IJERST

International Journal of Engineering Research and Science & Technology

ISSN : 2319-5991
Vol. 4, No. 4
November 2015

www.ijerst.com

Email: editorijerst@gmail.com or editor@ijerst.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 4, No. 4, November 2015
© 2015 IJERST. All Rights Reserved

Review Article

LINE AND WORD SEGMENTATION OF HANDWRITTEN GURUMUKHI TEXT DOCUMENT: A REVIEW

Payal Jindal[1]* and Balkrishan Jindal[1]

*Corresponding Author: **Payal Jindal** ✉ payal26jindal@gmail.com*

The process of segmentation is having an immense importance in handwritten script recognition and widely used field of OCR. This research area is focused on the line segmentation of Gurumukhi script to get proper recognition as an output. Segmentation is important to improve the accuracy of handwritten script identification, as recognition system is mainly dependent on segmentation phase. For proper character segmentation, line segmentation must be done earlier in efficient manner. Variations in handwritings of different writers and presence of touching and broken lines make this task more challenging. In this research work, major concern is to improve the efficiency of existing techniques and to overcome the limitations of existing segmentation techniques. In this paper we present the review on line segmentation techniques.

Keywords: Line Segmentation, Segmentation, OCR, Gurumukhi Script

INTRODUCTION

OCR is a process of automatic reading of optically sensed document images of printed and handwritten text materials to convert human-readable characters to machine readable codes (Karmakar *et al.,* 2014). Characters are encoded in the form of ASCII or some standard digital format. OCR applications are used in banks, business automation, defense applications and reading aid for blind. Important application of OCR is used to encode or digitalize various old manuscripts. OCR is one method to prevent these old documents from deteriorating. OCR is the procedure of perceiving a fragmented a piece of the checked image as a character. The general methodology comprises of three real sub procedures like pre-processing, segmentation and afterward recognition. Out of these three, segmentation is most important phase of OCR system. Incorrect results of segmentation phase leads the problem of wrong recognition. OCR deals with the problem of recognizing optically sensed characters. Optical recognition is performed on handwritten and printed documents after the completion of writing process. On other hand, it is also performed on computer drawn characters through any electronic media.

Origin of OCR

With development of digital computer and improving scanning devices, OCR technology

[1] YCOE, Talwandi Sabo.

took major turn in middle of 1950. From that era, OCR is widely used for data processing approach in business world. David Sheppard, founder of Intelligent Machine Research Co. can be considered as an innovator of OCR systems (Karmakar *et al.,* 2014). Older OCR systems, work by matching the scanned images against stored bitmaps based on specific fonts. The hit or miss results of such pattern recognition systems helped establish OCR's reputation for inaccuracy. Today, OCR software can recognize a wide variety of fonts; Developers are taking different approaches to improve script and handwriting recognition. Advances are being made to recognize characters based on the context of the word in which the software will use knowledge of the parts of speech and grammar to recognize individual characters.

Segmentation: Segmentation is the process to segment digital image into multiple segments. These multiple segments form different zones to extract features of the characters for recognition. Multiple segments are splitting from large document object into smaller objects proceeding through various types of segmentation.

Line Segmentation: It is the technique to extract lines from document image which is further used for character recognition. Accuracy of this phase put impact on the accuracy of recognition system. Lines of a text block are detected from scanned image by calculating the frequency of black pixels in each row constructing row histogram. When the frequency of black pixels in a row is appear to zero, then it denotes the boundary between two consecutive lines.

Word Segmentation: It is the process of extracting words from lines detected in the above phase. Words are extracted with the help of column histogram. Number of black pixels in each column is calculated to construct column histogram. When the frequency of black pixel is zero, then it denotes the boundary between two words.

Character Segmentation: After word segmentation, each word segmented to get individual characters. These individual characters are further used as an input for recognition. There are various approaches present to extract characters from single word. In Indian scripts, most of the languages are containing header line on the top of the word which is removed before segment the word into characters.

Line Segmentation

Line segmentation is one of the challenging and crucial fields of optical character recognition. It is the process to detect horizontal text blocks to identify boundaries of each line. Lines are detected based upon the frequency of row pixels. Incorrect line segmentation put impact on the accuracy of recognition system (Kaur and Himani, 2014). The text line extraction commonly make two assumptions: firstly gap between two neighboring lines is important and secondly, lines are acceptably straight. Line segmentation of handwritten documents is a difficult task as many problems are faced during line segmentation. Segmentation of handwritten text line is complicated because of inter-line gap variability and base line skew variability (Tang *et al.,* 2015).

