Segmentation of Touching, Overlapping, Skewed and Short Handwritten Text Lines

Gomathi @ Rohini. S.
Sri Ramakrishna Engineering College, NGGO Colony PO, Coimbatore, India.

Uma Devi. R.S.
GRG School of Applied Computer Technology, Coimbatore, India.

Mohanavel. S.
Dr.N.G.P. Business School, Dr.N.G.P. Institute of Technology, Coimbatore, India.

ABSTRACT
Text line segmentation is an inherent part of document recognition system and important preprocessing step for word and character segmentation. Presence of touching or overlapping text lines, short-lines, curvilinear or skewed lines and small or variant gaps between the text lines make the segmentation challenging. These variations cause errors in recognition phase. This paper describes the top-down approach of handwritten text line segmentation. 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. The system was evaluated using 200 images from the IAM database and 100 documents collected from different writers. From the experiments conducted, it was observed that the system has 91.92% accuracy and imbibes in its reliability.

General Terms
Document Image Processing, Image Segmentation.

Keywords
Line Segmentation, Connected Component, Distance Metrics, Run Length.

1. INTRODUCTION
Text line segmentation is an important step in document image processing. There is no universally accepted solution in automatic handwritten document recognition systems. Segmentation of text lines could be a crucial step in document image processing tasks due to varying text characteristics, fluctuating lines and irregularity in geometrical properties of the line. It is a complicated and diverse problem by the nature of handwriting. Hence it represents a leading challenge in document image processing. Moreover, the touching of characters across the lines and overlapping spatial envelopes of text lines make the problem more demanding. The text lines in handwritten documents are often multi-skewed and curved and the space between the lines are not even. They can also vary greatly depending on the user skill, disposition and cultural background. Methods based on connected components are faster, but suffer from touching or close proximity of components.

This paper is organized as follows: Section 2 briefs the related works carried out by the researchers, Section 3 demonstrates the preprocessing and proposed segmentation method, Section 4 presents the experimental results and Section 5 concludes the paper with future works.

2. RELATED WORKS
Projection-based method is suitable for clearly separated lines. This method could cope with a few overlapping or touching characters. From the vertical projection profile, the gaps between the text lines in the vertical direction could be observed [7]. Short lines would provide low peaks and very narrow lines; and overlapping characters would not produce significant peaks [8].

The stochastic method seems to be suitable to separate overlapping characters, for which they could generate non linear segmentation paths and even more to derive non linear cutting paths from touching characters by identifying the shortest paths. But, it avoids cross overlapping characters and results in non linear paths turn around into obstacles. This method might fail, if the touching point contains a lot of black pixels [14].

Pixels of the image act as attractive forces for baselines and already extracted baselines act as repulsive forces. The baseline to extract is initialized under the previously examined one, in order to be repelled by it and attracted by the pixels of the line below [10]. Concerning text line fluctuations, baseline-based representations seem to fit naturally. More fluctuated the text line, the more refined local criteria must be. The lines must have similar lengths.

Grouping method consists of building alignments by aggregating units in a bottom-up strategy. Pixels lying within a given baseline and a median line are clustered in the corresponding text line; but ascenders and descenders are not segmented. The joining scheme relies on both local and global criteria, which are used for checking local and global consistency respectively [5].

In smearing method, consecutive black pixels along with the horizontal direction are smeared; i.e., the white space between them is filled with black pixels, if their distance metric is within a predefined threshold. The bounding boxes of the connected characters in the smeared image enclose the text lines [16].

In simple cases of handwritten pages, the center of gravity of the connected character is used either to associate the character to the current line or the following line or to cut the character into two parts. This works well, if the component is a single character [8]. It may fail, if the component is a word, part of a word or several words.

3. PROPOSED METHOD
The proposed method for line segmentation has the following two phases.

3.1 Preprocessing
A cursive handwritten text image is obtained from the IAM database. This grayscale image is converted into binary level representation using fixed threshold algorithm, where the pixels with value greater than the threshold are set to 0 (black) and the rest are set to 1 (white). Using standard java functions getHeight() and getWidth(), height and width of the image are read.

