Line Segmentation of Handwritten Gurmukhi Manuscripts

Simpel Jindal
Yadwindra College of Engineering,
Talwandi Sabo, Punjab, India
+91-9855283188
simpel_jindal@rediffmail.com

Gurpreet Singh Lehal
Department of Computer Science
Punjabi University, Patiala
+91-9815473761
gslehal@gmail.com

ABSTRACT
The development of an OCR system for recognition of old Gurmukhi handwritten manuscripts is a complex task involving many difficulties. Historical documents are affected by problems of ageing and repeated use and many other uncontrollable factors. Segmentation is one of the important phase of an OCR, as accuracy of an OCR depends upon the accuracy of segmentation. The writing styles of historical documents make the activity of segmentation extremely difficult. Segmentation includes line, word and character segmentation. In this paper, we have discussed a method for segmenting lines for Gurmukhi handwritten manuscripts.

Keywords
OCR, Line Segmentation, Gurmukhi script, segmentation.

1. INTRODUCTION
Historical documents contain a lot of important information and are available in almost each script all over the world. Since these documents are deteriorating day by day, digital preservations is important for these kinds of documents. Optical Character Recognition (OCR) system is one of the methods to preserve these documents and making them useful for further use. Making an Optical character recognition system for historical manuscripts is a challenging task as historical documents contains a number of problems to be taken care of.

We are working on recognition of ancient handwritten historical documents written in Gurmukhi script, which is one of the important Indian scripts used in north India. There are huge amount of historical documents in handwritten Gurmukhi script containing ancient manuscripts, early printed books and typewritten documents of the twentieth century. But major composition is the handwritten manuscripts mainly written by spiritual persons of fifteenth to eighteenth centuries.

The recognition of historical documents is a challenging problem and application of existing technology for this purpose is not successful. More robust methods are required to develop to cope with this challenging problem. While making an OCR system, line segmentation is an important step, as the accuracy of OCR heavily depends upon the correct segmentation. Incorrect segmentation leads to incorrect feature extraction and incorrect classification.

DAR '12, December 16 2012, Mumbai, IN, India

Ntzios *et. al.* [1] have proposed a segmentation free technique for the detection and recognition of characters and characters ligatures in old Greek handwritten documents. Firstly, they detected the open and closed cavities, and then the classification of a specific character or character ligature is done based on the protrusible segments that appear in the topological description of the characters skeletons. Gatos *et. al.* [2] have discussed the results of handwriting segmentation contest organized during ICDAR 2007. The various methods presented in the contest include BESUS method, DUTH-ARLSA method, ILSP-LWSeg method, PARC method and UoA-HT method for segmenting the handwritten text.

Lu and Shridhar [3] have studied the segmentation of hand-printed words, handwritten numerals and cursive handwritten words. Casey and Lecolinet [4] have divided their survey into four categories: dissection techniques, recognition-based segmentation, mixed strategies (oversegmentation) and holistic strategies. Jindal *et. al.* [5-7] have discussed some of the major problems encountered during recognition of degraded Gurmukhi script characters.

A pertinent literature related to text line segmentation in historical documents has been thoroughly reviewed by Sulem *et. al.* [8]. The review encompasses techniques for segmenting printed or handwritten documents, broken and touching characters and a comparative study of segmentation results are addressed as well. Improvements are still necessary to obtain reasonable segmentation rates.

There are mainly three basic categories that these text line detection methods fall in [9]. Methods lying in the first category make use of the Hough transform, second category make use of projections and the third category deals with methods that use a kind of smearing. In some methods that do not lie in these categories, the text line extraction problem is seen from an Artificial Intelligence perspective. The aim is to cluster the connected components of the document into homogeneous sets that correspond to the text lines of the document.

2. PROBLEMS DURING LINE SEGMENTAION
Traditionally used techniques for handwriting recognition cannot be applied to old handwritten manuscripts of Gurmukhi script. The reason behind this is that the aforementioned manuscripts entail several unique characteristics:

1. Headlines of the words are not straight causing problem in recognizing words or lines and failing traditional algorithm for segmentation.
2. The character set of traditional handwritten Gurmukhi script documents contains symbols that are not used in modern Gurmukhi script.
3. There are lots of touching characters in a single word. Sometimes more than two characters touch each other, making the algorithm process more complicated. Also, we can find many touching words in a line. Further overlapping (touching) lines are frequently found in these kinds of documents.
4. Heavy printed characters are also found in abundance in these documents.
5. There is uneven space between lines, between words and even between characters.
6. There is much shape variation in different occurrences of a single character in terms of height, width and many important features like loops and sidelines etc.

