
WWJMRD2017; 3(10): 148-152
www.wwjmrd.com
*International Journal*
*Peer Reviewed Journal*
*Refereed Journal*
*Indexed Journal*
*UGC Approved Journal*
*Impact Factor MJIF: 4.25*
e-ISSN: 2454-6615

**Ravneet Kaur**
Research Scholar, Yadavindra College of Engineering, Talwandi Sabo, Punjab, India

**Ashok Kumar Bathla**
Assistant Professor, Yadavindra College of Engineering, Talwandi Sabo, Punjab, India

**Correspondence**:
**Ravneet Kaur**
Research Scholar, Yadavindra College of Engineering, Talwandi Sabo, Punjab, India

# A Review on Different Handwritten Line Segmentation Techniques

**Ravneet Kaur, Ashok Kumar Bathla**

**Abstract**
OCR is Optical Character Reader is to read the Gurumukhi old documents. These documents are having very old age also involves large amount of complexities like skewness, not equal spaced, over lapped lines etc. to overcome this issue different techniques are studied on segmentation. Various techniques have their own set of complexities. Each technique provides solution in given direction but some time results get deteriorated if the un-condition occurs while reading Gurumukhi text. So from study it is clear that hybrid technique remain more suitable for different situations. So that optimal performance in terms of segmentation can be achieved.

**Keywords:** Hybrid, Segmentation, Gurumukhi

## Introduction
Historical document written by historians contain vital information. This information is hand written. The pages onto which this information is written has large amount of problems. These problems can be due to the age old documents. Also while writing the documents different people where involved, so different scripts written with different font sizes and different line width. To overcome this problem different level of processing is performed so that digitized document can be normalized according to the technique.
Once the document is normalized it is to be segmented, so that required text lines can be extracted. The segmentation always is the important issue as far as selection of the text from document is concerned. There are various techniques of segmentation. They can be categorized into four parts. 1) Classical approach, 2) recognition based approach, 3) Holistic approach, 4) Hybrid approach. Various classical approaches are based on subdivision of the large document into smaller parts dealing with each individual part on individual basis. Hybrid approach is more sorts out techniques for such conditions when document is not having uniform distribution of contents.
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.

## Segmentation Techniques
Several recognition techniques have been defined by various researchers. Each technique has their own set of success and failures.

### a. Line segmentation
Line segmentation is a procedure to segment the lines from the hand written or computer written documents.

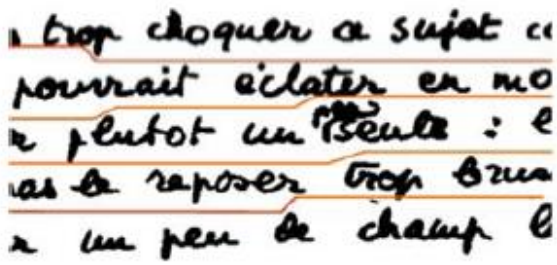

**Fig. 1:** Line Segmentation

**b. Word Segmentation**

In word segmentation a line is broken into words. These words are differentiated with space. In word segmentation vertical line technique is used where two words will be differentiated on the basis of space.

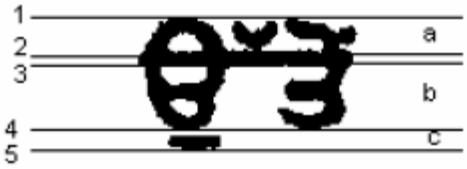


**Fig. 2:** Word Segmentation

**c. Character segmentation**

In character segmentation character is segmented from word**.** Each character is segmented based in upper line or base line segmentation technique**.** In this either water reservoir technique or upper line or base line technique is used.

**Fig. 3:** Character segmentation

**Gurumukhi Language**

Its meaning is straight from the much of Guru.

It is language being standardized by Second Guru Angad Dev Ji. Guru Granth Sahib is written in Gurumukhi Language. Modern Gurumukhi language has 35 consonants. 10 vowels symbols, two symbols for nasal sounds and one sound for duplicate sound.(Adak).

**Problems during line segmentation**

Traditional techniques cannot be applied on to the old manuscript reading because various types of problem exist while reading such documents.

1. Headline of the words is not straight. So traditional techniques may be failed while recognition of non-straight lines.
2. The character set used in Gurumukhi is large consists of different symbols. These symbols are not written or used in now a days Gurumukhi script.
3. There are more touching characters in single word. Even there are more than one touching characters in single word.
4. There are heavy printed letters and words in such old documents.
5. Lines written are such documents are not equal spaced.
6. There is large amount of variations of single characters in shape and size front.

