© 2016 IJSRSET | Volume 2 | Issue 2 | Print ISSN : 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

Line Segmentation of handwritten text documents written in Gurumukhi Script-A Review

Sheetal Mongla, Rajneesh Narula
A.I.E.T Faridkot, Punjab, India

ABSTRACT

Optical Character Recognition (OCR) is a process to recognize the handwritten or printed scanned text with the help of a computer. Segmentation is an important task of any OCR as accuracy of overall OCR system depends on the segmentation algorithm. Line segmentation of handwritten text documents is a very crucial step in optical character recognition because every user has a different writing pattern and style. Incorrect line segmentation leads to wrong results in word segmentation and finally in recognition. In printed text, line segmentation is quite easy due to consistent pattern and style. While segmenting the handwritten text documents into lines there are various problems that must be satisfied by the proposed algorithm. The objective of this paper gives the review on various problems present in Line segmentation of a handwritten document written in Gurumukhi Script by various authors.

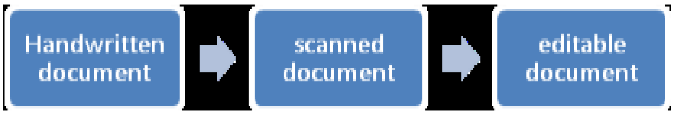
Keywords: Line Segmentation, OCR

I. INTRODUCTION

O.C.R (optical character recognition) has been one of the most challenging research areas in the fields of image processing. The main aim of O.C.R. is to convert the scanned documents into editable format. O.C.R helps us to read and recognize the scanned documents. It involves computer software designed to translate images of typewritten text into machine-printed editable text, or to translate pictures of characters into a standard encoding scheme representing them in ASCII or Unicode. If you scan a text document, you might want to use optical character recognition (OCR) software to translate image into text that you can edit. When a scanner first creates an image from page, image is stored in computer's memory as a bitmap. A bitmap is a grid of dots; one or more bits represent each dot. The job of OCR software is to translate that array of dots into text that computer can interpret as letters and numbers.

The typical phases of OCR system Pre-processing, Segmentation and Recognition. The Pre-processing phase includes the conversion of gray scale image into binary, noise removal and skews detection and correction. Segmentation phase includes the segmentation of text image into lines, word and characters. The final recognition phase consists of feature extraction, selection and classification.

The main process of O.C.R is shown here



Line Segmentation

Line segmentation is a technique to extract lines from a scanned document. Line segmentation in printed text is much more easier than the handwritten text an in printed text, there is a equal font size and equal line spacing between the text lines which makes the text line segmentation very much easier, but in handwritten text document, text line segmentation is not that much easy due to presence of some problems like to skewed, overlapped lines and touching lines and also due to different writing style of a writer.. 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.

IJSRSET162185 | Received: 09 February 2016 | Accepted: 09 March 2016 | March-April 2016 [(2)2: 35-39]

The text line extraction commonly make two assumptions: firstly gap between two neighbouring lines is important and secondly, lines are acceptably straight. Lines are segmented before word and character segmentation. In this, lines are detected by scanning of image in horizontal manner. Count the 0’s and 1’s here 0 means white 1 means black. Create a row histogram by calculate total no of 1’s for detect the lines. When there is all 0’s means there is no black pixel, it denotes a boundary between two consecutive lines.

According to Text type, Line segmentation is categorized into two parts: Machine Printed Text Document and Handwritten Text Document Machine printed text includes the materials such as books, newspapers, magazines, documents. Machine printed characters are uniform in height, width, and pitch assuming the same font and size are used. Handwritten text can be further divided into two categories: cursive and hand printed script. Recognition of handwritten characters is a much more difficult problem. Characters in the handwritten text documents are non-uniform and can vary greatly in size and style. Even characters written by the same person can vary considerably. In the location of characters is not predictable, nor the spacing between them.

Gurumukhi Script

Gurumukhi script is the script used for writing Punjabi language invention of Sikh Gurus. Writing style is from left to right. Gurumukhi script composed of forty one consonants, twelve vowels. Some characters in the form of half characters are present in the feet of characters .Gurumukhi script can be partitioned into three horizontal zones that is upper zone, middle zone and lower zone.

Some Definitions:

Baseline: A line that connects the lower part of character bodies is known as baseline.