3.2 Line segmentation
The lines that do not share pixels with neither touching nor overlapping are segmented using horizontal projection

profile: and the lines that share pixels with neither touching nor overlapping are segmented using connected component. These techniques could not segment the following challenges.

(1) Touching and overlapping lines

(2) Skewed or slanting lines

(3) Short lines

This phase addresses the above issues and solutions for these issues are illustrated with examples.

3.2.1 Touching and overlapping lines

In this case, line segmentation begins with finding the core region for each line in the image. At first, horizontal projection profile is obtained for the preprocessed image. Core region is plotted in the constructed profile as the area from the point, where the pixel distribution is greater than the threshold to the point, where the pixel distribution is less than the threshold. Here trimmed mean is used as threshold value (Tm) (See eqn .1).

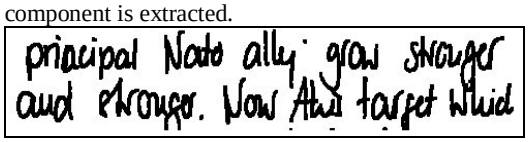
$$Tm = \frac{1}{width - 2k} \sum_{m=k+1}^{width-k} pix \longrightarrow 1$$

Where k is α*width and α is degree of freedom. A separator line is drawn at the center of successive core region of the adjacent text lines. Midpoint (MU/ML) lines (see Figure 1b) are drawn between the separator line and core region (Upper baseline / Lower baseline of adjacent text lines). Midpoint line between upper baseline and MU separator and midpoint line between lower baseline and ML separator are marked as segmentation points for the adjacent text lines. These lines are extracted by using standard java function getSubimage(), which takes segmentation point metrics as parameters. The segmented output will have noise due to touching or overlapping components (see Figure 1c and 1d).

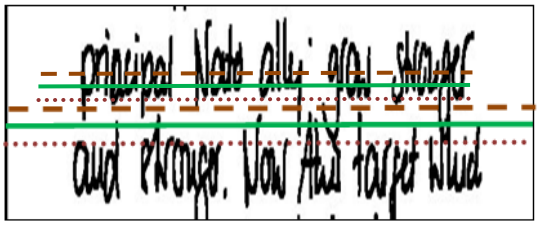
The overlapping may occur due to component with loop overlapping on component with or without loop. These overlapping areas are extracted (see Figure 1e).

In order to segment these components, following operations are performed.

1. Thinning is applied on the extracted overlapping component to capture the structural knowledge. This morphological operation is done using ImageJ tool.
2. The thinned image is extracted into two separate components using run-length. Applying run-length transition between foreground pixel to background pixel and vice-versa are captured. Some foreground pixels might be skipped, if their number does not exceed a pre-defined threshold. The threshold value is calculated as inter quartile range. Thus one of the components is extracted.
3. The processed image is subtracted from the thinned image obtained from the step-1. Thus the other component is extracted.

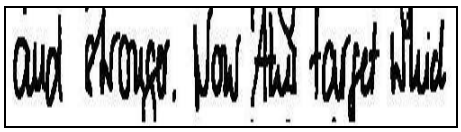


a) Input image with overlapping components

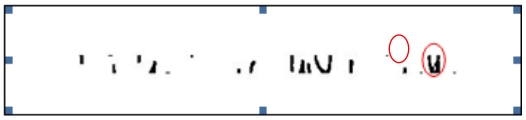


b) Separator line (solid line) with MU(dashed lines) and ML(dotted lines)

c) Output with disturbances

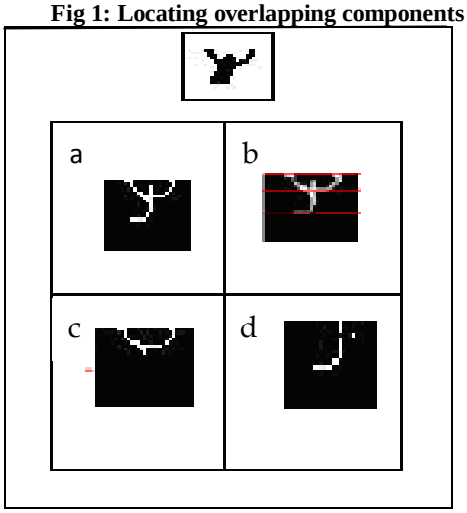


d) Output with disturbances



e) Locating overlapping area

Fig 1: Locating overlapping components



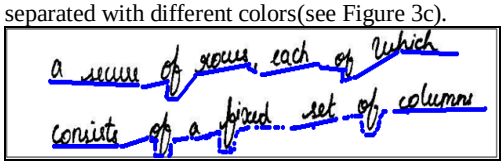
a) Output after thinning
b) Calculating run-length
c) Output after extracting one component using run-length
d) Output after extracting another component using subtraction

