Text Line Segmentation of Handwritten Documents using Clustering Method based on Thresholding Approach

M.Ravi Kumar
Assistant Professor
Dept. of Computer Science,
Kuvempu University
Shankaraghatta-577451

Nayana N Shetty
Student
Dept of Computer Science,
Kuvempu University
Shankaraghatta-577451

B.P.Pragathi
Student
Dept of Computer Science,
Kuvempu University
Shankaraghatta-577451

ABSTRACT
Segmentation of the text lines in an un-constrained handwritten documents still a challenging task because handwritten text lines are often un-uniformly skewed and curved, and the space between lines is not obvious. In this paper, we propose a text-line segmentation algorithm based on clustering using threshold. The connected components of document image are grouped, from which text-lines are extracted dynamically by coloring all the text-lines.

Keywords
Handwritten document, segmentation, Bounding Box, Threshold, clustering.

1. INTRODUCTION
Document image analysis involves the tasks of text block segmentation, text line separation, character segmentation and recognition. Text line extraction is generally seen as a preprocessing step for tasks such as document structure extraction, printed character or handwriting recognition.
The segmentation of handwritten documents has also been addressed with the segmentation of address blocks on envelopes and mail pieces [3], and for authentication or recognition purposes[5]. More recently, the development of handwritten text databases (IAM database, [6]) provides new material for handwritten page segmentation. Some efforts have been devoted to the difficult problems of handwritten text line segmentation [8] . The approaches can be roughly categorized into top-down and bottom-up ones. Top-down methods partition the document image recursively into text regions, text lines, and words/characters. Bottom-up methods group small units of image (pixels, connected components, characters, words, etc.) into text lines and then text regions. Bottom-up grouping can be viewed as a clustering process, which aggregates image components according to proximity and does not rely on the assumption of straight lines. Both top-down and bottom-up methods have their disadvantages. Top-down methods do not perform well on curved and overlapping text lines. Bottom-up grouping is more complicated in computation than top-down partitioning, and its erformance relies on some heuristic rules or artificial parameters, such as the between component distance metric for clustering. Moreover, the spaces between handwritten text lines are often not obvious compared to the spaces between within-line characters, and some text lines may interfere with each other. Methods partition the document image recursively into text regions, text lines, and words/characters with the assumption of straight lines. Bottom-up methods group small units of image (pixels, connected components, characters, words, etc.) into text lines and then text regions. Bottom-up grouping can be viewed as a clustering process, which aggregates image components according to proximity and does not rely on the assumption of straight lines. Hybrid methods combine bottom-up grouping and top- down partitioning in different ways.
.

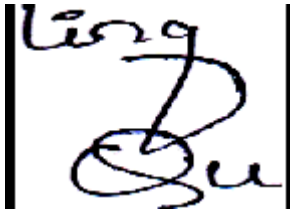
2. RELATED WORK
Document structure is a hierarchy of text regions, text lines, words, characters and connected components. Text lines can be extracted by either top-down region partitioning or bottom-up component aggregation. Some representative top-down and bottom-up segmentation Methods are reviewed below. The X-Y cut algorithm [10] is a projection-based top-down Segmentation method but performs well only on printed documents because of the assumption of parallel text lines and large between-line gaps. Some researchers modified the projection-based method to deal with slightly curved text lines. To do this, the document image is partitioned into several vertical strips [11]. The text lines in each strip (assumed to be approximately straight) are extracted according to horizontal projection profiles and then connected with the lines of other strips by heuristic rules. Zahour et al[12] proposed a partial projection-based method combined with slant detection and partial contour tracing. From a different viewpoint, several researchers proposed smearing-based top down methods. Shi et al [13] use an adaptive local connectivity map (ALCM), in which the value of each pixel is the sum of all pixels in the original image within a specified horizontal distance [13]. After thresholding the smeared image, the connected components then represent probable regions of text lines. Kennard and Barrett use a similar method with slight extension [14] to deal with free-form handwritten historical documents. The recently proposed level set based method[15] Is an effective top-down approach for unconstrained handwritten documents. On converting a binary image to gray-scaled, the level set method is exploited to determine the boundary between neighboring text lines. An obvious flaw of this algorithm is its high computation complexity. The Docstrums (document structure) method of O'Gorman [16] is typical of bottom-up grouping. It merges neighboring connected components using rules based on the geometric relationship between K nearest neighbor units, and performs well on printed documents as well as slightly curved handwritten documents. Likforman-Sulem and Faure [17] developed an iterative method based on perceptual grouping using three Gestalt criteria, namely, proximity, similarity and direction continuity ,to group connected components. The grouping of components to text lines can be considered as a clustering problem, and has been treated using minimal spanning tree (MST) clustering [19]. The performance of clustering relies on the distance metric between components. The Hough transform algorithm has also been