Middle line: A line that connects the upper parts of character bodies is known as middle line

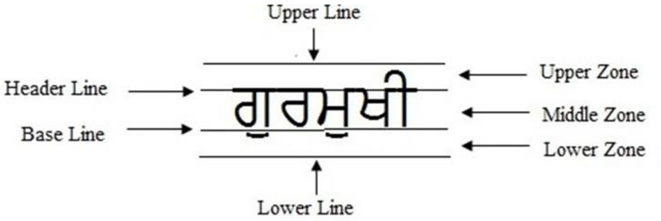
Upper line: A line that connects top of ascenders or upper modifiers.

Lower line: A line that connects bottom of descenders or lower modifiers

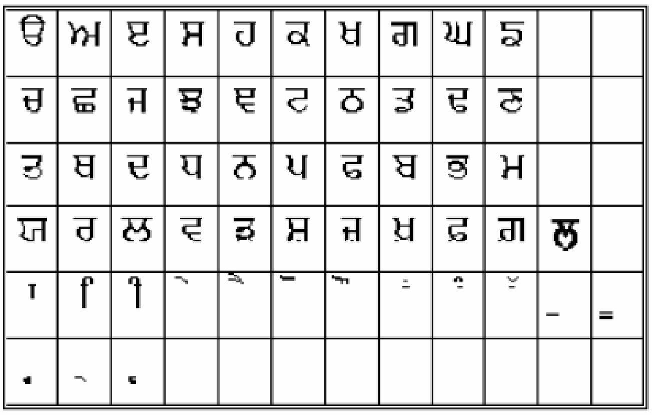
Upper Zone: The region above headline is called upper zone. Vowels reside in this area.

Middle Zone is the area below the head line where the consonants and some sub-parts of vowels are present.

Lower Zone : The area below the middle zone is represented by lower zone. This is the part where some vowels and certain half characters lie in the feet of consonants.



Three zones of Gurumukhi word



ੳ	ਅ	ੲ	ਸ	ਹ	ਕ	ਖ	ਗ	ਘ	ਙ		
ਚ	ਛ	ਜ	ਝ	ਞ	ਟ	ਠ	ਡ	ਢ	ਣ		
ਤ	ਥ	ਦ	ਧ	ਨ	ਪ	ਫ	ਬ	ਭ	ਮ		
ਯ	ਰ	ਲ	ਵ	ੜ	ਸ਼	ਜ਼	ਖ਼	ਫ਼	ਗ਼	ਲ਼	
ਾ	ਿ	ੀ	ੇ	ੈ	ੋ	ੌ	ਂ	ੰ	ੱ	ੁ	ੂ
਼	੍	਼									

Character Set of Gurumukhi Script

Literature Review

This section describes the work done carried out by the various researchers in the field of handwritten text line segmentation in OCR. A good research about problems of segmentation according to three different zones like Skewed, overlapped lines and touching lines and also provide a review of methods of handwriting or printed text line segmentation like projection profiles, Hough transform, smearing method, fuzzy run length and many others given in [1],[2].Feature extraction of Gurumukhi script has shown and proposed methods and techniques for this. A Survey by Pritpal Singh, Sumit Budhiraja. This paper presents an overview of the various O.C.R. systems for Gurumukhi [3]. In this paper author discussed a new algorithm that can perform line segmentation in handwritten text. This algorithm mainly deals with skewed text but also with overlapping and touching of characters based on projection profile technique [4]. Namisha Modi proposed an algorithm can segment skewed and touching lines of fixed size lines using Probable text line detection algorithm with 75% result [5]. Summary of some other paper with results shown in the form of table:-

S. No.	Author Name/Year	Technique Used	Problem Solved	Limitation (Gaps)	Results
1	Snehdeep / 2014	Mid Detection Algorithm	Segment the line overlapped lines and lines having connected components with fixed size	1. Algorithm works only for fixed sized text lines 2. Algorithm can segment maximum two adjacent overlapped lines 3. Algorithm doesn't work for broken parts	90% (Overlapped lines having fixed Size)
2	Amreen Sigh/2013	Projection Profile Technique	Algorithm segment the skewed lines and isolated lines	1. Algorithm segment the lines only of Fixed Size 2. Algorithm cannot segment overlapped lines , touching lines and lines containing broken parts	93% (For skewed and Simple Lines)
3	Rajiv Kumar /2010	Top down projection technique for segmentation	Algorithm can segment the isolated lines words and characters	1. Algorithm cannot segment touching, overlapping lines and lines containing broken parts	92% (For Line Segmentation) 90% (For Word segmentation) 88% (For Character segmentation)
4	Rahul Garg /2014 (Devanagri Script)	Piecewise projection profile method	Algorithm can segment isolated lines , touching lines, and overlapping lines with the fixed size	1. Works on fixed sized line 2. Algorithm can segment isolated, touching lines and overlapping lines of fixed size 3. Algorithm must be improved to segment variable sized lines	91% (for Devanagri script)