Line segmentation of historical documents is a difficult task as many problems are faced during line segmentation *e.g.*, lines of text in general are not straight, the inter-line distance variability and inconsistent distance between the components may vary due to writer movement. It may be straight, straight by segments, or curved. There are three types of skew which exists in documents making the process of line segmentation more difficult:

1. A document may have global skew, in which all the page blocks have the same orientation as shown in figure 1.

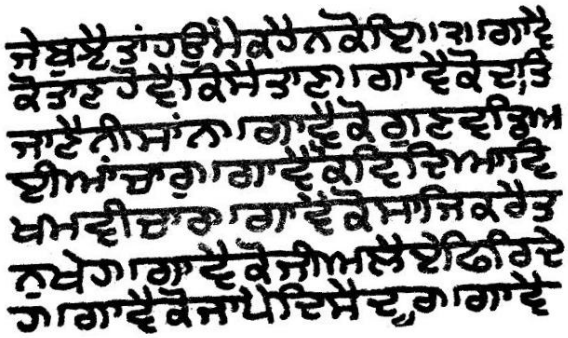

Figure 1. Global skew.

2. Documents containing multiple skew, *i.e.*, unaligned paragraphs or slant are different in different blocks of the page as shown in figure 2, there is anticlockwise rotation on top of the page while there is clockwise rotation at bottom and there is almost no skew in middle of the figure 2.
3. Other than these kind of skewness a document may contain non uniform text line skew or varying text line slope in which slant is different along the same line of text, for example curvilinear text lines as shown in figure 3.

Other then these, touching or overlapping of lower zone and upper zone characters with middle zone characters is another problem.

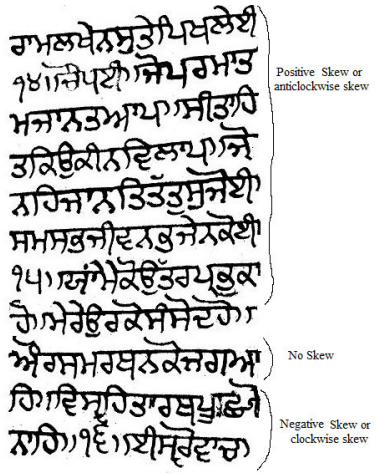

Figure 2. Multiple skew.

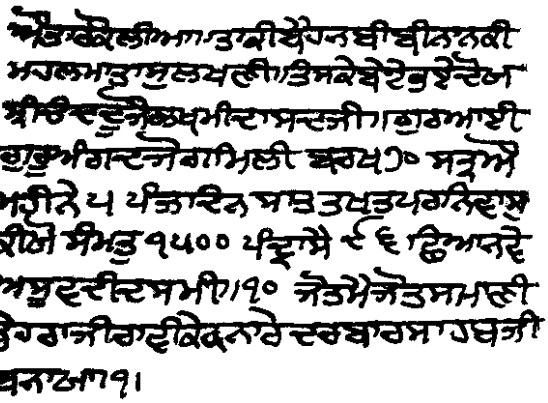

Figure 3. Non uniform text line skew or varying text line slope.

3. PROPOSED TECHNIQUE OF LINE SEGMENTATION

For line segmentation, we have used proposed the following algorithm linesegmanucsript.

Algorithm: Linesegmanuscripts

Step 1: We have divided the document into non overlapping vertical stripes. We have experimented on text documents shown in figures 1-3.

Step 2: For each strip, we have projected all black pixels on the Y-axis and selected positions whose number of accumulative pixels is minimal. The pixels between two minimal positions constitute one text block.

Step 3: Further the text blocks have been divided into three categories, Small Text Blocks (STB) containing upper zone or lower zone characters or some part of middle zone. Average Text Blocks (ATB) containing middle zone along with upper and/or lower zone. Large Text Blocks (LTB) contain overlapping lines.

Step 4: Text line extraction is carried out by segmenting the larger text blocks and assigning each resulting text block to a single line. For segmenting larger text blocks, first the average size of text blocks has been calculated. We have identified the larger text blocks having height greater than a threshold value from average size of text blocks. For segmenting the LTB we have used the connected components. Starting from top of the LTB, find a position where connected component separates. That position is marked as segmentation point. The same process is applied iteratively on same strip until the size of LTB reduces below the threshold value. This step solves the problem of a LTB containing many overlapping text strips. With the breakage of LTB into many text blocks we have increased the total number of text blocks in that strip by the same value.

Step 5: Also the problem of small text blocks has been solved by merging the small text block to nearest average text block. First, the process of identification of STB starts. A text block having height below a threshold value from average text block size is considered to be STB. For each STB the nearest ATB is noted, it can be the previous one of STB or next one in same strip. The nearest STB is merged with the nearest ATB and no of text blocks is reduced accordingly.

