Volume 4, Issue 7, July 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Review on Text Line Segmentation Problems and Techniques of Gurumukhi Handwritten Scripts

Er. Snehdeep
Student, M.Tech* (CE)
YCOE, Talwandi Sabo, Punjab, India

Er. Manoj Chaudhary
AP, Dept. of Computer Engineering
YCOE, Talwandi Sabo, Punjab, India

***Abstract*—OCR is the process of recognition characters from scanned documents. Line segmentation is a very important step in OCR. Accuracy of OCR depends upon correct line segmentation. Segmentation of text document image is a big challenge in OCR Systems. The problem becomes more complex in handwritten documents due to Skewed, overlapped lines and touching lines. The objective of this paper to provide a review of most complicated problems present in segmentation and also provide a review of methods of handwriting or printed text line segmentation proposed by various authors.**

***Keywords*— OCR; Segmentation; Line Segmentation; Overlapping Line segmentation; connected components line segmentation; Gurumukhi script**

I. INTRODUCTION

In character recognition field, OCR is a crucial area. It is a technique to process scanned document and make the document editable. OCR field take a major turn in 1950 with the development of technology. OCR field is very popular in business world and banking industries. OCR system convert the text image in machine encoded form which reduces the space required for storage. OCR is the one of the methods used to digitalize the handwritten documents that makes transmission of documents easy. OCR process consists of three major sub-phases like Pre-processing, Segmentation and Recognition [7].

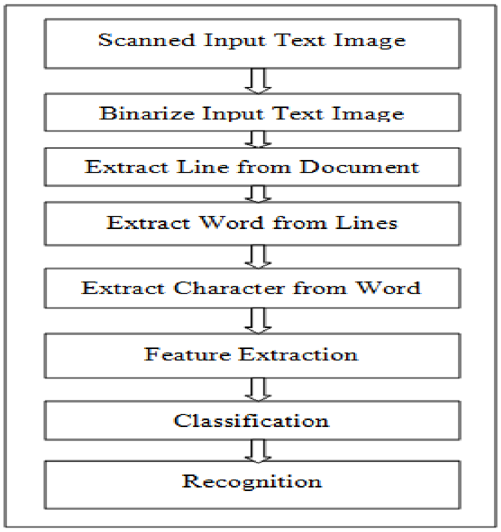

Fig.1: Steps of OCR

SEGMENTATION

Segmentation process is the backbone of the overall OCR process [7]. Segmentation is a technique to divide a text image document into line, words and character and recognize character and its features for further processing. OCR system accuracy depends upon segmentation. For correct recognition of characters there is a need to perform segmentation correctly.

LINE SEGMENTATION

Line segmentation is a technique to divide a paragraph into number of lines. Incorrect line segmentation leads towards incorrect character recognition. Line segmentation is a technique to extract number of lines and boundaries of each line in any input image document before word and character segmentation. Line segmentation is an important step in OCR. Line segmentation is difficult in handwritten text documents as there is unequal spacing between lines and different writing styles.

These problems make line segmentation of handwritten documents more complex. A lot of research work has been done by various researchers and various techniques developed for line segmentation. But there is still need improvement in existing techniques to increase efficiency of OCR systems. In handwritten documents, Problem of skewed text lines, overlapping lines and connect components makes line segmentation more complicated.

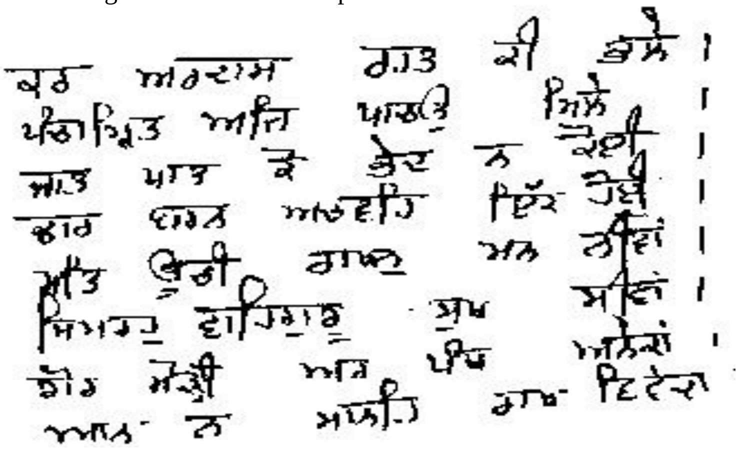

Fig. 2: Handwritten Text Document

II. CHARACTERISTICS OF GURUMUKHI SCRIPTS AND TEXT LINE

Gurumukhi Scripts is commonly used to write Punjabi language which consists of 41 consonants and 12 vowels. There is no concept of upper and lower case letter in Gurumukhi scripts. Some characters in the form of half characters are also present in lower zone of characters called half characters [3]. Writing style is from left to right. A line of Gurumukhi script can be partitioned into three horizontal zones that is upper zone, middle zone and lower zone [3].

