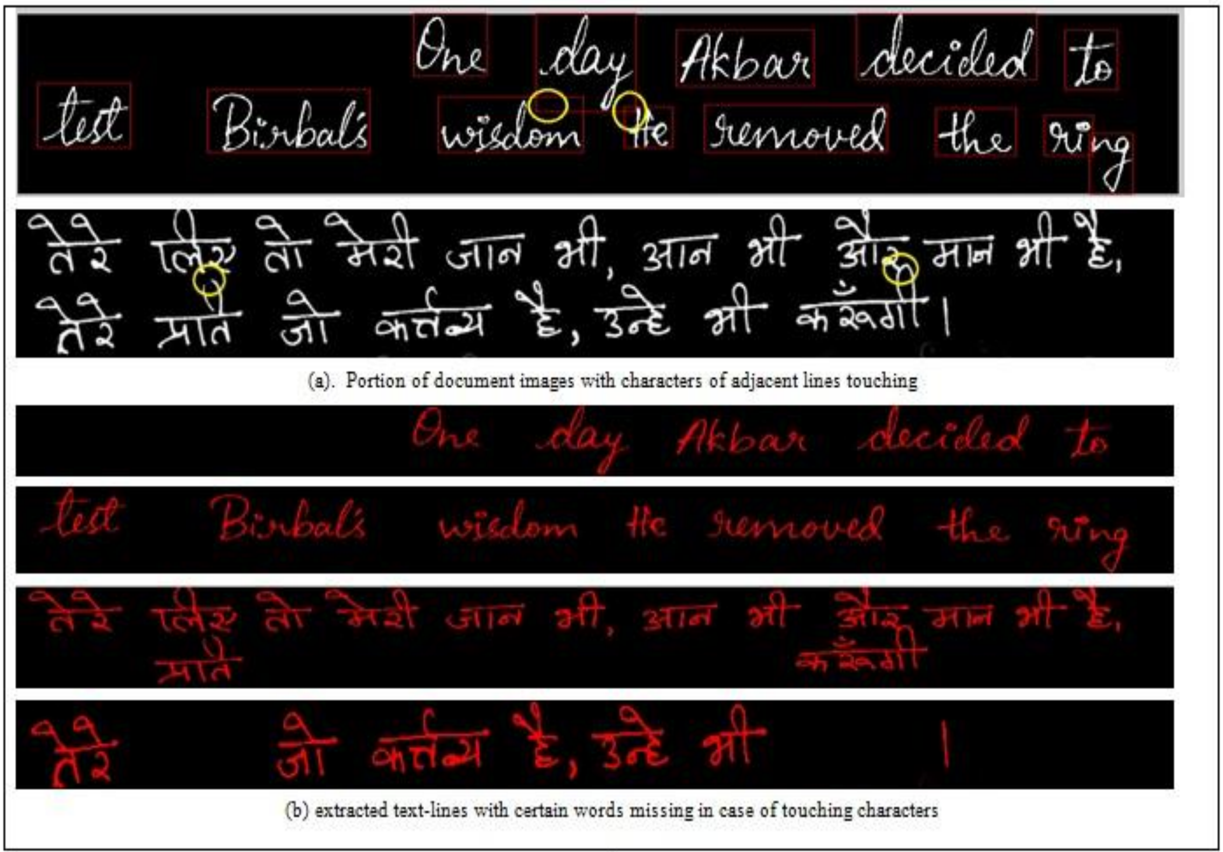
Horizontal profile of the document

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


One day Akbar decided to
test Birbal's wisdom He removed the ring
तेरे लिए तो मेरी जान भी, आन भी और मान भी है,
तेरे प्राति जो कर्तव्य है, उन्हे भी करूँगी।

(a). Portion of document images with characters of adjacent lines touching

(b) extracted text-lines with certain words missing in case of touching characters

Fig 5: Sample images with touching characters and results after stage 2 of the proposed method