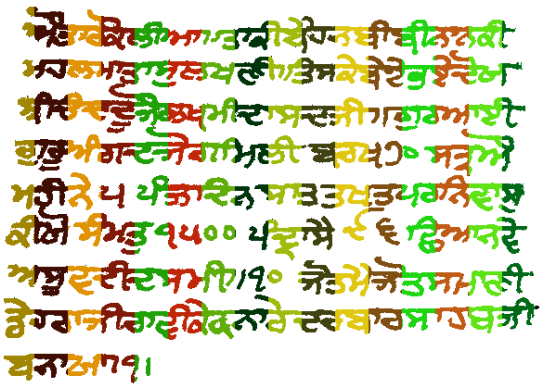

Figure 4. Vertical strips of input document.

Figure 4 and figure 5 shows the results of the algorithm applied on three problem documents. Figure 4 contains vertical strips corresponding to problem document (shown in figure 3) as discussed in step 1 of the proposed algorithm.

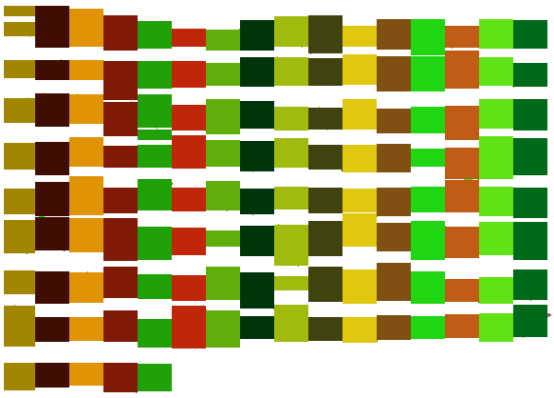

Figure 5. Text blocks identified within strips.

Figure 5 contains the segmented lines using the proposed technique as discussed in step2. Also, figure 6 contains the connected components of the problem document. In figure 7, we have shown large sized text blocks which are spanning between adjacent lines as discussed in step 4 of the algorithm.

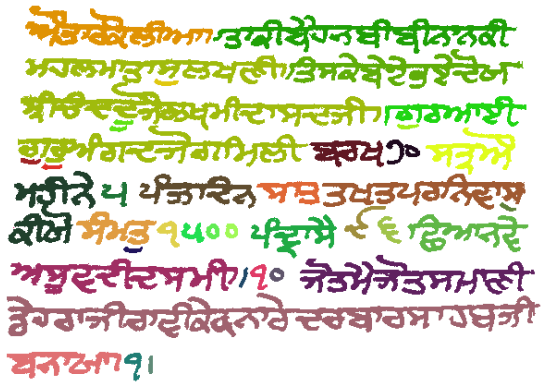

Figure 6. Connected components.

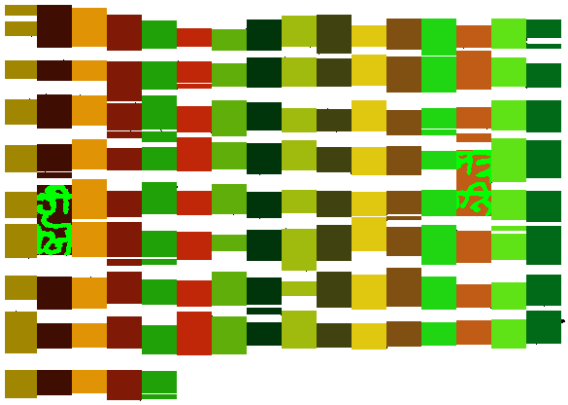

Figure 7. Large sized text blocks.

After implementing the steps 4 and 5 of the algorithm we have segmented the lines of the problem document and the result has been shown in figure 8. One can see that each line has been shown in different color, and all the lines have been correctly segmented.

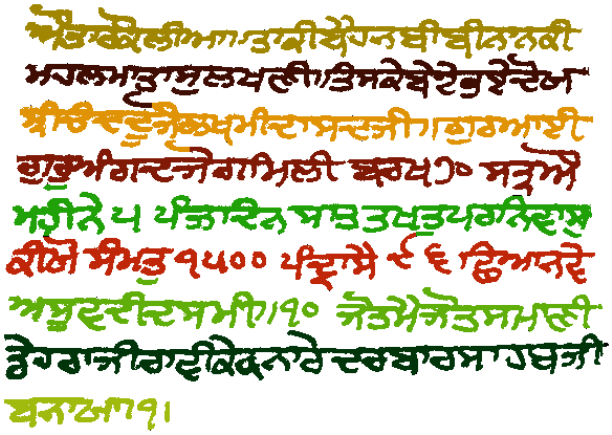

Figure 8. Segmented lines using the proposed algorithm.