In Text line representation, text line is categorized in different sections. These sections are:

1) Baseline: A line that connects the lower part of character bodies is known as baseline as shown in Fig.3.
2) Middle line: A line that connects the upper parts of character bodies is known as median or middle line as shown in Fig.3.
3) Upper line: A line that connects top of ascenders or upper modifiers as shown in Fig.3.
4) Lower line: A line that connects bottom of descenders or lower modifiers as shown in Fig.3.

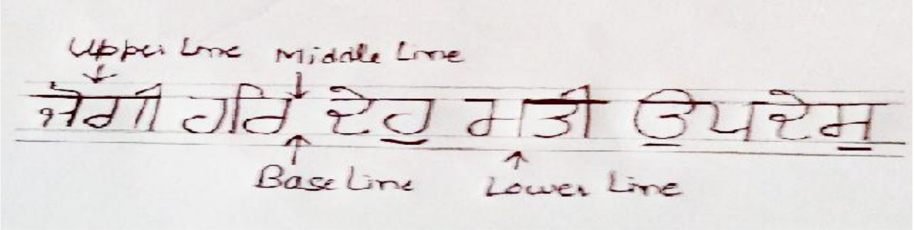

Fig. 3: Various Text Line Regions

III. LINE SEGMENTATION PROBLEMS

Hand-written document leads to many problems in segmentation due to overlapped lines, touching lines, connected components and skew text lines. Freestyle and unconstrained handwriting text line segmentation is considered a complex and challenging task due to the following characteristics [13]:

i) Skewed Lines Skewed lines make document analysis, document understanding and visualization tasks more difficult [4]. Skewness is present in the document vary with varying hand-writing style of the writer. Skewed text lines categorized into three types based upon skew text.

a. Single Skew
b. Multi skew
c. Non-uniform Skew

a. Single skew: Single skew is again categorized into two parts: Global Skew and Local Skew.
Global skew mostly comes into existence in printed text. But Local (varying) Skew, Multi-skew and Non uniform skews are found in handwritten text. Global Skew detection of printed text is easier than Varying skew and Multi-skew. Because all line are parallel to each other and have same value of skew angle [4].

Global Skew: The entire page block has same orientation as shown in figure [9].

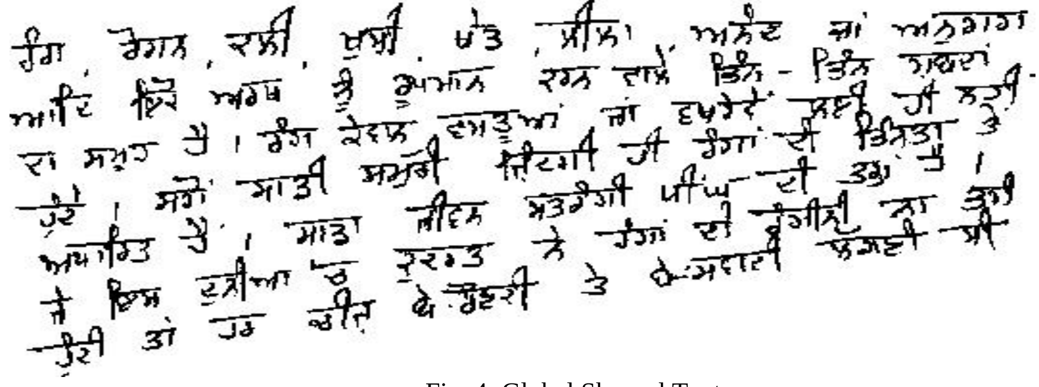

Fig. 4: Global Skewed Text

b. Multiple Skew: Document containing multiple skew that is unaligned paragraphs are different in different blocks of page [9].

