ISSN 2348 – 8034
Impact Factor- 3.155

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

A REVIEW PAPER ON LINE SEGMENTATION OF HANDWRITTEN TEXT DOCUMENTS WRITTEN IN GURUMUKHI SCRIPT

Er. Sheetal[*1] and Er.Rajneesh Narula[2]
[*1]Student, M.tech (cse), A.I.E.T Faridkot, Punjab
[2]Dean Academic, A.I.E.T Faridkot,punjab

ABSTRACT

The scanned image can't be edited, if required it can be done by OCR(optical character recognition). It is electronic conversion of scanned images of handwritten, typewritten or printed text into machine-encoded text. Line Segmentation is one of the important phase of an OCR, as accuracy of an OCR depends upon the accuracy of segmentation. Incorrect segmentation leads to incorrect recognition. Line segmentation of handwriiten document makes process difficult due to skewed,overlapped lines and touching lines.In this paper, we are presenting a brief introduction on OCR. The objective of this paper gives the review on various methods used to segment a line of a handwritten document written in Gurumukhi Script by various authors.

Keywords- *OCR, segmentation, Line segmentation, Word segmentation, Character segmentation, segmentation techniques.*

I. INTRODUCTION

OCR stands for optical character recognition that Converts scanned image of handwritten and printed text into editable text i.e. machine encoded text . OCR software has tools for acquiring image from the scanner and recognising the text. . Accuracy of line segmentation plays very important role in OCR, because in OCR system there are many other phases which depends on the accuracy of the segmented line, various parts are word segmentation, character segmentation and feature extraction. If there is error in line segmentation then other parts can't perform their task correctly. The recognition of handwritten documents is more complicated in comparison to printed documents because handwritten documents contains unconstrained variations of written styles by different writers even different writing styles of same writer on different times .
The typical phases of OCR system

- Pre-processing
- Segmentation
- Recognition

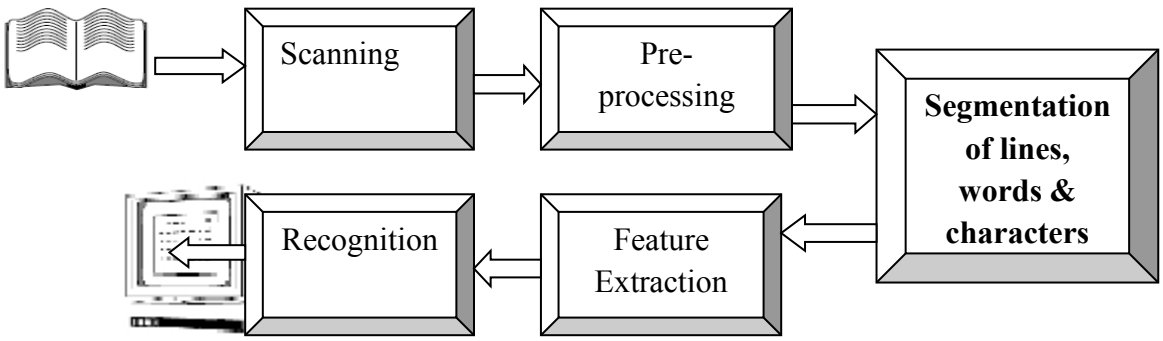


Fig 1: Components of OCR

Scanning:- In this the document is converted into scanned image with the help of scanner. It is the first step of OCR scan the input text image. Also create the database of scanned images.

ISSN 2348 – 8034
Impact Factor- 3.155

Pre-processing:- Scanned document may contain a certain amount of noise depending on the resolution on the scanner. This process convert's raw data into a form that is usable for recognizer.Pre-processing phase includes Binarization, Noise detection and removal, Skew detection and correction.

Segmentation: -The word "segmentation" means to divide. It is most crucial part of the over OCR system. Accuracy of any OCR system depends upon this phase. For correct recognition of characters there is a need to perform segmentation correctly. 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.