applied to handwritten text line detection [21], with the gravity centers or minima points of connected components as the points to be fitted, but needs a sophisticated post-processing procedure to extract the lines. On the other hand, [22] piece-wise projections are sensitive to character's size variation within text lines and significant gaps between successive words. These occurrences influence the effectiveness of smearing methods too. In such cases, the results of two adjacent zones may be ambiguous, affecting the drawing of text- line separators along the document width. To deal with these problems we introduce a smooth version of the projection profiles to over segment each zone into candidate text and gap regions. Then, we reclassify these regions by applying an Hidden Markov Model (HMM) formulation that enhances statistics from the whole document page. Starting from left and moving to the right we combine separators of consecutive zones considering their proximity and the local foreground density. These piece-wise [23] projection based methods have a few shortcomings: (a) they generate too many potential separating lines, (b) the parameter of stripe width is predefined, (c) text-lines should not have significant skew as mentioned in and (d) if there is no potential piece-wise line in the first and last stripes, drawing a complete separating line will become impossible in the any algorithms. These shortcomings are observed based on an experiment that we have conducted with a number of text-pages. Some authors also used skew information for text-line separation. In an unconstrained handwritten text-page, it is very difficult to detect the orientation of each line on the basis of the skew calculated for the entire page. Therefore, these methods may not work properly. The concept of the Hough transform [24] is employed in the field of document analysis for many purposes such as skew detection, line detection, slant detection and text-line segmentation. The Hough transform is employed for text-line segmentation in different scripts. A block-based Hough transform is presented which is a modification of the conventional Hough transform methodology. The algorithm includes partitioning of the connected component domain into three spatial sub- domains and applying a block-based Hough transform to detect the potential text lines. Many efforts have been devoted to the difficult problem of handwritten text line segmentation [25]. The methods can be roughly categorized into three classes: top-down, bottom-up, and hybrid.Top-down methods partition the document image recursively into text regions, text lines, and words/characters with the assumption of straight lines. Bottom-up methods group small units of image (pixels, connected components, characters, words, etc.) into text lines and then text regions. Bottom-up grouping can be viewed as a clustering process, which aggregates image components according to proximity and does not rely on the assumption of straight lines. Hybrid methods combine bottom-up grouping and top-down partitioning in different ways. All the three approaches have their disadvantages. Top-down methods do not perform well on curved and overlapping text lines. The performance of bottom-up grouping relies on some heuristic rules or artificial parameters, such as the between-component distance metric for clustering. On the other hand, hybrid methods are complicated in computation, and the design of a robust combination scheme is non-trivial.

A thinning operation [23] has also been used by other researchers for text-line segmentation of Japanese and Indian text documents. Thinning algorithms followed by post-processing operations are employed for the entire background region of an input text image to detect the separating borderlines. Recently, some techniques have used level set, active contour and a variational Bayes method for text-line segmentation. Density estimation and the level set method (LSM) were utilized for text- line segmentation. A probability map is estimated from an input document image, where each element represents the probability of the original pixel belonging to a text line. The level set method (LSM) is utilized to determine the boundary evolution of neighboring text lines. At first, a matched filter bank approach is used for smoothing the input text image. The central line of text- line components is then computed using ridges over the smoothed image. Finally, the active contours (snakes) over the ridges are adapted to obtain the text-line segmentation result.