According to Text type, Line segmentation is categorized into two parts: Machine Printed Text Document and Handwritten Text Document.

Machine Printed Text: It includes the materials such as books, newspapers, magazines, documents, and various writing units in the video or still image. Machine printed characters are uniform in height, width, and pitch assuming the same font and size are used. These problems for fixed- font, multi-font and Omni-font character segmentation is relatively well understood and solved with little constraint. Writing patterns are mostly straight.

Handwritten Text: It can be further divided into two categories: cursive and hand printed script. Segmentation of handwritten documents is a much more difficult problem. Text lines are non uniform and can vary greatly in size and style. There is the problem in locating header and base lines. Different writers have different writing style. Even lines written by the same person can vary considerably. In this, the location of line is not predictable, nor the spacing between them. In unconstrained writing, average height of characters varies. Due to this, characters and symbols of two neighboring lines are connected, touched and overlapped. These problems make the task of segmentation complicated.

EXISTING APPROACHES

Line segmentation is an important phase of OCR. There are different approaches used to segment the lines and words which are discussed as follows:

Projection Profile Approaches: Projection-based methodologies are making utilization of the structural aspects of the documents. They are top-down systems, straightforward and simple in usage. Histogram constructing by using the horizontal and vertical profile projections which consists of the regions of larger and lower concentrations of pixels (Angadi and Kodabagi, 2014). 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. These approaches are mainly used to extract lines both in printed as well as handwritten documents (Karmakar *et al.,* 2014).

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 (Karmakar *et al.,* 2014).

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. 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 (Karmakar *et al.,* 2014).

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 (Garg and Kumar, 2014). 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.

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 (Karmakar *et al.,* 2014).

HISTOGRAM APPROACH

This method is based on pixel histogram obtain. Here a Y histogram projection is performed which results in text line position. To divide a line into different regions a threshold is applied. After that another threshold is used to eliminate false lines. These procedures however, cause some loss on the text line area. So, recovery method is proposed to minimize the effect (Karmakar *et al.,* 2014).

LITERATURE SURVEY

Karmakar *et al.* (2014), has explained a simple segmenting technique for a line and word segmentation of a script document has been proposed. In this space recognition technique the main objective is to recognize the spaces that separate two text lines and the similar procedure is followed for the word segmentation procedure. The space recognition based approach proposed here is simply based on recognition of spaces that separates two lines or two words.

Kaur and Himaniz (2014), has introduced detection of skew in scanned document images is the main stage of preprocessing recognition which in turn helps in improving the quality of scanned image so that document can be easily extracted. Whenever we scan a document, skew is automatically introduced in the image even if we consider all precautions well. Various technique have been discussed for skew angle detection in scanned document images like Hough transform, projection profile analysis, discrete Fourier transform, fast Fourier transform, etc.

Tang *et al*. (2014) has described a novel text line segmentation method based on matched filtering and top-down grouping for handwritten documents. The proposed method consists of three distinct steps. Firstly, the foreground pixel density of handwritten document image is estimated, and then foreground pixel density is used to decide the size of the generated filter which is the convolution of a band-shape filter and an isotropic LoG filter. Secondly, the centers of the text lines are extracted by performing filtering, binarizing, thinning and top-down grouping operation on handwritten document image. We propose a new text line segmentation method for handwritten documents based on matched filtering and top-down grouping techniques.

Garg and Kumar (2014) has 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. This algorithm is based on projection profile technique. If the text have sufficient gap between text lines and the document is properly scanned then the accuracy in line segmentation will be very much. Here by using this algorithm, skewed text lines are segmented correctly however on overlapped and touched text lines.

We have proposed an algorithm for handwritten text line segmentation. This algorithm is based on projection profile method. Here we have used piecewise projection profile. Segmentation is done by using the gap between the text lines. We have divided the whole document image into 6, 7 and 8 strips and have calculated results on many different document images. This algorithm efficiently deals with skewed text. It can also promisingly deal with overlapped and touched text lines.

