A Simple Text-line segmentation Method for Handwritten Documents

M.Ravi Kumar
Assistant professor
Dept. of P.G.Studies in Computer Science Kuvempu University Jnana Sahyadri, Shankaraghatta-577451

R. Pradeep
Dept. of P.G.Studies in Computer Science Kuvempu University Jnana Sahyadri, Shankaraghatta-577451

B.S.Puneeth Kumar
Dept. of P.G.Studies in Computer Science Kuvempu University Jnana Sahyadri, Shankaraghatta-577451

Prasad Babu
Dept. of P.G.Studies in Computer Science Kuvempu University Jnana Sahyadri, Shankaraghatta-5774514th

ABSTRACT

Text line segmentation is an important step because inaccurately segmented text lines will cause errors in the recognition stage.. The nature of handwriting makes the process of text line segmentation very challenging. Text characteristics can vary in font, size, orientation, alignment, color, contrast, and background information. These variations turn the process of word detection complex and difficult. Since handwritten text can vary greatly depending on the user skills, disposition and cultural background. In this work we have proposed the method which works on the different intensity values for extracting the text-lines.

General Terms

Image processing, Handwritten document analysis

Keywords

Handwritten document, segmentation, text-lines

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, handwriting recognition.

Segmentation of a document image into its basic entities, namely, text lines and words, is considered as a non-trivial problem to solve in the field of handwritten document recognition. The difficulties that arise in handwritten documents make the segmentation procedure a challenging task. Different types of difficulties are encountered in the text line segmentation and the word segmentation procedure. In case of text line segmentation procedure, major difficulties include the difference in the skew angle between lines on the page or even along the same text line, overlapping words and adjacent text lines touching. Furthermore, the frequent appearance of accents in many languages (e.g. French, Greek) makes the text line segmentation a challenging task. In word segmentation, difficulties that arise include the appearance of skew in the text line, the existence of punctuation marks along the text line and the non- uniform spacing of words which is a common residual in handwritten documents [1]. Text line segmentation from document images is one of the major problems in document image analysis. It provides crucial information for the tasks of text block segmentation, character segmentation and recognition, and text string recognition. Whereas the difficulty of machine-printed document analysis mainly lies in the complex layout structure and degraded image quality, handwritten document analysis is difficult mainly due to the irregularity of lay- out and character shapes originated from the variability of writing styles. For unconstrained handwritten documents, text line segmentation and character segmentation-recognition are not solved though enormous efforts have been devoted to them and great advances have been made. Text line segmentation of handwritten documents is much more difficult than that of printed documents. Unlike that printed documents have approximately straight and parallel text lines, the lines in handwritten documents are often un-uniformly skewed and curved. 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. Therefore, many text line detection techniques, such as projection analysis and K-nearest neighbor connected components (CCs) grouping, are not able to segment handwritten text lines successfully[2].

On the other hand, 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 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 [3].

These piece-wise 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 [4].

The concept of the Hough transform 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 [5].

Many efforts have been devoted to the difficult problem of hand- written text line segmentation. 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, CCs, 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 [2].

A thinning operation 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 [4].

2. RELATED WORK

In this section, we give a brief review of recent work on text line and word segmentation in handwritten document images. As far as we know, the following techniques either achieved the best results in the corresponding test datasets, or are elements of integrated systems for specific tasks. One of the most accurate methods uses piece-wise projection pro- files to obtain an initial set of candidate lines and bivariate Gaussian densities to assign overlapping CCs into text lines [7]. Experimental results on a collection of 720 documents (English, Arabic and children's handwriting) show that 97.31% of text lines were segmented correctly. The writers mention that “a more intelligent approach to cut an overlapping component is the goal of future work”. A recent approach [8] uses block-based Hough transform to detect lines and merging methods to correct false alarms. Although the algorithm achieves a 93.1% detection rate and a 96% recognition rate, it is not flexible to follow variation of skew angle along the same text line and not very precise in the assignment of accents to text lines. Li et al. [9] discuss the text-line detection task as an image segmentation problem. They use a Gaussian window to convert a binary image into a smooth gray-scale. Then they adopt the level set method to evolve text-line boundaries and finally, geometrical constrains are imposed to group CCs or segments as text lines. They report pixel-level hit rates varying from 92% to 98% on different scripts and mention that “the major failures happen because two neighboring text lines touch each other significantly”. A similar method [10] evaluates eight different spatial measures between pairs of CCs to locate words in handwritten postal ad- dresses. The best metric proved to be the one which combines the result of the minimum run-length method and the vertical overlapping of two successive CCs. Additionally, this metric is adjusted by utilizing the results of a punctuation detection algorithm (periods and commas). Then, a suitable threshold is computed by an iterative procedure. The algorithm tested on 1000 address images and performed an error rate of about 10%. Manmatha and Rothfeder [11] propose an effective for noisy historical documents scale space approach. The line image is filtered with an anisotropic Laplacian at several scales in order to produce blobs which correspond to portions of characters at small scales and to words at larger scales. The optimum scale is estimated by three different techniques (line height, page averaging and free search) from which the line height showed best results. Much more challenging task is line segmentation in historical documents due to a great deal of noise. Feldbach

