Review: A Literature Survey on Text Segmentation in Handwritten Punjabi Documents

Namisha Modi*
GZS PTU Campus
Bathinda, Punjab, India
namisha.2207@gmail.com

Ritu Dewan
NIET, Greater Noida
UP, India
ritu1706134@gmail.com

Vaneet Mohan
GZS PTU Campus
Bathinda, Punjab, India
er.vaneetkaur@gmail.com

Jashanpreet Kaur
Punjabi University
Patiala, Punjab,India
jashn00042@gmail.com

Abstract— *Gurumukhi script is used for Punjabi language, which is a two dimensional composition of symbols with connected and disconnected diacritics. Handwritten Gurumukhi script has some complexities like connected, overlapped text lines, words and characters. It is one of the foremost issues for errors during the recognition process. Text segmentation is a challenging job in unconstrained writer independent handwritten document image processing. There is a huge need for research in the domain of Punjabi handwriting recognition to resolve these challenging problems. In this paper we have done a review on various methods for line, word and character segmentation in handwritten Punjabi document.*

Keywords—*Text Line Segmentation, Word segmentation, character segmentation, Gurumukhi script*

I. INTRODUCTION

Gurumukhi script is used primarily for the Punjabi language. The Gurumukhi word exactly means "from the mouth of the guru". Punjabi speakers are not only confined to north Indian states such as Punjab, Haryana and Delhi but are spread over all parts of the world. There is rich literature in this language in the form of scripture, poetry, and books. It is, therefore, important to develop offline handwriting recognition for such a rich and widely used language which may find many practical uses in various areas. The writing style of Gurumukhi script symbol is from left to right side of the paper. In Gurumukhi script, there is no concept of case sensitivity as no upper case or lower case perception exists. There are 41 consonants and 12 vowels in the Gurumukhi script.

In optical character recognition, segmentation is a significant phase. Accuracy of character recognition highly depends on accuracy of text segmentation. Incorrect text segmentation leads to incorrect character recognition. Segmentation stage includes text line, word, and character segmentation. Text detection and separation in digital image documents is a challenging job for handwritten document analysis and character recognition. The problem becomes compounded if the text lines, words and character in the document image are connected or overlapped. Emergence of these problems is familiar in handwritten documents in contrast of printed documents because of writer's varying handwriting styles.

Although some researchers practiced to solve the handwritten text detection problem for Gurumukhi script but the results are not encouraging. Some methods are still not up to the mark in detecting connected component and to separate those connected component from right place, placing disconnected components with their own text lines, words and characters. In most cases, diacritics are incorrectly separated due to overlapping of two adjacent text lines or less vertical gap between the two lines.

II. RELATED WORK

This section describes the work done carried out by the various researchers so far in the field of handwritten text detection in OCR. The observations from the research so far have also been illustrated. The various issues related with text line segmentation in OCR are critically analysed in the literature survey and these help the researchers to understand and carry out the work further in this domain. A variety of text line segmentation methods for handwritten documents has been reported based on projection profiles, fuzzy run length, Hough transform, smearing method, and many others. A. Nicolaou et al. (2009) proposed [5] technique to segment handwritten document images into text lines by shredding their surface with local minima tracer. It is assumed that a path exists from one side of the image to other that traverses only one text line. Tracers are used to follow the white-most and black-most paths from both left to right and right to left direction in order to shred the image into text line areas after blurring the image. Xiaojun Du et al. (2009) presented [4] a new text line segmentation approach based on the Mumford–Shah model. The algorithm is language independent, use piecewise constant approximation of the MS model to segment handwritten text images. In addition, morphing is used by the author to remove overlaps between neighbouring text lines and connect broken text lines. Yi-Feng Pan et al. (2011) presented [17] a hybrid approach to robustly detect and localize texts in natural scene images. Text detection and extraction in natural scene images is significant for content-based image study. This problem is difficult due to the complex background, the non-uniform illumination, and the variations of text font, size and line orientation. A text region detector is designed to estimate the text existing and scale information in image pyramid, which assist segment candidate text components by local binarization. To proficiently filter out the non-text components, a conditional random field (CRF) model considering unary component properties and binary contextual component relationships with supervised parameter learning is proposed. Lastly, text components are grouped into text lines/words with a learning-based energy minimization method. G. Louloudis et al. (2008) presented [1] a text line detection method for handwritten documents. The proposed technique is based on three distinct steps. The initial step includes image pre-processing and connected component extraction, division of the connected component domain into three spatial sub-domains and average character height estimation. Secondly, author used a block-based Hough transform for the detection of potential text lines while third step is to correct possible splitting, to spot text lines that the previous step did not expose and, finally, to cut off vertically connected characters and assigns them to text lines. Yi Li et al. (2008)