Fig 2: Segmentation of overlapping components

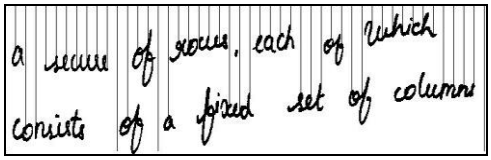
3.2.2 Skewed or slant lines

In this case, in order to segment the text lines from the image having irregular baselines (see Figure 3a), the following procedure is applied.

(1) At the top, a horizontal line is drawn from left to right of the image, where there is no black pixel hit.
(2) From left to right of the horizontal line, at regular intervals (say width of the line is 15 pixels), going from top to bottom, the distances from top extreme to the first black pixel hit are found (see Figure 3b) and its average is calculated.
(3) If the distance from the horizontal line to the first black pixel is lesser than the average, a line is drawn from the top extreme to this point (This is the skewed or slanting line identified). Otherwise this line is skipped. This step is repeated until the right end is reached.
(4) Now using 'connected component' script of MATLAB, these skewed or slanting lines are separated with different colors(see Figure 3c).



a) Input image with irregular baseline



b) Mapping distance metrics

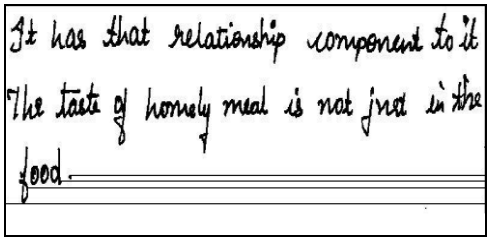
c)Segmenting lines within threshold

Fig 3: Segmentation of irregular baseline

3.2.3 Short lines

In this case, in order to segment the short text lines from the image (see Figure 4a), the following procedure is applied.

(1) At the right extreme, a vertical line is drawn from top to bottom of the image, where there is no black pixel hit.
(2) From top to bottom of the vertical line, at regular intervals (say height of the line is 15 pixels), going from right to left, the distances from the right extreme to the first black pixel hit are found and its average is calculated.
(3) If the distance from the vertical line to the first black pixel is greater than the average, a line is drawn from the right extreme to this point (This is the shorter line identified). Otherwise this line is skipped. This step is repeated until the bottom is reached.
(4) Now using 'connected component' script of MATLAB, these shorter lines are separated with different colors (see Figure 4b).



a) Input image with short lines

b) Shortline identification

Fig 4: Segmentation of short lines

4. EXPERIMENTAL RESULTS

The proposed segmentation system incorporates the above three cases. For all images, corresponding ground truth in terms of text lines is described and segmentation results were manually checked for errors. The experiments are performed on the handwritten documents, randomly selected from IAM database and collected samples. The experimental results of the proposed method are encouraging in giving accurate segmentation points for different line length and spacing (See Table 1 or Figure 5).

Table 1. Accuracy of segmentation

Line Type	Total No. of Lines	No. of Lines Correctly Segmented	Percentage of Accuracy
Touching and overlapping lines	1100	950	86.36%
Skewed or Slant lines	1500	1360	90.66%
Short lines	400	395	98.75%
Total	3000	2705	Average 91.92%