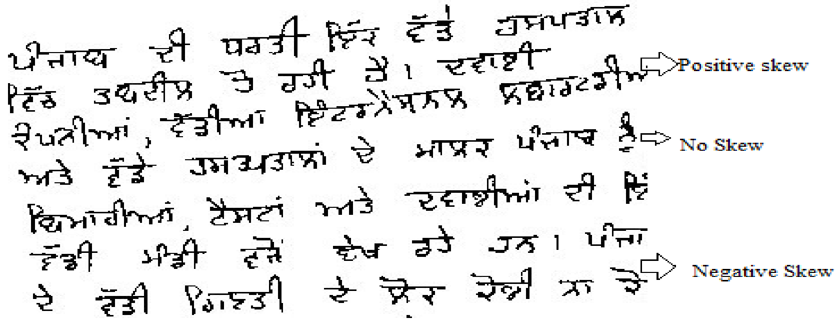

Fig. 5: Multiple Skewed Texts

c. Non-Uniform Skew: Document contains varying text line slope in which slant is distinct along the same line of text as shown in figure [9].

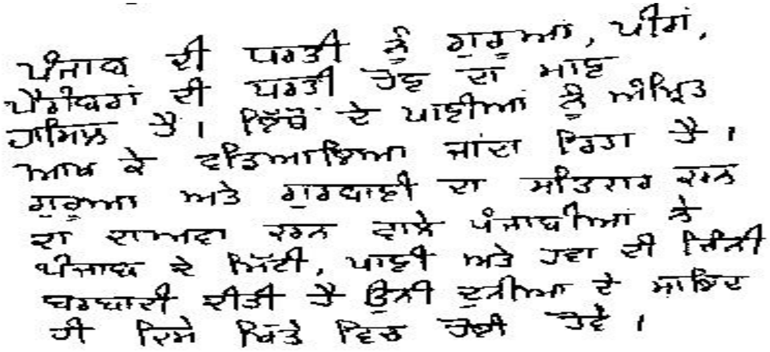

Fig. 6: Non-uniform Skew Text

ii) Lines Adjacency Due to inter-line distance between neighbouring lines will cause the problems like touching , overlapping and connected components between lines.

a. Overlapped Lines

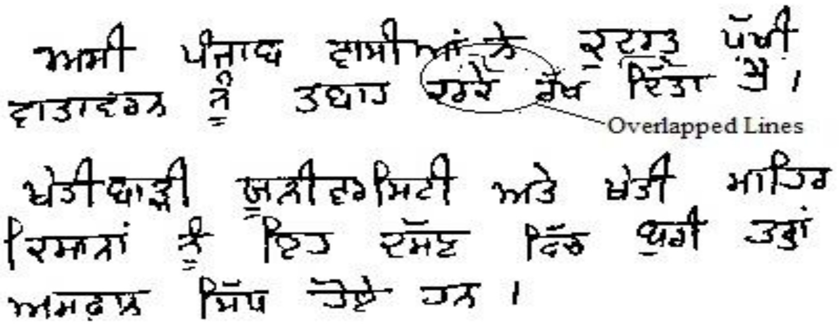

Fig. 7: Overlapped lines

b. Connected Component Problem: Connected component problem arises because of upper and lower modifiers of neighboring lines as shown in figure.

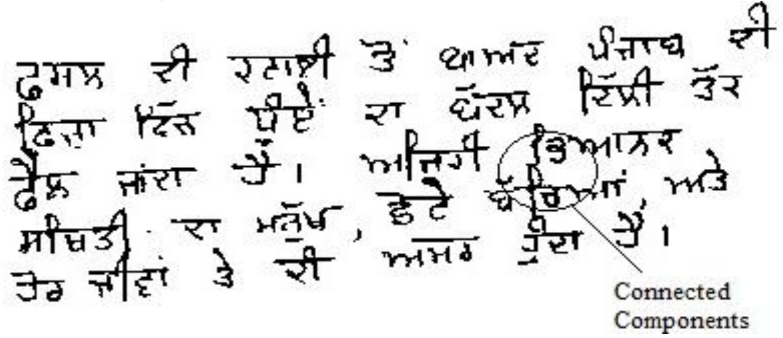

Fig. 8: Text containing connected components

c. Touching Lines: These are overlapped lines in which lower modifiers of a line touching at various points to neighboring lines with upper and middle modifiers.