presented an approach [2] based on density estimation and a state-of-the-art image segmentation method, the level set method. A probability map is predicted from an input document image where each element represents the probability of the underlying pixel belonging to a text line. Then level set method is developed to determine the boundary of neighbouring text lines by evolving an initial estimate. Fei Yin et al. (2009) proposed [3] a text line segmentation algorithm based on minimal spanning tree (MST) clustering with distance metric learning. The connected components (CCs) of document image are grouped into a tree structure, from distance metric text lines are extracted by dynamically cutting the edges using a new hyper volume reduction criterion and a straightness measure. The proposed algorithm handles a variety of documents with multi-skewed and curved text lines. Vassilis Papavassiliou et al. (2010) presented [8] two approaches to extract text lines and words from handwritten document. The line segmentation algorithm is based on locating the optimal succession of text and gap areas within vertical zones using Viterbi algorithm. A text-line partition drawing technique is applied and then finally the connected components are assigned to text lines. M. K. Jindal et al. (2007) proposed [6] a solution for segmenting horizontally overlapping lines and solve the problem of eight most widely used printed Indian scripts. In this whole document is divided into strips and proposed algorithm is applied for segmenting horizontally overlapping lines and associating small strips to their respective lines. Dhaval Salvi et al. (2013) proposed [7] a method that finds the text segmentation with the maximum average likeliness for the obtained characters. A graph model is used so as to describe the feasible locations for segmenting adjacent characters. Later, an average longest path algorithm is applied to identify the globally optimal segmentation. Nikolaos Stamatopoulos et al. (2008) presented [9] a combination method of different segmentation techniques. It is done to make use of the segmentation results of complementary techniques and specific features of the initial image so as to generate improved segmentation results. The combination method is composed of the following five steps: Average feature extraction, Detect acceptably segmented regions, partition sub regions into groups, Create correctly segmented regions from each group, Last process of the new segmentation result. Alireza Alaei el al. (2011) proposed [16] a painting scheme to accomplish the objective of line segmentation of unconstrained handwritten text. The new method has been devised by studying the cursive Persian text scripts widely. The proposed line segmentation algorithm is applicable to handwritten text in any script. The text block is vertically separated into parallel pipe structures called as strip. Every row in each strip is painted with the average gray intensity value of all pixels intensity present in that particular row-strip. Later, the painted pipes are converted into two-tone painting and then smoothed. The white/black places in each pipe of the smoothed image are analyzed to get a short line of separation called as Piece-wise Potential Separating Line (PPSL), between two consecutive black spaces. The PPSLs are joined to produce the segmentation of text lines.

III. CONCLUSIONS

In this paper we present a review for text line, word and character segmentation from text of complex handwritten Gurumukhi document images. Handwritten Gurumukhi script has some complexities like connected, overlapped text lines, words and characters. It is one of the major reasons for errors during recognition process. From the above review of text line, word and character segmentation it is clear that it is very useful for segmenting lines words and characters but more study is required to work on overlapped and broken word and characters. Thought we have not achieved required level of accuracy but the results obtained are encouraging and satisfactory. Incorrect word segmentation will further leads to incorrect character segmentation. Improvement in character segmentation method will lead to more accurate classification and recognition of Punjabi Text.

REFERENCES

[1] G. louloudis, B. Gatos, I. Pratikakis, and C.Halatsis, ”Text Line Detection in handwritten documents,” Pattern Recognition vol.41, pp. 3758 – 3772, 2008.