We have made following assumptions about the data:

1. The minimum height of a constant in a text line is 10 pixels.
2. The average height (AVGHYT) of a text line is between 20 to 40 pixels.
3. The maximum height of a text line (consonant + modifier) is 25 pixels.
4. If a text block is of less than 10 pixels, it would be merged with previous or next text block depending upon condition.
5. If a text block is of more size than AVGHYT+10 pixels, it must be broken into two parts.

Procedure for text line segmentation:

We have initialized the following arrays:

1. m_MyImage: It contains pre-processed binarized image.
2. HProfiles: It is array that stores the total number of black pixels per row.
3. START: It is a 2-D array and contains the starting address of each text block in each strip.
4. END: It is a 2-D array and contains ending address of each text block in each strip.

Step-1: Read the input image and then binarized it and store it in a 2-D array m_MyImage.

Step-2*:* Get information about the height (h) and width of the image (w) which is the size of the 2-D array.

Step-3: Divide the image into vertical strips and for each strip, calculate the number of black pixels in each row and store the result in HProfiles[y] array.

Step-4: For each strip follow these steps:

i. if HProfiles[y] <=0 or HProfiles[y] ==1 or HProfiles[y] ==2, assign RED color to that pixel and assign that value of HProfiles[y] to a new variable p.

ii. if HProfiles[y]>p, put the starting pixel address into START array, now a while loop is called that processes all the pixels till the HProfiles[y] ==0. Put the last pixel address into END array.

iii. Now we have got the starting and ending addresses of all the text blocks. By using a loop, display the first text block of all the stipes and then after some gap display second text blocks.

Step-5: Repeat the step-4 until full text document image is displayed.

Deep and Kumar (2014) has presented a algorithm that is based on mid-point detection. The algorithm deals with these problems and gives effective results 90% in case of overlapped lines and 94% accurate results for segmentation of connected components between neighboring lines. The proposed algorithm is used to segment skewed lines, overlapped lines and connected components between the neighboring lines. This technique provides effective results for text line segmentation. This algorithm distinguishes the concept of segmenting connected components from overlapped lines. This algorithm distinguishes the concept of segmenting connected components from overlapped lines.

ALGORITHM

Step 1: Divide the document into vertical stripes by assuming the strip size equal to 100 pixels.

Step 2: Using the concept of Horizontal Projections, find the white spaces between the consecutive lines.

Step 3: Calculate the mid-points of these white spaces to segment the consecutive non-overlapping lines.

Step 4: Identify the overlapped lines and connected components by analyzing the difference between two mid-points, which is explained in further sub steps:

Step 4.1: By taking the assumption that average line height is of 30 pixels.

Step 4.2: If the difference between two mid-points is greater than 70 pixels, then it is assumed that document contains some overlapped lines and connected components.

Step 5: Segment the overlapped lines and connected components by calculating the mid-value from previous and next mid-point value.

Step 6: Draw horizontal black lines to represent the segmentation.

Step 7: Repeat same process for each strip.

Angadi and Kodabagi (2014) has introduced a new approach for segmentation of text lines, words and characters from Kannada text in low resolution display board images is presented. The proposed method uses projection profile features and on pixel distribution statistics for segmentation of text lines. The method also detects text lines containing consonant modifiers and merges them with corresponding text lines, and efficiently separates overlapped text lines as well. The character extraction process computes character boundaries using vertical profile features for extracting character images from every text line.

Further, the word segmentation process uses k-means clustering to group inter character gaps into character and word cluster spaces, which are used to compute thresholds for extracting words. The method segments lines, words and characters without applying techniques for removal of noise and other degradations. This aspect of work makes it more robust and efficient. The proposed line extraction method gives very good performance for the images with variability in font size and style, free segmentation path, overlapped text lines, and degraded images.

Jain *et al.* (2014) has introduced a novel approach for word segmentation in OCR system. Segmentation is one of the substantial sub-processes of the OCR system. The meaning of the word can be changed if segmented word is not correct. An approach of segmentation is formulated in which textual area of image is crimped as one large window .Then large window is divided into small windows of different lines and words are segmented out of each line as sub windows to each small window. Then characters are segmented from sub-windows for recognition. The proposed word segmentation technique works efficiently for variable word spaces.

Mehdi *et al.* (2013) has described the result of research carried out to enhance the efficiency of cursive handwriting word based segmentation for sigma based offline cursive handwriting recognition. During the recognition phase the speed was compromised for accuracy. After success in recognition process the optimization process was undertaken. A new algorithm with the combination of smart data structure techniques was successfully developed and tested over various samples. Also the comparative analysis was taken in extensive research between bitmap and bitmap-data (binary images). The algorithm was tested on both type of images and results under different circumstances were compared.