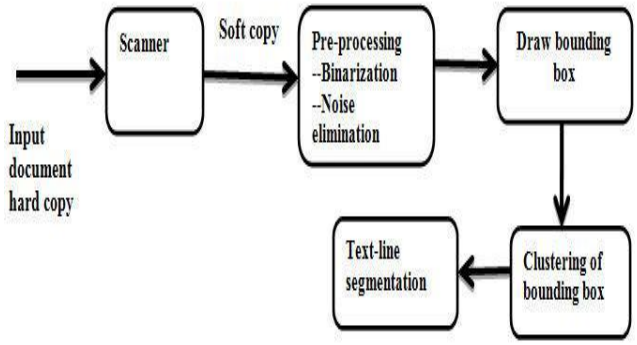
3. CHALLENGES

Smearing methods can't deal well with touching and overlapping components.



Horizontal projections can't deal well with skewed, curved and fluctuating lines. Bottom-up approaches: connected components or even pixels are connected to their close ones based on geometrical criteria to form text lines [26]. Due to many challenges in text line segmentation, although many methods have been proposed, the problem still remains open. Various methods on text line segmentation doesn't works for multiple skew lines. The above challenges in analyzing the handwritten documents are the motivation to take this work.

4. BLOCK DIAGRAM OF PROPOSED METHOD



5. PROPOSED WORK

In the proposed method, we present a method for segmentation of text-lines of hand written documents using clustering based on thresholding approach. The method consists of three stages,(i) Drawing bounding box ,(ii) Clustering of Bounding box and (iii) Text line segmentation.

5.1 Drawing bounding box

The pre-processing operations like Binarization and Noise elimination is done by using morphological operations. After removing the noise, the groupings of wards are done by drawing the bounding box. It is because the connect components are easily identified and it helps in clustering.

5.2 Clustering

In this step, we cluster the different bounding box by measuring the area covered by each bounding box. Here we compare the distance between two neighbor bounding boxes for threshold value. The two bounding boxes are

combined by calculating the threshold value. This threshold value groups the different bounding boxes into one cluster. As clusters are formed, it helps in text line segmentation.

5.3Text-line segmentation

Since the different clusters are obtained in previous step, it is easy to segment the text-line by assigning the different colors for different clusters. The coloring on clusters is done by assigning the various RGB intensity values.

6. RESULTS

We have conducted the experiments on the languages like Kannada and English of 200 documents (Kannada- 100 and English-100) and obtain the accurate result.

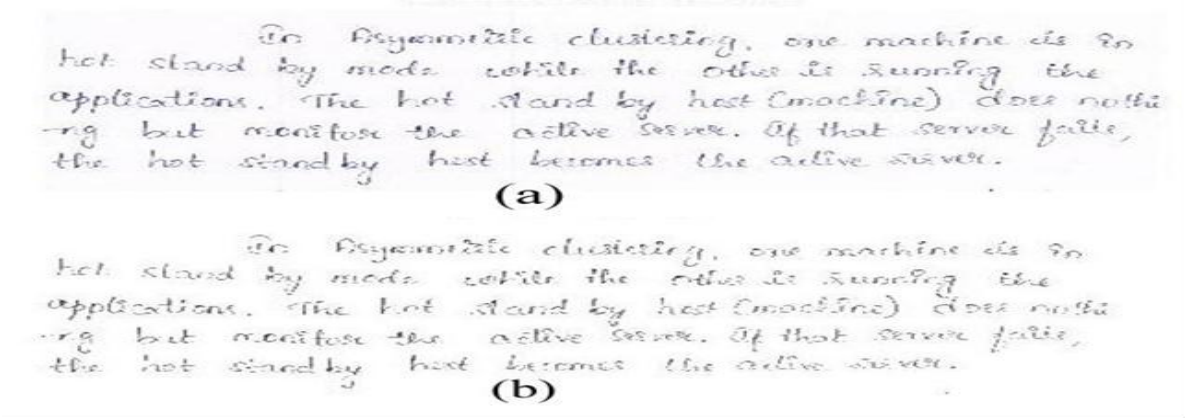


Figure (1): Illustration of Noise removal. (a) Original image with noise (b) Image without noise