and Tonnies [12] have proposed a bottom up method for historical church documents that requires parameters to be set according to the type of handwriting. They report a 90% correct segmentation rate for constant parameter values which rises to 97% for adjusted ones.

Another integrated system for such documents [13] creates a foreground/background transition count map to find probable locations of text lines and applies min-cut/max-flow algorithm to separate initially connected text lines .The method performs high accuracy (over 98%) in 20 images of George Washington's manuscript.

3. SEGMENTATION CHALLENGES

In this section we the challenges involved in the segmentation of the text-lines. When dealing with handwritten text, line segmentation has to solve some obstacles that are uncommon in modern printed text [6]. Among the most predominant are

3.1 Skewed Lines

Lines of text in general are not straight.These lines are not parallel to each other

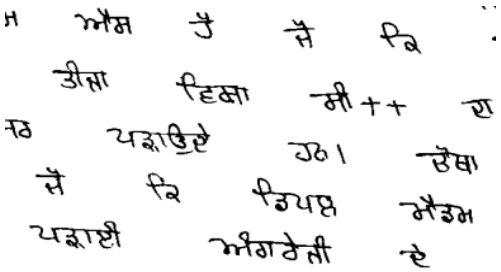


Figure 3.1: Skewed Lines [6]

3.2 Fluctuating lines

Lines of text are partially or fully connected to other text-lines

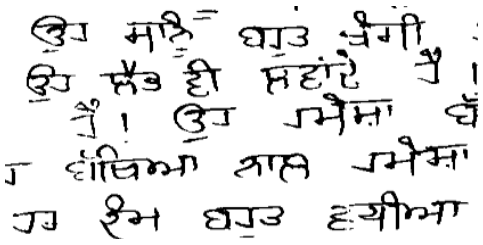


Figure 3.2: Fluctuating Lines [6]

3.3 Line Proximity

Small gaps between neighboring text lines will cause touching and overlapping of components, usually words or letters, between lines and irregularity in geometrical properties of the line, such as line width, height, distance in between words and lines, leftmost position etc.

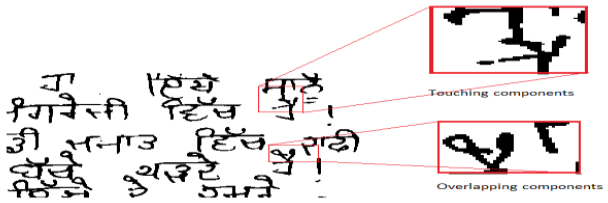


Figure 3.3: Line proximity [6]

4. MOTIVATION

The above challenges are motivated us to take this challenging work.

5. PROPOSED METHOD

In the proposed method our objective is to identify the boundary of the text-lines, which consist two steps (i) Generating partial boundary line and (ii) Generating complete boundary line. By using these partial lines it is very difficult to differentiate between the two lines. Because these partial lines are having gaps and they are broken. Thus we need to construct the complete line which acts as differentiator for identifying the text-lines. In constructing the complete line we will consider the highest frequent y co-ordinate value in each partial boundary lines, y co-ordinate values are the complete boundaries for each text-lines. In last step we segment the each text-line by representing the each text-line with different colors.

5.1 Generating Partial Text-Line Boundary

The partial text-line boundary is generated by blocking the text-lines. Using morphological operations like erosion. Here blocking is done for filling the holes and gaps between the words. This helps in drawing the partial line. Here the lines are drawn at the edge of the every blocked text-line. However these partial lines are not sufficient to differentiate the text-lines. This is shown in figure 5.1.