Each image type had its advantaged and disadvantages. The segmentation time was well reduced by using basic iteration for just reading the image else every calculation and word segmentation was handled by different data structures depending upon their use and function they handle the best in different situations. Binary image made it faster with loss of some quality while with retaining the quality the speed was still better than before. The idea of using during-process image scaling and re-scaling is also under consideration and experiment. The input images to the algorithm were already processed, normalized and noise removed by edge detection techniques using very careful threshold to have the difference between a word and a noise. During the comparison search with other algorithms only one paper was found which used almost the same technique but it was for character based handwriting recognition.

Jindal *et al.* (2012) has described that segmentation is one of the important phases 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. 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.

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.

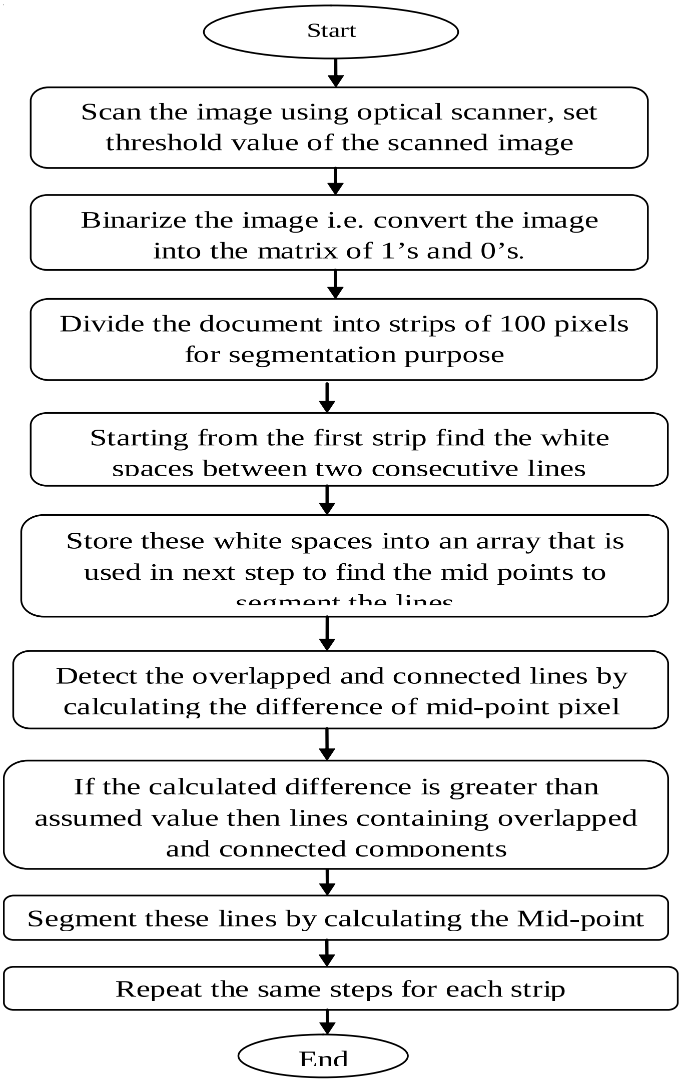
Step 3: Further the text blocks have been divided into three categories, Small Text Blocks containing upper zone or lower zone characters or some part of middle zone. Average Text Blocks containing middle zone along with upper or lower zone. Large Text Blocks contains overlapping lines.

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.

Kumar and Jindal (2012) described a new technique to segment a handwritten document into distinct lines of text is presented. The proposed method is robust to handle line

Flowchart Representing the Implementation of Line Segmentation


Start
Scan the image using optical scanner, set threshold value of the scanned image
Binarize the image i.e. convert the image into the matrix of 1's and 0's.
Divide the document into strips of 100 pixels for segmentation purpose
Starting from the first strip find the white spaces between two consecutive lines
Store these white spaces into an array that is used in next step to find the mid points to segment the lines
Detect the overlapped and connected lines by calculating the difference of mid-point pixel
If the calculated difference is greater than assumed value then lines containing overlapped and connected components
Segment these lines by calculating the Mid-point
Repeat the same steps for each strip
End

Comparison of Existing Techniques on the Basis of Accuracy

S.No.	Author	Segmentation Type	Document Type	Problem Type	Document Language	Accuracy
1	Nallapareddy Priyanka (2010)	Word	Printed	Isolated	Multiscript	99.5%
2	Nallapareddy Priyanka (2010)	Line	Printed	Isolated	Multiscript	99.5%
3	Sonam Jain(2014)	Word	Printed	Isolated	English	99%
4	Muhammad M. Mehdi(2013)	Word	Handwritten	Isolated	English	85%
5	Munish Kumar (2010)	Word	Handwritten	Isolated	Gurmukhi	98.2%

fluctuation. The experiments are performed on various handwritten text images in Gurmukhi Script.