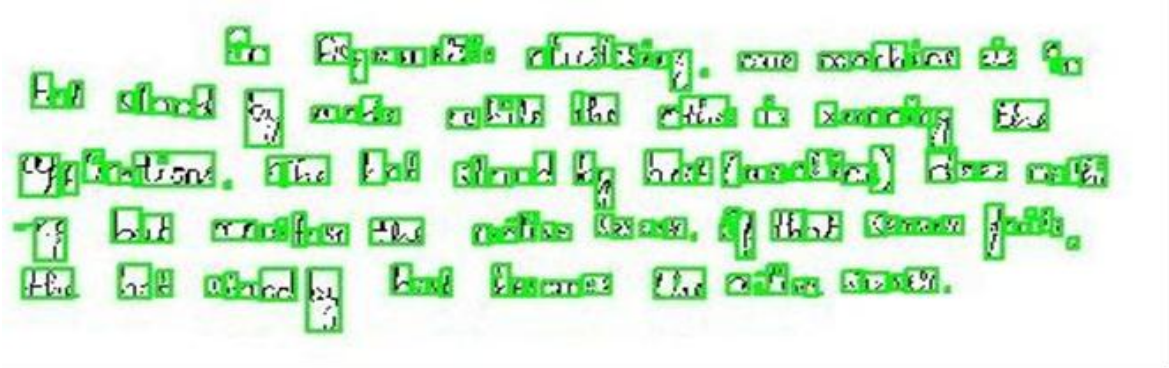


Figure (2): Illustration of drawing bounding box

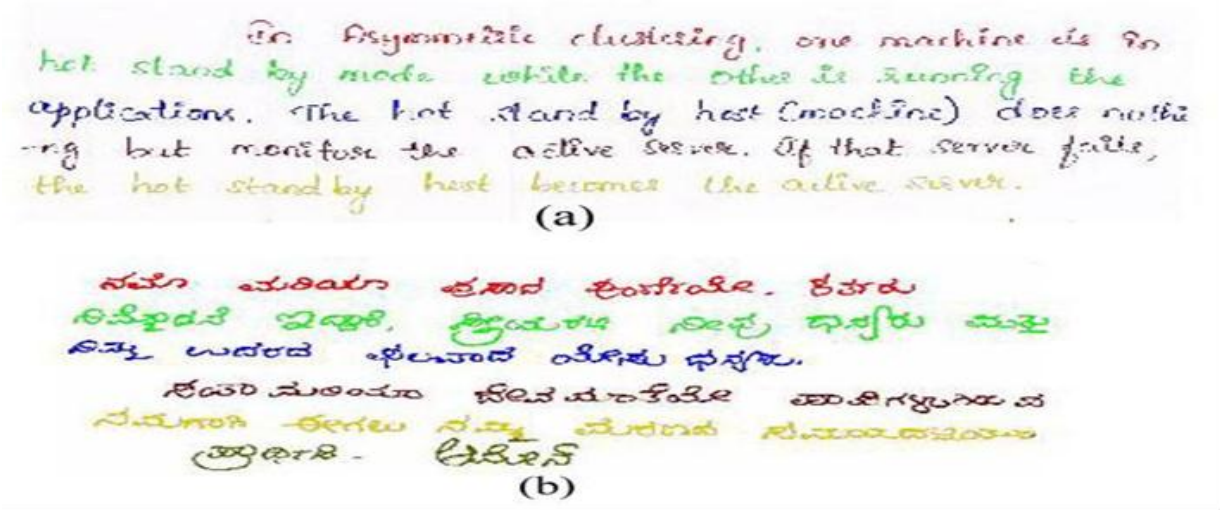


Figure (3): Illustration of Text-line Segmentation. (a) English Document (b) Kannada Document

7. CONCLUSION AND FUTURE WORK

The proposed method extracts the text lines of a document accurately. In this work, we are mainly concentrated on the document with single languages and proposed method is very simple our method works if the text lines are separated each other, otherwise accuracy will be reduced. In future we work on the documents with more than one language and the document contains the text lines which are not separated well.