Similarly we have applied the algorithm on the problem documents shown in figure 1 and 2 and the results have been shown in figure 9 and figure 10 respectively.

Figure 9. Segmented lines using the proposed algorithm for problem document shown in figure 1.

Figure 10. Segmented lines using the proposed algorithm for problem document shown in figure 2.

4. RESULTS AND DISCUSSIONS

There are few limitations while working on this algorithm:

1. We are still working to find optimal number of strips in which the whole document should be divided.
2. Division of larger text blocks into smaller text blocks is not 100% accurate.
3. Sometimes smaller text lines are not merged into their required text lines.
4. The threshold values for calculating larger text blocks and smaller text blocks are to be optimized.

Due to these reasons, sometimes characters in upper or lower zone of a line have been mis-segmented as shown in figure 11.

This algorithm has produced very good results for line segmentation. We have received 100% accuracy in line segmentation as shown in table 1 using the proposed line segmentation algorithm. Although, the problem of incorrectly segmented characters in lower and upper zone exists to some extent.

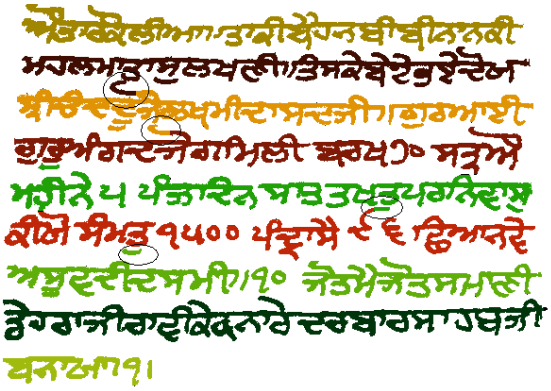

Figure 11. Mis-segmented characters.

Table I. Line segmentation accuracy.

Document	Total lines	Correctly segmented lines	Incorrect segmented upper /lower zone characters
Doc 1	9	9	4
Doc 2	7	7	5
Doc 3	11	11	1
Doc 4	10	10	3
Doc 5	9	9	7
Doc 6	8	8	2
Doc 7	10	10	4
Doc 8	8	8	6

The major advantages of this algorithm are:

- This strategy of segmenting lines is very useful for segmenting text document containing multiple skew.
- Extremely useful for segmenting heavily overlapping text lines.
- Can be implemented for segmenting handwritten text of Gurmukhi script.
- Can be useful for segmenting text lines from historical text documents of another Indian script.

5. CONCLUSIONS

The major kinds of problems encountered during line segmentation of historical documents in Gurmukhi script have been discussed in this paper. We have applied the idea of text blocks for segmenting the lines. We have received very good accuracy for line segmentation using the proposed algorithm, but the problem of incorrectly segmented lower/upper zone characters remains there. We are trying to solve this problem also using the concept of connected components.

6. REFERENCES

[1] K. Ntzios, B. Gatos, I. Pratikakis, T. Konidaris, and S. J. Perantonis. An old greek handwritten OCR system based on an efficient segmentation-free approach. *IJDAR*. 9: 179-192, 2007.

[2] B. Gatos, A. Antonacopoulos, and N. Stamatopoulos.Handwriting Segmentation Contest. In *Proceedings of 9th ICDAR*, pages 1284-1288, 2007.

[3] Y. Lu, and M. Shridhar. Character segmentation in handwritten words - an overview. *Pattern Recognition*. 29(1): 77-96, 1996.

[4] R. G. Casey, and E. Lecolinet. A survey of methods and strategies in character segmentation. *IEEE Transactions on PAMI*. 18(7): 690-706, 1996.

[5] M. K. Jindal, R. K. Sharma, and G. S. Lehal. On Segmentation of touching characters and overlapping lines in degraded printed Gurmukhi script. *International Journal of Image and Graphics (IJIG)*. World Scientific Publishing Company, 9(3): 321-353, 2009.

[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 (IJCIR)*. Research India Publications, 3(4): 277-286, 2007.

[7] M. K. Jindal, R. K. Sharma, and G. S. Lehal. A Study of Different Kinds of Degradation in Printed Gurmukhi Script. In *Proceedings of the IEEE International Conference on Computing: Theory and Applications (ICCTA'07)*, pages 538-544, 2007.

[8] L. L. Sulem, A. Zahour, and B. Taconet. Text line segmentation of historical documents: a survey. *IJDAR*. 9, 123-138, 2007.

[9] G. Louloudis, B. Gatos, I. Pratikakis, and C. Halatsis. Text line and word segmentation of handwritten documents. *Pattern Recognition.* 42(12): 3169-3183, 2009.