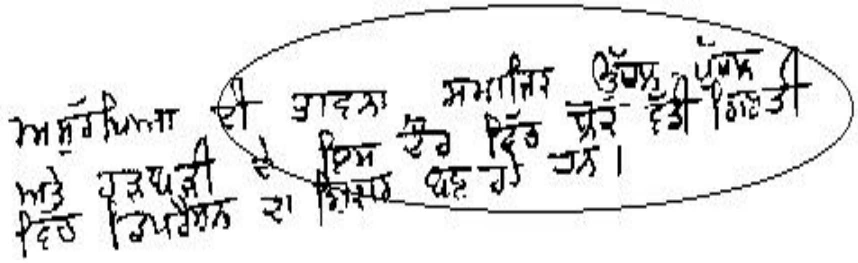

Fig. 9: Text containing touching lines

IV. EXISTING TECHNIQUES OF LINE SEGMENTATION

Several approaches are available for line segmentation and skew detection and correction. Handwriting text line segmentation approaches can be categorized according to the different strategies used [13].These can be classified as-

4.1 Projection Based Approach

Projection Profile methods are commonly used for printed as well as handwritten text segmentation with little overlap. These approaches used the structural characteristics of the document [2]. These approaches distinguish into two types- Vertical Profile Projection and Horizontal Profile Projection. These projections are used to insert gap between lines. Vertical Projection profile is obtained by summing pixel values along the horizontal axis [13].

In M.K.Jindal, et al.[3], describes the line segmentation technique based upon horizontal profile projection. In this approach, document is divided into different stripes. Each strip uses the concept of headline and average height for segmentation. This proposed algorithm solves the problem of horizontally overlapping lines. This method is implemented in those scripts which have the concept of headline.

In Vikas J Dongre, et.al [11], a technique has been proposed based on global horizontal projection method and constructs histogram. This technique provides efficient results for line and word segmentation.

In Naresh Kumar Grag, et.al [6], authors presents a new method for line and word segmentation that is based on base line and header line detection. Two-stripe projections have been used by authors for detection of header line and base line. Header line identified with the help of maximum number of pixels. Detection of header line becomes complicated in skew text line because slope of header line vary.

4.2 Grouping Approach

Grouping approach is the process of grouping the pixels according to specific constrains designed to result to a layer of text lines [2].

Feldbach et.al [14], proposed a method for line detection and segmentation in historical church registers. This method is based on local minima detection of connected components and is applied on a chain code representation of connected components. Line segments are constructed until a unique line is formed. This method is able to segment text lines closed to each other, touching and fluctuating lines.

4.3 Hough-based Approach

Hough Transform is used for locating straight lines in scanned text image. This technique based on the geometric parametric geometric shapes and identifies geometric locations. These geometric parametric suggest the existence of the sought shape.

This technique is proposed to detect fuzzy snapshots of objects in a certain category of shapes and under a voting procedure. This method is not very popular because of its computational complexity [2].These approaches applied to skew detection and correction and handwritten as well as printed text document segmentation.

Louloudis et.al [15], proposed a text line detection method for unconstrained handwritten documents. It uses the concept of connected component removal and estimating the average height of character. For Connected component removal, authors applied block-based Hough transform for detecting the potential lines. The proposed technique provides satisfactory results for text line segmentation of unconstrained handwritten documents.

4.4 Graph Based Approach

Graph Based Methods use the concept of graphs. In this, the representation of document images by graphs is an important tool of line segmentation procedure. The graph is constructed as vertices of pixel or more complex connected components [2].

Vasant Manohar et.al [16], presented a method for line segmentation. Authors describe the approach in terms of the structure of the graph on the document images connected components. The edge-costs of the graph are normally associated with weighted edges that depict distances between connected components and clustering nodes in the graph to obtain text lines.

4.5 CTM Approach

CTM (Cut Text Minimum) approach finds a path or cut line in between the text lines to be separated which minimizes the text line pixels cut by the segmentation line that is minimize the separation of descenders from the upper line and ascenders from lower line. The projection is used for rough estimate of text line separations [10].

4.6 Smearing Approach

Smearing methodology used for printed and handwritten documents and solves the problems such as overlapped, touched and connected components. These methods generally use many thresholds and heuristic rules. It is the process to modify a set of background pixels located between foreground pixels into foreground pixels. Smearing methods used many local techniques to solve specific problems and overlapping touched connected component. These methods give extraordinary results with documents that contain characters of variable height [2].

