Volume 4, Issue 4, April 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Problems and Review of Line Segmentation of Handwritten Text Document

Rahul Garg, Naresh Kumar Garg
GZS-PTU Campus, Bathinda
India

Abstract*— Optical character recognition (OCR) is a very popular research area since 1950's. Many people has done a lot of work on various scripts. Line segmentation is a very important step in OCR as the accuracy of the recognition algorithm highly depends on the correct line segmentation. Incorrect line segmentation not only decreases the accuracy but also may lead to some other errors. The objective of this paper is to provide some information of problems that usually comes during line segmentation and also to provide a survey of existing methods of line segmentation on handwritten Indian scripts.

Keywords*— Segmentation, handwritten text, problems in line segmentation, header line, base line, a survey.

I. INTRODUCTION

Optical character Recognition is one of the most challenging areas in the field of character recognition and image processing. It is process of electronic conversion of scanned document image of typed, printed or handwritten text to machine encoded text. A lot of work has been done by many people in past few years and many techniques have been developed for segmentation. Segmentation is process of partitioning the text document into lines further to words and further to characters. Line segmentation is a very important part of character recognition as the accuracy of algorithm highly depends on correct line segmentation. Line segmentation is quite easy in printed text as there are equal line spacing in printed text and almost similar writing style. but in hand written text, line segmentation is not so easy due to many problems. In handwritten text, line spacing is not equal also there is a varying font size and style which makes the line segmentation a very tough job as shown in Fig.1. In handwritten text, problem of skewness, overlapping and touching of text also makes line segmentation difficult. So for the proper and good line segmentation all these problems must be corrected before starting the line segmentation.

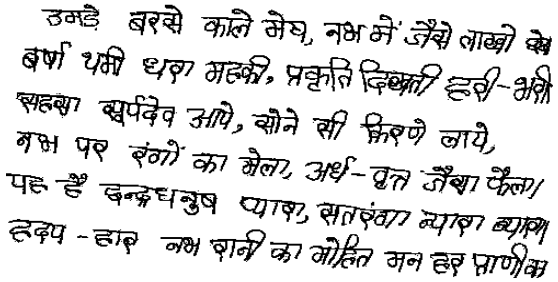

Fig.1 Handwritten Text Document

II. SOME CHARACTERSTICS AND REPRESENATION OF TEXT LINE

Here are some definitions related to line and some factors that makes line segmentation difficult in hand written text.

A. Definitions

1) *Baseline*: In a text line, a line that joins the lower part of character bodies is known as baseline as shown in Fig.2.
2) *Median line*: In a text line, a line that joins the upper parts of character bodies is known as median line as shown in Fig.2.
3) *Upper line*: Line that joins top of ascenders as shown in Fig.2.
4) *Lower line*: Line that joins bottom of descenders as shown in Fig.2.

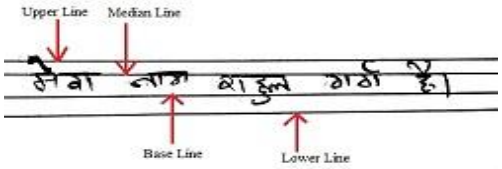

Fig.2 Various line regions.

5) *Touching Component*: These are connected ascenders and descenders belonging to consecutive lines as shown in Fig.3.
6) *Overlapping Component*: Ascenders and descenders belonging to the region of an adjacent line as shown in Fig.3.

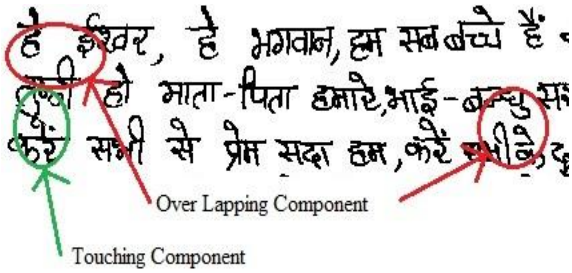

Fig.3 Reference lines and interfering lines with overlapping and touching components.