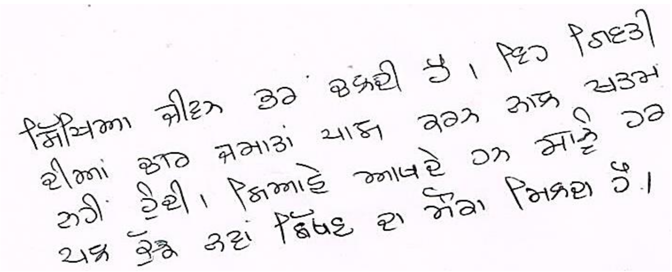
Problems During Line Segmentation

Due to unequal spacing between lines, Variability of writing style, both between different writers and between separate examples from the same writer overtime, absence of focus, ill-advised learning about language various problems are faced in segmentation of handwritten document. Line segmentation of handwritten text document is a complex task because of irregularities in geometrical properties such as line height, width, and distance in between line.

These problems are not present in printed text because in case of printed documents, the lines are aligned properly and all the characters have same width and height. In line segmentation, free style handwriting text is considered a complex and challenging task due to the following characteristics.

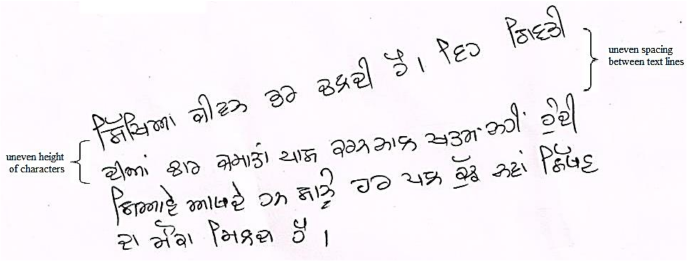
1. **Skewed Lines:-** The lines of text that are not straight as that of paper i.e. that are not parallel to paper header are called skew lines. Skewed lines present in the text can also have uneven spaces in between the adjacent lines. The main problem of skewed lines is due to its

non-straight gap with respect to the paper, due to which horizontal profile projection technique got fail to segment the skewed lines. Lines in the paragraph can be skewed upward and downward due to variability of writing styles of different writers. Upward Skew means when while writing, the lines start going upwards and Downward Skew is when line starts going downwards while writing. The skewed text makes the document analysis and segmentation more complex.



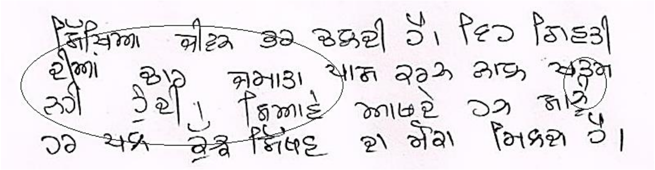
Skewed Lines

2. There is uneven space between lines, between words and even between characters.


uneven spacing between text lines
uneven height of characters

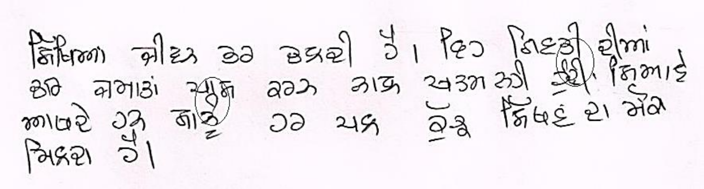
Unequal Spaces

3. Touching of adjacent lines:- in this problem, the words in a line cross the zones of words in other lines. Some authors, perform touching of the characters in a single or multiple words from one line into another line due to the writing style. Hence this creates an problem for the segmentation algorithm to perform segmentation.



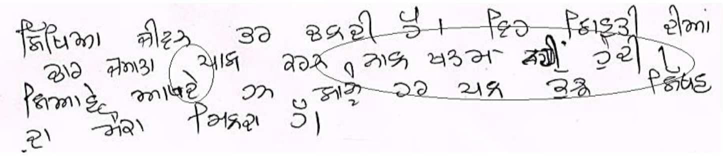
Touching Lines

4. Connected Components :- Adjacent lines said to have Connected components , when lower modifier of one character connects with upper modifiers of next line due to handwriting styles of different writers.