TYPES -. A text document can be segmented into Line, word, character

LINE SEGMENTATION is used to segment the paragraph into lines. It is a technique which helps to extract lines from document. Segmentation of lines is easier in printed text compare to handwritten text. 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.

WORD SEGMENTATION is used to segment the lines into words. In this a text line has taken as a input. When a line has been detected, then each line can be scanned vertically for word segmentation. Create a column histogram by calculate total no of black pixels in each column. if there is no black pixel found in vertical scan that is considered as the space between two words.

CHARACTER SEGMENTATION is a process of segmenting the words into separate characters. These individual characters are further used as an input for recognition. The presence of touching characters in handwritten documents further decreases correct segmentation as well as recognition rate . Accuracy in extracting the features is highly depends upon the segmented character

Feature Extraction: Feature extraction is used to extract certain features of character. Its major function is to capture the characteristics of the symbols.It is the process to extract relevant data used for classification purpose.example of feature extraction

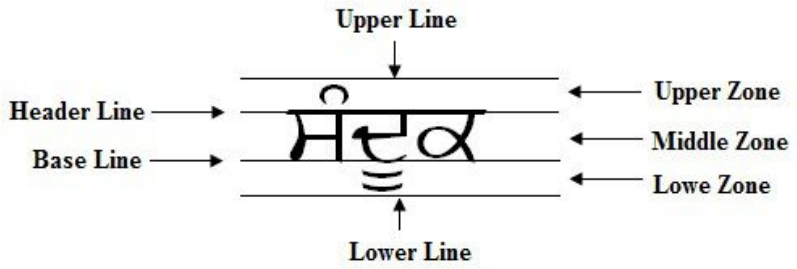


Fig 2: Text Line Regions

For feature extraction, word in Gurumukhi script can be partitioned into three horizontal zones upper zone, middle zone and lower 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. 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.

Recognition: In this phase we actually recognised thecharacter from extract the features of character.

ISSN 2348 – 8034
Impact Factor- 3.155

II. REVIEW OF LITERATURE

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. Algorittm must be improved to segment variable sized lines	91% (for Devanagri script)
5	Namisha Modi/2013	Probable text line detection algorithm	Algorithm can segment skewed and touching lines of fixed size lines	It is based on two assumptions 1. It depends upon the gap between two neighboring lines 2. Lines must be straight not skewed	75%

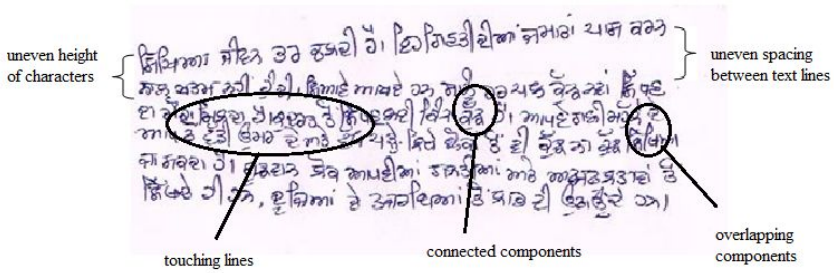
ISSN 2348 – 8034
Impact Factor- 3.155

				These two assumptions are very less applicable on the handwritten text documents	

Problems faced during Text line segmentation

There are various problems in segmentation of handwritten documents. We will discuss the problems with the help of following example as shown in figure :

1. In handwritten documents, majority of writing patterns are not straight which cause problems in locating header line and base line.
2. Space between lines is uneven.
3. Characters and symbols of neighboring lines are connected, touching or overlapping.
4. 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.[4]

Fig 3: Image of punjabi handwritten document

III. EXISTING TECHNIQUES OF LINE SEGMENTATION

1 Projection –profile based method: projection profile methods are commonly used for printed as well as handwritten document. These methods are of two types 1. Horizontal profile projection 2. Verticalprofileprojection. They are top down technique, easy to implement. In this we start scanning from the first line and find the pixel intensity of each line. Basically, Horizontal projection profile concept is used to extract lines from a document image. Vertical projection profile concept is implemented in the case of skewed text lines. Vertical projections are applied on the document image to divide the image into number of stripes o detect the accurate position of header line.