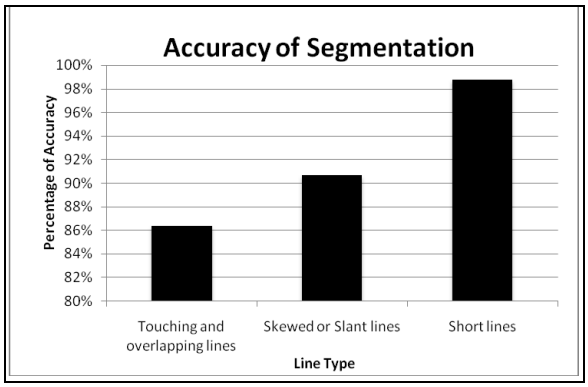


Fig.5 Accuracy of segmentation

5. CONCLUSION AND FUTURE WORK

In this paper, a new approach for handwritten text line segmentation has been presented and applied to the IAM database and documents collected from different writers. The input image was preprocessed and the images were segmented into lines. Horizontal projection profile was constructed to detect core region. Following this, separator line was drawn and lines were segmented with noise, which was corrected by run-length for overlapping components. For short lines and skewed or slant lines, distance measures were taken. Threshold value for the distance measure was set to segment lines using connected component recursively. From the experiments, 91.92% accuracy and imbibed reliability were observed in the system. The results obtained by this segmentation, thus show that the proposed system is capable of locating accurately the text lines in images and documents. Future work mainly concerns the improvement of the proposed line segmentation method, by using feedbacks from word segmentation and recognition modules.

5. REFERENCES

[1] Abo Samra. K, *et al*., (2011), A Comprehensive Algorithm for Segmenting Handwritten Arabic Scripts in Off-Line Systems, *Document Recognition and Retrieval XVIII,* Electronic Imaging, United States.

[2] Brodic. D, (2011), Methodology for the Evaluation of the Algorithms for Text Line Segmentation Based on Extended Binary Classification, *Measurement Science Review,* Volume 11, No. 3.

[3] Jayant Kumar, *et al*., Segmentation of Handwritten Text lines in Presence of Touching Components.

[4] Laurence Likforman-Sulem, *et al*., Text Line Segmentation of Historical Documents: A Survey.

[5] Likforman Sulem L. and Faure C., (1994), Extracting Lines on Handwritten Documents by Perceptual Grouping, *Advances in Handwriting and Drawing: A Multidisciplinary Approach*, pp. 21-38, Europia, Paris.

[6] Louloudisa, *et. al*., (2009), Text line and Word Segmentation of Handwritten Documents, *Science Direct Pattern Recognition Journal*, Vol. 42, pp 3169-3183

[7] Manmatha R. and Srimal N, (1999), Scale Space Technique for Word Segmentation in Handwritten Manuscripts, *Proceedings of 2nd International Conference on Scale Space Theories in Computer Vision*, pp. 22-33.

[8] Marti U. and Bunke H., (2001), The Influence of Vocabulary Size and Language Models in Unconstrained Handwritten Text Recognition, *Proceedings of ICDAR'01*, Seattle, pp. 260-265

[9] Nicolaou. A and Gatos. B, (2009), Handwritten Text Line Segmentation by Shredding Text into its Lines, *10th International Conference on Document Analysis and Recognition.*

[10] Oztop E., *et al*., (1999), Repulsive Attractive Network for Baseline Extraction on Document Images, *Signal Processing*, Vol. 75, pp.1-10.

[11] Rodolfo P., *et al*., Text Line Segmentation Based on Morphology and Histogram Projection.

[12] Shi Z. and Govindaraju V., (2004), Line Separation for Complex Document Images Using Fuzzy Run length, *Proceedings of the International Workshop on Document Image Analysis for Libraries*, Palo, Alto, CA.

[13] Syed Saqib Bukhari, *et al*., Script-Independent Handwritten Text line Segmentation Using Active Contours.

[14] Tseng Y.H. and Lee H.J., (1999), Recognition-based Handwritten Chinese Character Segmentation Using a Probabilistic Viterbi Algorithm, *Pattern Recognition Letters*, Vol. 20, No. 8, pp.791-806.

[15] Vassilis Papavassilioua, *et al*., (2010), Handwritten Document Image Segmentation into Text lines and Words, *Science Direct Pattern Recognition Journal*, Vol. 43, pp 369-377.

[16] Wong K., R. Casey and F. Wahl, (1982), Document Analysis Systems, *IBM Journal of research and development*, Vol. 26, No. 6.

[17] Yangdong Gao, Xiaoqing Ding and Changsong Liu, (2011), A Multi-scale Text Line Segmentation Method in Freestyle Handwritten Documents, *International Conference on Document Analysis and Recognition.*