Connected Components

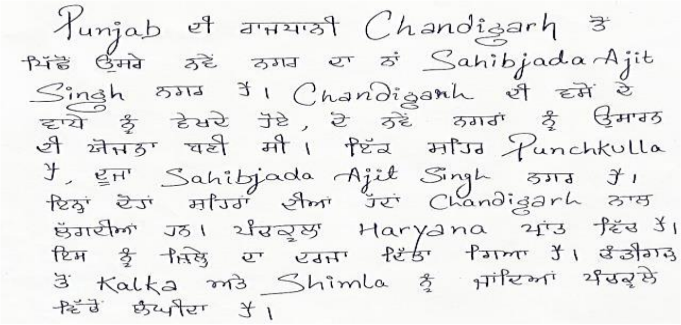
5. Overlapping :- Problem of overlapping lines and characters becomes more complex due to unequal spaces between neighbouring lines. Overlapping means upper and lower modifiers and characters of one line may cover the modifiers or characters of adjacent line. Problem of overlapping can be considered as a sub problem of touching problem. The only difference between the touching problem and the overlapping is that in touching problem only some pixels from one line touch with base of the characters in the another line but in case of problem of overlapping components a word or a character may compeletly overwrite the word or character of the adjacent line. In the overlapping lines, it is quite difficult to separate the characters of words than that of touching lines.



Overlapping

6. To calculate average height at which the connected lines to be chopped as even in single document height of a segment is not similar. Calculating right average height was the tedious job.

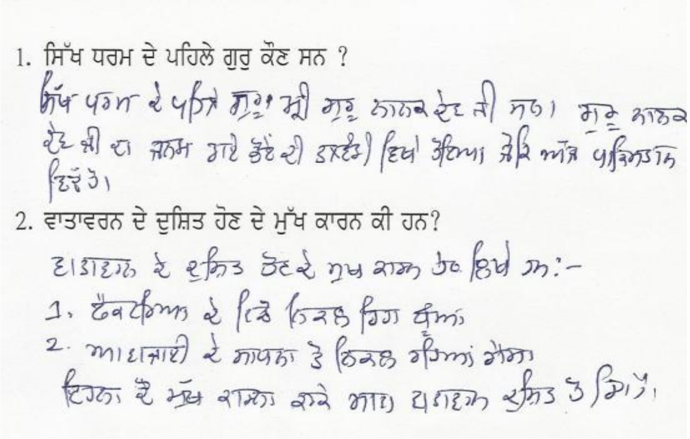
7. Multi Script Document :- When a handwritten text document contains more than one script then line segmentation becomes challenging because each of the script follows its own characteristics and those may not match for the two. Due to different pattern and styles of multiple scripts segmentation of text documents into lines becomes a more difficult task.



Punjab ਦੀ ਰਾਜਧਾਨੀ Chandigarh ਤੇ
ਪਿੰਡੇ ਉਸਾਰੇ ਨਵੇਂ ਨਗਰ ਦਾ ਨਾਂ Sahibjada Ajit
Singh ਨਗਰ ਹੈ। Chandigarh ਦੀ ਵਸੋਂ ਦੇ
ਵਾਧੇ ਨੂੰ ਵੇਖਦੇ ਹੋਏ, ਦੋ ਨਵੇਂ ਨਗਰਾਂ ਨੂੰ ਉਸਾਰਨ
ਦੀ ਯੋਜਨਾ ਬਣੀ ਸੀ। ਇੱਕ ਸ਼ਹਿਰ Punchkulla
ਹੈ, ਦੂਜਾ Sahibjada Ajit Singh ਨਗਰ ਹੈ।
ਇਨ੍ਹਾਂ ਦੋਹਾਂ ਸ਼ਹਿਰਾਂ ਦੀਆਂ ਹੱਦਾਂ Chandigarh ਨਾਲ
ਲੱਗਦੀਆਂ ਹਨ। ਪੰਚਕੂਲਾ Haryana ਪ੍ਰਾਂਤ ਵਿੱਚ ਹੈ।
ਇਸ ਨੂੰ ਜ਼ਿਲ੍ਹੇ ਦਾ ਦਰਜਾ ਦਿੱਤਾ ਗਿਆ ਹੈ। ਚੰਡੀਗੜ੍ਹ
ਤੋਂ Kalka ਅਤੇ Shimla ਨੂੰ ਜਾਂਦਿਆਂ ਪੰਚਕੂਲੇ
ਵਿੱਚੋਂ ਲੰਘੀਦਾ ਹੈ।

Multi Script Document

8. Mixture of Handwritten and printed text: This usually occurs in question papers and certain forms like bank cheques, application and admission form etc . Following figure explains this concept.