B. Factors That Makes Line Segmentation Difficult

There are many factors that makes line segmentation difficult in hand written text. Some of them are listed below:

1) *Overlapping and Touching Components*: These components makes the line segmentation difficult as due to presence of these factors it is becomes very difficult to identify the boundary of a text line and hence segmentation becomes difficult.
2) *Influence of author's writing Style:*It is always seen that handwritten document does not contains even style of writing. Line spacing, line orientation are not always unique while writing. While writing, author may provide more line spacing at some place in text while lesser at some place. Also there can be different line orientations especially where there are annotations or corrections.
3) *Influence of poor quality of document image:* Low Quality of document image may produce errors in binarization stage. Words, characters may be split into many connected components and hence it produce many connected components and thus they are very difficult to find. Also the poor scanning of document, poor quality of paper, variable intensity of background and seeping ink from other side of document lowers the quality of image and hence produce errors at binarization stage.

III. TEXT LINE SEGMENTATION

Text line segmentation is the process in which the image of the text is divided into units of patterns that seems to form a character. The accuracy of all the recognition algorithm highly depends on the accuracy of the segmentation algorithm to break the image of text into individual characters. Here is survey of some methods used for line segmentation.

A. Projection-Based Methods

Projection profile method is commonly used for printed text segmentation but it can also be used for handwritten text with little overlap. Here in this method, vertical projection profile is obtained by summing the pixel values along the horizontal axis for each value of y as shown in Fig.4. This projection profile is then used to observe the gap between the text lines.

Fig.4 Horizontal Projections.

In Itay Bar-Yosef, et al. [3], firstly using the piecewise projection profile, the document is divided into vertical strips and then in each strip the local minima of the projection profile is searched. Then these minima are grouped to form a complete text line. This method has advantage that it can be directly applied to a gray scale image. They have also extended their algorithm to admit any skew angle. They have adapted the skew of each line as it progress in the document.

In Yangdong Gao, et al. [4], a technique has been proposed that takes the advantage of both algorithms in large scale and small scale. Between each pair of neighbouring text lines a path is detected dynamically to separate them. During the process, a three- stage multi-scale algorithm is used to find out the line separating path's coordinates. This algorithm includes (i) a simple algorithm for local minima search, (ii) the technique based on following the contour of the foreground component, (iii) a simple piecewise projection profile.

In Vassilis Papavassiliou, et al. [5], Global projection based approach is very good for machine printed text but it fails when the text have different skew angles. So piecewise projection technique is used which can handle text with skewness but it also sensitive to gaps within words and size of character within a text line. So to deal with these problems they have introduced a smooth version of projection profile to over segment each zone into candidate text and gap region. Then they reclassify regions by applying an HMM formulation that enhances statics from whole document page. Then starting from left and moving towards right, they combines the separators of consecutive zones considering their proximity and foreground density.

In Naresh Kumar Garg, et al. [6], authors has proposed a new method for line, word and character segmentation which is based on base line and header line detection. Two-stripe projection have been used by authors for etection of header line and base line. Header line is the most visible part of the text. If there is a skew in the text line, detection of header line becomes a very challenging task. Many authors are detecting header lines by finding the row with maximum pixel density but this fails when there is a skew in the text line.

In N. Tripathy, et al. [7], a water reservoir-concept based scheme is proposed for the segmentation of unconstrained oriya handwritten text into individual characters. For the purpose of line segmentation, the document image is divided into vertical stripes. The heights of the water reservoir are then calculated from the different components of the document. By analysing that heights, the width of the stripe is calculated. Then stripe-wise horizontal histograms are computed and the relationship between the peak-valley points of that histograms is used for line segmentation.

B. Run Length Smearing Method

Run-Length smoothing algorithm can be applied on printed and binarized. Along the horizontal direction, consecutive black pixels are smeared: i.e. white spaces between them are filled with black pixels if their distance is within a predefined threshold. Smearing methods use fuzzy RLSA and adaptive RLSA.