5.2 Generating Complete Text-Line Boundary

As in the previous step we obtained the partial boundary lines, these broken boundary lines are not sufficient for segmenting the text-lines, thus we need to generate complete line which is continuous. So now by using these partial lines, we draw the complete boundary lines with the help of the frequent vertical points. These y co-ordinate values are used to differentiate between text-lines, and these complete boundary lines helps in segmentation. This is showed in figure 5.2

5.3 Text-Line Segmentation

Once the complete boundary lines were drawn, it is easy to segment the text-lines by assigning different values to each character in between the lines. In this section, we give the different intensity values for the characters in a text-line between the two drawn lines. Through the different intensity values of the characters in a text-line, this method recognizes

the different lines. The segmentation for different languages are showed in figure 5.3

Figure 5.1: Illustration of Partial text-line boundary detection. (a) A handwritten document. (b) Blocked text-lines. (c) Partial boundary lines

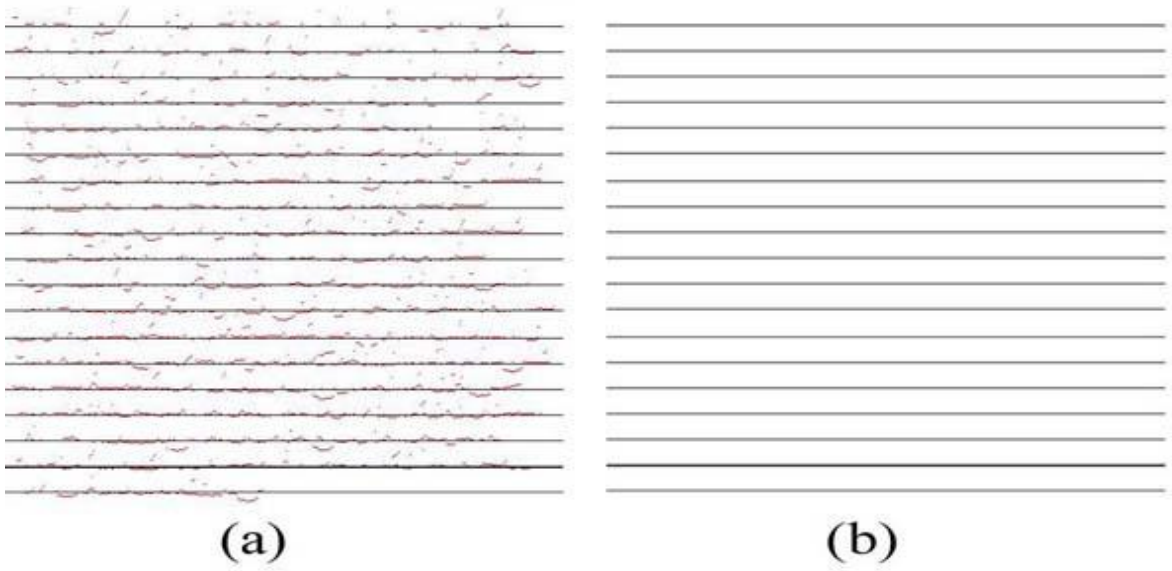
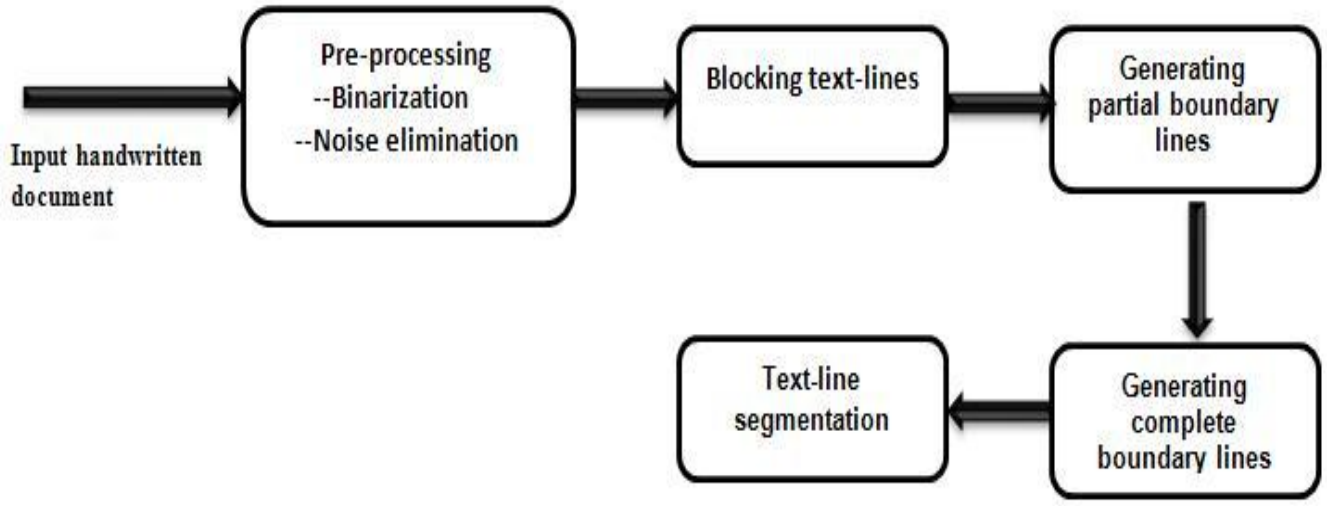