The method of horizontal projection of the whole text is suitable for segmentation of the text with straight lines and with large gap in lines. This method cannot segment handwritten document because it contains touching lines, overlapping lines or fluctuating lines. So to segment this type of text, Here, we are modifying the method to segment text lines based on histogram projection and this technique is called piece-wise projection.

CONCLUSION

In this paper we present a review on line and word segmentation. Various techniques for line and word segmentation have been discussed in this paper. It is concluded that there are various problems in the process of line segmentation which are multiple touching components, variable sized lines, etc. Existing approaches cannot solve these problems. Hence a new algorithm is required to solve all these problems.

REFERENCES

1. Angadi S and Kodabagi M (2014), “A Robust Segmentation Technique for Line, Word and Character Extraction from Kannada Text in Low Resolution Display Board Images”, Proc. Signal and Image Processing, Fifth International Conference on, Jeju Island, Vol. 102, No.13, pp. 10-14.
2. Garg R and Kumar N (2014), “An algorithm for Text Line Segmentation in Handwritten Skewed and Overlapped Devanagari Script”, *International Journal of Emerging Trends in Engineering and Development,* Vol. 4, No. 5, pp. 114-118.
3. Jain S and Singh H (2014), “A Novel Approach for Word Segmentation in Correlation based OCR System”, *International Journal of Computer Applications,* Vol. 99, No.18, pp. 12-20.
4. Jindal S and Lehal G (2012), “Line Segmentation of Handwritten Gurmukhi Manuscripts”, Proc. Advance Computing Conference (IACC), Institute of Electrical and Electronics Engineers, 3rd International, Mumbai, pp. 1797-1801.
5. Karmakar P, Nayak B and Bhoi N (2014), “Line and Word Segmentation of a Printed Text Document”, *International Journal of Computer Science and Information Technologies*, Vol. 5, No. 1, pp. 157-160.
6. Kaur N and Himani (2014), “A Review of Different Skew Detection Techniques”, *International Journal of Emerging Trends in*

Engineering and Development, Vol. 2, No. 4, pp. 108-115.

7. Kumar A and Jindal S (2012), "Segmentation of handwritten Gurmukhi text into lines", Proc. International Conference on Recent Advances and Future Trends in Information Technology, pp. 13-17, 2012.

8. Kumar A, Jindal S and Singla G (2012), "Line Segmentation Using Contour Tracing", *Journal of Global Research in Computer Science*, Vol. 3, No. 1, pp. 50-54.

9. Mehdi M and Riaz A (2013), "Optimized Word Segmentation for the Word Based Cursive Handwriting Recognition", *Institute of Electrical and Electronics Engineers*, pp. 299-304.

10. Sharma N, Shivakumara P, Pal U, Blumenstein M and Limtan C (2012), "A New Method for Word Segmentation from Arbitrarily-Oriented Video Text Lines", *Institute of Electrical and Electronics Engineers,* pp. 978-985.

11. Sneh and Kumar M (2014), "Segmentation of Connected Components and Overlapping Lines in Gurumukhi Handwritten Documents", *International Journal of Emerging Trends in Engineering and Development,* Vol. 4, No. 5, pp. 114-118.

12. Tang Y, Wu X and Bu W (2014), "Text Line Segmentation Based on Matched Filtering and Top-down Grouping for Handwritten Documents", Proc. 11th IAPR International Workshop on Document Analysis Systems, Chennai, India, pp. 365-369.

IJERST

International Journal of Engineering Research and Science & Technology
Hyderabad, INDIA. Ph: +91-09441351700, 09059645577
E-mail: editorijlerst@gmail.com or editor@ijerst.com
Website: www.ijerst.com

ISSN 2319-5991
04
9 772319 599001