In Zhixin Shi, et al. [10], the proposed algorithm uses the application of fuzzy directional runlength. This proposed technique was tested on many complex handwritten document images which includes postal parcal images and some historical handwritten documents such as Newton's manuscripts. Here on initial image, fuzzy RLSA measure is calculated for every pixel which describes that along a horizontal path, how far one can see while standing at a pixel.

C. Hough Transform

Hough Transform is a feature extraction technique used in Digital Image Processing and Optical Character Recognition. The classical Hough Transform was used to identify lines in the text document image but later on it has been extended to identify positions of arbitrary shapes such as circles or ellipses. The technique works on a voting procedure.

In G. Louloudis, et al. [11], the proposed technique consists of three distinct steps. In first step, image pre-processing, connected component extraction, also partitioning of connected components into three different sub-domains is done. Each of this sub-domain is treated in different manner. Average character height estimation is also done in first step. In second step, for detection of potential text lines, author used a block-based Hough Transform. In Third step, correctness of the splitting is checked means to check for the lines that are not expose in previous step and then finally the vertically connected characters are disconnected and are assigned to text lines.

IV. **CONCLUSIONS**

Projection profile method is best suited for well printed text. It can show 100% accuracy in printed English text but it shows less accuracy in handwritten text due to problems of skewness, overlapping and touching of text. However Run Length Smearing Method can handle the problem of Skewness but it fails when two lines touch each other. Hough- based algorithms fails when there are different skew angles in the same text line. So this paper will provide a basic idea of methods of line segmentation to the person who wants to start their research in this field.

REFERENCES

[1] N. K. Garg, L. Kaur and M. K. Jindal, "A New Method for Line Segmentation of Handwritten Hindi Text", In: Proceedings of the 7th Internation IEEE Conference on Human Technology: New Generations (ITNG), pp. 392-397, 2010.

[2] L. Likforman-Sulem, A. Zahour, B. Taconet, "Text line segmentation of historical documents: a survey", International Journal of Document Analysis and Recognition. Vol. 9, pp. 123-138, 2007.

[3] Bar-Yosef, N. Hagbi, K. Kedem, I. Dinstein, "Line segmentation for degraded handwritten historical documents", International Journal of Document Analysis and Recognition. Vol. 10, pp. 1161-1165, 2009.

[4] Y. Gao, X. Ding, C. Liu, " A Multi-scale Text Line Segmentation Method in Freestyle Handwritten Documents", International Conference on Document Analysis and Recognition. pp. 643-647, 2011.

[5] V. Papavassiliou, T. Stafylakis, V. Katsouros, and G. Carayannis, "Handwritten document image segmentation into text lines and words," Pattern Recognition. Vol. 43, pp. 369–377, 2010.

[6] N. K. Garg, L. Kaur and M. K. Jindal, "Segmentation of Handwritten Hindi Text", International Journal of Computer Applications (IJCA), Vol. 1, pp. 19-23, 2010.

[7] N. Tripathy, "Handwriting Segmentation of Unconstrained Oriya Text", International Workshop on Frontiers in Handwriting Recognition. pp. 306-311, 2004.

[8] N. Modi and K. Jindal, "Text line Detection and Segmentation in Handwritten Gurumukhi Scripts," International Journal of Advance Research in Computer Science and Software Engineering, Vol. 3, pp. 1075-1080, 2013.

[9] M. K. Jindal, R. K. Sharma, and G. S. Lehal, "Segmentation of Horizontally Overlapping Lines in Printed Indian Scripts," International Journal of Computational Intelligence Research, Vol. 3, no.4, pp. 277–286, 2007.

[10] Z. Shi, and V. Govindaraju, "Line separation for complex document images using fuzzy runlength," First International Workshop on Document Image Analysis for Libraries, pp. 306-312, 2004.

[11] G. louloudis, B. Gatos, I. Pratikakis, C.Halatsis, "Text Line Detection in handwritten documents," Pattern Recognition Vol. 41, pp. 3758 – 3772, 2008.