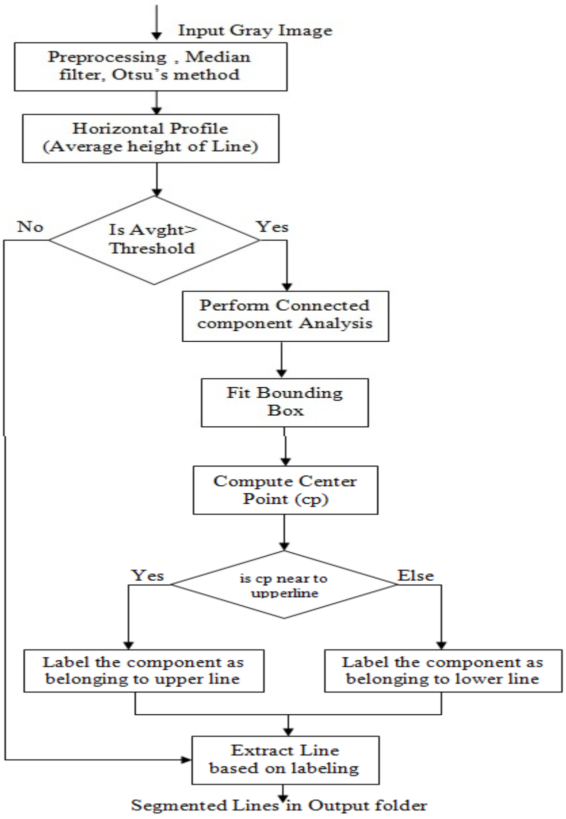

Input Gray Image
Preprocessing , Median filter, Otsu's method
Horizontal Profile (Average height of Line)
Is Avght> Threshold
No
Yes
Perform Connected component Analysis
Fit Bounding Box
Compute Center Point (cp)
is cp near to upperline
Yes
Else
Label the component as belonging to upper line
Label the component as belonging to lower line
Extract Line based on labeling
Segmented Lines in Output folder

Fig 6: Block diagram of the proposed method

4. CONCLUSION

An efficient two stage method has been proposed to extract text-lines from handwritten document images. Extraction of text-lines from document images with lines appearing curved (oriented) poses difficulty in segmentation. Connected Component Analysis is used on adjacent lines, in the second stage, wherein on pixel lines of these text-lines overlap. Bounding box is inserted on the component 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 documents written in different scripts and the results obtained are encouraging. The proposed method also extracts the lines with touching characters. However, in certain cases the resulting segmentation is not accurate. Our future aim is to improve upon the proposed method to extract text-lines accurately in case of touching characters.

5. REFERENCES

[1] FEI YIN, CHENG-LIN LIU. "Handwritten Text-line Extraction Based on Minimum Spanning Tree Clustering". Proceedings of the 2007 International Conference on Wavelet Analysis and Pattern Recognition, Beijing, China, 2-4 Nov. 2007.

[2] Yin, Fei, and Cheng-Lin Liu. "Handwritten Chinese text-line segmentation by clustering with distance metric learning." Pattern Recognition 42.12 (2009): 3146-3157.

[3] Saabni, Raid, and Jihad El-Sana. "Language-independent text-lines extraction using seam carving." Document Analysis and Recognition (ICDAR), 2011 International Conference on. IEEE, 2011.

[4] 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.

[5] 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

[6] Gomathi@ Rohini.S, Umadevi.R.S and Mohanavel.S. “Segmentation of Touching, Overlapping, Skewed and Short Handwritten Text-lines”. International Journal of Computer Applications (0975 – 8887) Volume 49– No.19, July 2012

[7] 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.

[8] 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.

[9] Zhinxin Shi, SrirangarajSetlur and VenuGovindraju.“A Steerable Directional Local Profile Technique for Extraction of Handwritten Arabic Text-lines”. 2009 10th International Conference on Document Analysis and Recognition.

[10] 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.

[11] SaiprakashPalakollu, 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.

[12] SaiprakashPalakollu, 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.

[13] 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).

[14] 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).

[15] 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

[16] Samir Malkarao et al and Nibran et al “ Text-line extraction from handwritten document pages using spiral run length smearing algorithm”.978-1-4673-4698-6 ©2012 IEEE.

[17] NazihOuwayed, 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)".

[18] SaiprakashPalakollu, 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.

[19] Kumar, Jayant, et al. "Segmentation of handwritten textlines in presence of touching components." Document Analysis and Recognition (ICDAR), 2011 International Conference on. IEEE, 2011.

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

[21] Ram Sarkar et al. “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.

[22] 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.