2. **Hough Transform:** Hough transform is also a popular methodology in the area of text line segmentation. It describes parametric geometric shapes and distinguishes geometric areas that recommend the existence of the sought shape. The purpose of this technique is to identify fuzzy snapshots of objects in a certain category of shapes and under a voting procedure. The voting procedure takes place in a parametric space where the candidate objects are acquired as local maxima in a table made explicitly by the Hough transform. Serious drawback of this method is the computational complexity.

3.**Smearing methodology:** Smearing methodology is a bottom-up technique. It uses the concept of converting a group of background pixels located between foreground pixels into foreground pixels based on the threshold value. Smearing methods strengthen by local techniques, solve specific problems and overlapping touched connected component. In addition, these strategies work effectively with documents that hold characters of variable height.

ISSN 2348 – 8034
Impact Factor- 3.155

However, they may have problems in the presence of skewing. Additionally, they can't deal with the variability in separations in the middle of words and characters. They usually make use of many thresholds and heuristic rules

4.**Grouping methods**: Grouping methods are also bottom-up technique. It is the process of grouping the pixels according to specific constrains designed to result to a layer of text lines . From the lower level, the pixel, starts a process of grouping according to specific constrains designed to result to a layer of text lines. The process is relatively easy in the case of printed documents, but it may be proved to be difficult and problematic in manuscripts. This method is effectively used to segment connected components, fluctuating and touching lines.

5.**Graph-based method:** It is technique of line segmentation in which document images are represented with the help of graphs. The graph is constructed as vertices of pixel or more complex connected components. The vertices are normally associated with weighted edges that depict distances between connected components.

6.**Active Contour method:** This method uses the concept of difference between the foreground and the background through characteristics such as brightness or color that occurs at the border contours of the object. In case of text lines, curved line edges specify all the properties and characteristics which specify the shapes. When a specific curve around contour of an object is created then impose the appropriate equation of motion, it will force it to reach the curve forming the outline border of the object. These curves are called Active Contours and they have been used widely in text line segmentation.

IV. CONCLUSION

In this paper, we have present the various techniques for line segmentation of Handwritten text documents. Various problems of line segmentation has been discussesd in context of handwritten text documents. Comparison of various authors has been presented in terms of accuracy, techniques used and problem solved. It is concluded that existing techniques can not segment the line from a handwritten text document independent of size. Various line segmentation problems like overllaped lines, skewed lines , touching lines are still required to be solved. In future we are planning to develop an algorithm for line segmentation of handwritten text documents that can solve most of these problems in an efficient manner.

REFERENCES

1. *Rajiv Kumar, and Amardeep Singh, "Detection and Segmentation of Lines and Words in Gurumukhi Handwritten Text", IEEE, 2010*
2. *M.K. Jindal, R.K. Sharma, G.S.Lehal, "Segmentation of Horizontally Overlapping lines in Printed Gurmukhi Script", IEEE, 2006.*
3. *Amreen Singh and Er. Sukhpreet Singh "Line Segmentation of Handwritten Documentswritten in Gurumukhi Script",International Journal of Application or Innovation in Engineering & Management Volume 2, Issue 8, August 2013.*
4. *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.*
5. *U.Pal, Sagarika Datta, "Segmentation of Bangla Unconstrained Handwritten Text",IEEE, 2003.*
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. *Er. Snehdeep, Er. Manoj Chaudhary"A Review on Text Line Segmentation Problems and Techniques of Gurumukhi Handwritten Scripts" , International Journal of Computer Science, Engineering and Information Technology. Volume 4, Issue 7, July 2014*
8. *Er.Naunita " Segmentation of Handwritten Text Document- A Review"International Journal of Advanced Research in Computer Engineering & Technology Volume 2,Issue 3, March 2013.*