1. ਸਿੱਖ ਧਰਮ ਦੇ ਪਹਿਲੇ ਗੁਰੂ ਕੌਣ ਸਨ ?
ਸਿੱਖ ਧਰਮ ਦੇ ਪਹਿਲੇ ਗੁਰੂ ਸ੍ਰੀ ਗੁਰੂ ਨਾਨਕ ਦੇਵ ਜੀ ਸਨ। ਗੁਰੂ ਨਾਨਕ
ਦੇਵ ਜੀ ਦਾ ਜਨਮ ਰਾਏ ਭੋਇੰ ਦੀ ਤਲਵੰਡੀ ਵਿਖੇ ਹੋਇਆ ਜੋ ਕਿ ਅੱਜ ਪਾਕਿਸਤਾਨ
ਵਿੱਚ ਹੈ।
2. ਵਾਤਾਵਰਨ ਦੇ ਦੂਸ਼ਿਤ ਹੋਣ ਦੇ ਮੁੱਖ ਕਾਰਨ ਕੀ ਹਨ?
ਵਾਤਾਵਰਨ ਦੇ ਦੂਸ਼ਿਤ ਹੋਣ ਦੇ ਮੁਖ ਕਾਰਨ ਹੇਠ ਲਿਖੇ ਹਨ:-
1. ਫੈਕਟਰੀਆਂ ਦੇ ਵਿਚੋਂ ਨਿਕਲ ਰਿਹਾ ਧੂੰਆਂ
2. ਆਬਾਦੀ ਦੇ ਵਾਧੇ ਤੇ ਨਿਕਲ ਰਹੀਆਂ ਗੈਸਾਂ
ਇਹਨਾਂ ਦੋ ਮੁੱਖ ਕਾਰਨਾਂ ਕਰਕੇ ਅੱਜ ਵਾਤਾਵਰਨ ਦੂਸ਼ਿਤ ਹੋ ਗਿਆ ਹੈ।

Mixture of printed and Hand written text

II. CONCLUSION

OCR is a technology that deals with text documents to be converting them into E-Format. Line segmentation is a very important part of the overall OCR process because accuracy of other phases is highly dependent on this phase. In this paper we focus on the various problems occurs in the segmenting the handwritten text into lines written in Gurumukhi script. Various problems have been discussed with the help of examples representing these problems. Literature survey of previous authors has also be presented along with the limitations and overall accuracy. Although various algorithms and techniques has been presented by various authors a more robust algorithm is required to perform the segmentation of handwritten text documents into lines by solving the problems which are discussed in this paper.

III. REFERENCES

[1] Er. Snehdeep, Er. Manoj Chaudhary "A Review on Text Line Segmentation Problems and Techniques of Gurumukhi Handwritten Scripts" , (IJARCSSE) Volume 4, Issue 7, July 2014

[2] Rahul Garg, Naresh Kumar Garg" Problems and Review of Line Segmentation of Handwritten Text Document", (IJARCSSE) Volume 4, Issue 4, April 2014

[3] Pritpal Singh, Sumit Budhiraja,"Feature Extraction and Classification Techniques in O.C.R. Systems for Handwritten Gurmukhi Script – A Survey".

[4] RahulGarg,NareshKumarGarg"An algorithm for Text Line Segmentation in Handwritten Skewed and Overlapped Devanagari Script" (ijetae) Volume 4, Issue 5, May 2014)

[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] Snehdeep ,Manoj Kumar "Segmentation of Connected Components and Overlapping Lines in Gurumukhi Handwritten Documents" International Journal of Computer Applications Volume 102– No.13, September 2014

[7] .Amreen Singh and Er. Sukhpreet Singh "Line Segmentation of Handwritten Documents written in Gurumukhi Script", International Journal of Application or Innovation in Engineering & Management Volume 2, Issue 8, August 2013

[8] Rajiv Kumar, and Amardeep Singh, "Detection and Segmentation of Lines and Words in Gurumukhi Handwritten Text", IEEE, 2010

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

[10] Er.Naunita " Segmentation of Handwritten Text Document- A Review" International Journal of Advanced Research in Computer Engineering & Technology Volume 2,Issue 3, March 2013