8. ACKNOWLEDGEMENT

The authors would like to thank Mr.Pradeep.R, Prasad Babu and B.S Puneeth Kumar, Dept. of P. G Studies in Computer Science, Kuvempu University, for their help during this work

9. REFERENCES

[1] Downton A., Leedham C. G. (1990), Preprocessing and presorting of envelope images for automatic sorting using OCR, Pattern Recognition, 23(3-4):347-362.

[2] Govindaraju V., R. Srihari, S. Srihari (1994), Handwritten text recognition, Document Analysis Systems DAS

[3] Seni G., Cohen E. (1994), External word segmentation of off-line handwritten text line, pattern Recognition, 27, Issue 1, January, pp 41-52

[4] Srihari S., Kim G. (1997), Penman: a system for reading unconstrained handwritten page image, SDIUT 97, Symposium on document image understanding technology, pp. 142-153.

[5] Zhang B., Srihari S.N., Huang C. (2004), Word image retrieval using binary features, SPIE Conference on Document Recognition and retrieval XI, San Jose,California, USA, Jan 18-22. 2. Antonacopoulos A. (1994), Flexible Page Segmentation Using the Background, Proc. 12th Int. Conf. on Pattern Recognition (12th ICPR), Jerusalem, Israel, October 9-12, vol. 2, pp.339-344.

[6] Marti U., Bunke H. (1999), A full English sentence database for off-line handwriting recognition, Proc. 5th

[7] F. Yin, C.L. Liu,(2007), Handwritten text line extraction based on minimal spanning tree clustering, Proc. 5th Int. Conf. on Wavelet Analysis and Pattern Recognition, Vol.3, pp. 1123-1128.

[8] F.Chang,C.J.Chen,C.J.Lu,A linear-time component labeling algorithm using contour tracing technique, Computer Vision and Image Understanding, Vol.93, pp.206-220, 2004.

[9] Fei Yin, Cheng-Lin Liu, Handwritten Text Line Segmentation by Clustering with Distance Metric Learning, National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences

[10] G.Nagy, S.Seth, M.Viswanathan,(1992), A prototype document image analysis system for technical journals, Computer, Vol.25, pp. 10-22.

[11] U.Pal, S. Datta,(2003), Segmentation of Bangla unconstrained handwritten text, Proc.7th Int. Conf. on Document Analysis and Recognition, Vol.2, pp. 1128-1132.

[12] A.Zahour, B.Taconet, P. Mercy,S. Ramdane,(2001),Arabic handwritten text-line extraction, Proc 6th Int. Conf. on Document Analysis and Recognition, pp. 281-285.

[13] Z. Shi, S. Setlur, V. Govindaraju, (2005),Text extraction from gray scale historical document image using adaptive local connectivity map, Proc. 8th Int.Conf. on Document Analysis and Recognition, Vol.2, pp. 794-798.

[14] D.J. Kennard, W.A. Barrett,(2006), Separating lines of text in freeform handwritten historical documents, Proc. 2nd Int. Conf. on Document Image Analysis for Libraries, pp. 12-23.

[15] Y.Li, Y.Zheng, D.Doermann, S. Jaeger,(2008), Script independent text line segmentation in freestyle handwritten document, IEEE Trans.Pattern Analysis and Machine Intelligence, to appear.

[16] L O'Gorman, (1993),The document spectrum for page layout analysis, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol.15, No.11, pp. 1162-1173.

[17] L.Likforman-Sulem,(1994), C. Faure, Extracting lines on handwritten document by perceptual grouping,In: Advances in Handwriting and Drawing: A Multidisciplinary Approach, pp . 21-38.

[18] I.S.I. Abuhaiba,S.Datta,(1995),M.J.J. Holt, Line extraction and stroke ordering of text pages, Proc. 3rd Int. Conf. on Document Analysis and Recognition, Vol.1, pp. 390-393.

[19] A.Simon, J.-C. Pret , A.P. Johnson,(1997), A fast algorithm for bottom-up document layout analysis, IEEE Trans. Pattern Analysis and Machine Intelligence,Vol.19, No.3, pp. 273-277.

[20] Y.Pu,Z.Shi,(1998), A natural learning algorithm based on Hough transform for text lines extraction in handwritten document, Proc. 6th Int. Workshop on Frontiers in Handwriting