V. CONCLUSION & FUTURE WORK

Segmentation of handwritten documents is a difficult task due to upper and lower modifiers problems in Gurumukhi scripts. In this paper, we have discussed the major problems that make line segmentation more complicated. These problems create difficulty in OCR System processing. This paper has also provided a brief description of line segmentation approaches proposed by various researchers. These techniques provide excellent result in case of line segmentation. But still problem of connected components, overlapping text lines require improvement. We will try to improve existing techniques to solve line segmentation problems. So this paper will provide basic idea of problem concerning in line segmentation of handwritten Gurumukhi Scripts to the researchers and future study is carried out to solve such problems.

REFERENCES

[1] Ashu Kumar, Simpel Rani Jindal, “Segmentation of Handwritten Gurumukhi text into Lines”, International Conference on recent advances and Future Trends in Information Technology (iRAFIT2012), pp. 13-17,2012.

[2] Ergina Kavallirratou, Fotis Daskas, “Text Line Detection and Segmentation Uneven Skew Angles and Hill-and-Dale Writing”, Journal of Universal Computer Science, Vol.17, No.1, 2010.

[3] M.K. Jindal, R.K. Sharma, G.S.Lehal, “Segmentation of Horizontally Overlapping lines in Printed Gurmukhi Script”, IEEE, 2006.

[4] Mandeep Kaur, Lovnish Bansal, Jagdeep Singh, “Global and Local Skew Detection of Handwritten Gurumukhi Scriptt”,International journal of Computer Application, Vol. 82-No 17, November 2013.

[5] Namisha Modi, Khushneet Jindal, “Text line detection and segmentation in Handwritten Gurumukhi Scripts”, International Journal of Advanced Research in Computer Science and Software Engineering, vol.3, Issue 5, PP:1075-1080, May, 2013.

[6] Naresh Kumar Garg, Lakhwinder Kaur, M.K. Jindal, “Segmentation of Handwritten Hindi Text”, International Journal of Computer Applications, Vol.1, No.4, 2010.

[7] Rajiv Kumar, and Amardeep Singh, “Detection and Segmentation of Lines and Words in Gurumukhi Handwritten Text”, IEEE, 2010.

[8] Saiprakash Palakollu, Renu Dhir, Rajneesh Rani, “Handwritten Hindi Text Segmentation Techniques for Lines and Characters”, World congress on Engineering and Computer Science, Vol. 1, 2012.

[9] Simpel Jindal, Gurpreet Singh Lehal,”Line Segmentation of Gurumukhi Manuscripts”, December 2012.

[10] Naresh Kumar Garg, Lakhwinder Kaur, M.K. Jindal, “A New Method For Segmentation of Handwritten Hindi Text”, International Conference on Information Technology, IEEE, 2010.

[11] Vikas J Dongre, Vijay H Mankar, “Devnagari Document Segmentation Using Histogram Approach” , International Journal of Computer Science, Engineering and Infotmation Technology(IJCSEIT), Vol.1,No.3, August, 2011.

[12] Loveleen Kaur, Simpel Jindal,”Skew Detection Technique for Various Scripts”, International Journal of Scientific & Engineering Research, Vol. 2, Issue 9, Sept-2011.

[13] Zaidi Razak, et al. , “Off-line Handwriting Text Line Segmentation : A Review”, IJCSNS International journal of Computer Science and Network Security, VOL.8 No.7, July 2008.

[14] Markus Feldbach, Klaus D. Tonnies, “Line Detection and Segmentation in Historical Church Registers”, Preceedings of Sixth International Conference on Document Analysis and Recognition(ICDAR’01),IEEE,2001.

[15] G. Louloudis, B.Gatos, I. Pratikakis,C. Halatsis, “A Block-Based Hough Transform Mapping for Text Line Detection in Handwritten Documents”, Proceedings of the tenth International Workshop on Frontiers in Handwriting Recognition, La Baule,Oct.2006.

[16] Vasant Manohar, Shiv N. Vitaladevuni, Huaigu Cao, Rohit Prasad, Prem Natarajan, “Graph Clustering-based Ensemble Method for Handwritten Text Line Segmentation”,IEEE,2011.