[2] Yi Li, Yefeng Zheng, David Doermann, and Stefan Jaeger,” Script-Independent Text Line Segmentation in Freestyle Handwritten Documents.” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 8, pp. 1313 – 1329, Aug. 2008.

[3] Fei Yin, and Cheng-LinLiu,”Handwritten Chinese text line segmentation by clustering with distance metric learning,” Pattern Recognition 42, pp. 3146 – 3157, 2009.

[4] Xiaojun Du, Wumo Pan, and Tien D. Bui,” Text line segmentation in handwritten documents using Mumford–Shah model,” Pattern Recognition vol. 42, pp. 3136 – 3145, 2009.

[5] A. Nicolaou, and B. Gatos, “Handwritten Text Line Segmentation by Shredding Text into its Lines,” 10th International Conference on Document Analysis and Recognition, IEEE Computer society, 2009, pp. 626-630.

[6] 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.

[7] Dhaval Salvi, Jun Zhou, Jarrell Waggoner, and Song Wang, ”Handwritten Text Segmentation using Average Longest Path Algorithm,” Applications of Computer Vision(WACV), IEEE Workshop, pp. 505 - 512, 2013.

[8] Vassilis Papavassiliou, Themos Stafylakis, Vassilis Katsouros, and George Carayannis,” Handwritten document image segmentation into text lines and words,” Pattern Recognition, vol. 43, pp. 369 – 377, 2010.

[9] Nikolaos Stamatopoulos, Basilis Gatos, and Stavros J. Perantonis, ”A method for combining complementary techniques for document image segmentation,” Pattern Recognition vol. 42, pp. 3158 – 3168, 2009.

[10] Zhixin Shi, and Venu Govindaraju, “Line separation for complex document images using fuzzy runlength,” First International Workshop on Document Image Analysis for Libraries, pp. 306, 2004.

[11] B. Gatos, A. Antonacopoulos, and N. Stamatopoulos, ICDAR2007 handwriting segmentation contest, in: 9th International Conference on Document Analysis and Recognition (ICDAR'07), Curitiba, Brazil, pp. 1284 – 1288, Sept. 2007.

[12] Rajiv Kumar, and Amardeep Singh,” Algorithm to Detect and Segment Gurmukhi Handwritten Text into Lines, Words and Characters” IACSIT International Journal of Engineering and Technology, vol.3, no.4, pp. 392 - 395, Aug. 2011.

[13] Rajiv Kumar, and Amardeep Singh, “Detection and Segmentation of Lines and Words in Gurmukhi Handwritten Text” 2nd International Advance Computing Conference, 2010, pp. 353 - 356.

[14] Naresh Kumar Garg, Lakhwinder Kaur, and M.K.Jindal,” A New Method for Line Segmentation of Handwritten Hindi Text”

Seventh International Conference on Information Technology, 2010, pp. 392 - 397.

[15] A. Zahour, B. Taconet, L. Likforman-Sulem, and Wafa Boussellaa, “Overlapping and multi-touching text line segmentation by Block Covering analysis,” Pattern Analysis and Applications, vol. 12, pp. 335-351, 2008.

[16] Alireza Alaei, P. Nagabhushan, and Umapada Pal, “Piece-wise painting technique for line segmentation of unconstrained handwritten text: a specific study with Persian text documents,” Pattern Analysis and Application, vol. 14, pp. 381–394, 2011.

[17] Yi-Feng Pan, Xinwen Hou, and Cheng-Lin Liu, “A Hybrid Approach to Detect and Localize Texts in Natural Scene Images,” IEEE Transactions on Image Processing, vol. 20, no. 3, pp. 800 – 813, March 2011.

[18] Namisha Modi, and Khushneet Jindal, “Text Line detection and Segmentation in Handwritten Gurumukhi Scripts,” International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, no. 5, pp. 1075 – 1080, May 2013.

[19] Namisha Modi, and Khushneet Jindal, “Detection and Segmentation of text in Handwritten Punjabi Scripts,” ICECIT-2013, vol. 3, no.3, Oct 2013.