Figure 5.2: Illustration of complete text-line boundary detection. (a) Complete boundary lines with partial boundary lines. (b) Complete boundary lines without partial boundary lines

Figure 5.3: Illustration of segmented text-lines. (a) Segmented English document. (b) Segmented Kannada document (c) Segmented Hindi document (d) Segmented English document

6. BLOCK DIAGRAM OF PROPOSED METHOD



7. RESULTS AND DISCUSSION

We conducted the experiments on the modern set of handwritten documents of languages like, Kannada, English, Hindi and Arabic of 500 documents (kannada-150, English-200, Hindi-100 and Arabic-50) and obtained the accurate results. The accuracy of the proposed method is 98 %. In the proposed method, the variable structuring element is taken and the structuring element is varies from language to language. The proposed method works well only if the documents containing a single language, i.e monolingual documents. Proposed method works well if all the lines in the document are well separated. i.e. no two lines are touching each other. Otherwise the performance of the proposed method will be reduced and the accuracy will also be reduced. Different colors are assigned for different lines for clear separation of text lines

8. CONCLUSION AND FUTURE WORK

In this work we are mainly concentrated on extracting required text-line from the given document and obtained the specified text-line accurately. The proposed method works

well for segmenting the text-line of handwritten document. Our method only works on the fixed word length and text-lines without skew; in our future work we will improve the results by segmenting the text-lines with above described limitations. In the future work, we plan to conduct the experiments on multilingual documents. And also the method will be tested on unconstrained handwritten documents

9. REFERENCES

[1] G. Louloudisa, B.Gatosb, I.Pratikakisb, C.HalatsisaText line and word segmentation of handwritten documents. Pattern Recognition (2008) pp. 3169 – 3183.

[2] Fei Yin, Cheng-LinLiu. Handwritten Chinese text line segmentation by clustering with distance metric learning Pattern Recognition (2009) pp. 3146 -- 3157.

[3] Vassilis Papavassiliou, Themos Stafylakis, et al.., Handwritten document image segmentation into textlines and words. Pattern Recognition (2010) pp. 369 – 377

[4] Alireza Alaei UmapadaPal, et al..,. A new scheme for unconstrained handwritten text-line segmentation. Pattern Recognition (2011) pp. 917–928

[5] A Block-Based Hough Transform Mapping for Text Line Detection in Handwritten Documents G. Louloudis1, B. Gatos2, I. Pratikakis2, K. Halatsis1

[6] Ashu Kumar , Simpel Rani Jindal, Galaxy Singla Line segmentation using contour tracing, 2012 Vol 3

[7] Z. Razak, K. Zulkiflee, et al., Off-line handwriting text line segmentation: a review, International Journal of Computer Science and Network Security 8 (7) (2008) 12–20

[8] G. Louloudis, B. Gatos, C. Halatsis, Text line detection in unconstrained handwritten documents using a block-based Hough transform approach, in: Proceedings of International Conference on Document Analysis and Recognition, 2007, pp. 599–603

[9] Y. Li, Y. Zheng, D. Doermann, S. Jaeger, Script-independent text line segmentation in freestyle handwritten documents, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (8) (2008) 1313–1329

[10] G. Seni, E. Cohen, External word segmentation of off-line handwritten text lines, Pattern Recognition 27 (1994) 41–52.

[11] R. Manmatha, J.L. Rothfeder, A scale space approach for automatically segmenting words from historical handwritten documents, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (8) (2005) 1212–1225.

[12] M. Feldbach, K.D. Tonnies, Line detection and segmentation in historical church registers, in: Proceedings of International Conference on Document Analysis and Recognition, 2001, pp. 743–747.

[13] D.J. Kennard, W.A. Barrett, Separating lines of text in free-form handwritten historical documents, in: Proceedings of International Workshop on Document Image Analysis for Libraries, 2006, pp. 12–23.