**Handwritten character recognition system**

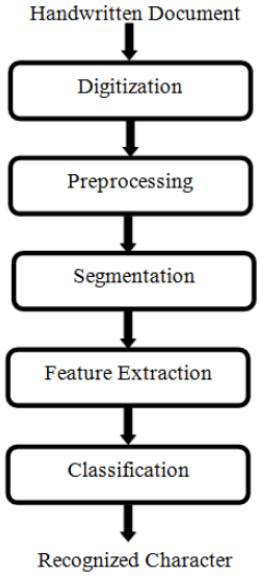


**Fig. 4:** Character recognition System

This character recognition system includes multiple steps followed one after the other. In first step there is hand written document which lies in the form of old or new hand written script.

In first step the digitization of this hand written document is performed. This digitization is using scanner or photograph.

While preprocessing there will be skewness removal and also noise removal. It will enhance the result of segmentation.

Segmentation is the process to segment the text character from the text recognized from the image. It is based classical of hybrid approach.

Features are being extracted from the segmented image. These features may be height, width, and font spacing etc.

Based on the features the characters recognized will be classified into various different categories.

**Design of OCR**

Different approaches utilized for the designs of OCR systems are discussed below [3]:

i. Matrix Matching: Matrix Matching converts each character into an example inside a matrix, and afterward compares the example with an index of known characters. Its recognition is strongest on monotype and uniform single column pages.

ii. Fuzzy Logic: Fuzzy logic is a multi-valued logic that permits intermediate values to be defined between conventional evaluations like yes/no, true/false, black/white and so on. An attempt is made to attribute a more human-like method for logical thinking in the programming of computers. Fuzzy logic is utilized when answers don't have a distinct true or false value and there is uncertainly involved.

iii. Feature Extraction: This method defines each character by the presence or nonattendance of key features, including height, width, density, loops, lines, stems

and other character traits. Feature extraction is an immaculate approach for OCR of magazines, laser print and amazing images.

iv. Structural Analysis: Structural Analysis identifies characters by looking at their sub features-state of the image, sub-vertical and level histograms. Its character repair capability is incredible for low quality text and newsprints.

v. Neural Networks: This technique simulates the way the human neural system works. It samples the pixels in each image and matches them to a known index of character pixel patterns. The capacity to recognize characters through abstraction is extraordinary for faxed documents and damaged text. Neural networks are ideal for particular types of problems, for example, processing stock market data or discovering trends in graphical patterns.

**Literature Survey**

**Richard G. Casey et.al (1996) [12]:** this paper is based on study of various character segmentation technique. In this each text recognition technique provides the varied amount of results. Also different types of constraints lie into different techniques. This paper has further proposed image compression using neural network with cumulative distribution function. This method or algorithm has higher amount of accuracy and generalization ability. While back propagation the weight is adjusted based on error correction rule. The synaptic weight are adjusted to make actual response to the any inputted image. This technique is applied on large image set.

**Edward K. Wong et.al (2000) [16]:** this paper has proposed a technique to extract the text from running video. This research paper has used the technique of scan line. First whole text will be extracted by running along the scan line. Later on multiple lines of text will be integrated to form a text block as well. This algorithm they have applied on the dataset of broadcast television. While working on this dataset this algorithm has given higher success rate. They have proposed an algorithm based on various steps. 1. Identify the expected line segment in the video frame. 2. Expected text block detection. 3. Filtering the text black found from the videos. 4. Boundary adjustment by aligning the text line along the axis of the frame. 5. Clustering based on two colors. One is black color and other is white color. 6. Last step is filtering the segmented image.

**Julinda Gllavata et.al (2003) [6]:** Text Detection in images or videos detection is the important step. In this text recognized is using an algorithm is done automatically. That means detect, localize and extract the horizontally aligned text from the image. The paper has used color reduction technique, method of edge detection and the localization of text regions using projection profile analyzer and geometric properties. The proposed algorithm is based on application of color reduction technique. Binarize the image. Later on edge detection technique is followed. Later on text is recognized using projection profile and geometric properties. This paper does not provide complete mechanism. Rather put various boxes with text with simplified background. This box can be feed into in OCR for reading the character lying into the box.

**B. Gatos, et.al (2005) [2]:** This paper is based on segmentation-free approach for keyword search in historical typewritten documents. Paper is to search for keywords typed by the user in a vast collection of digitized typewritten historical documents. User's feedback enables the transition from synthetic queries to real data queries. This transition is fast and resources saving. This paper has aim of identifying the keywords typed by the user in a large collection of digitized typewritten historical document. This paper has proposed an algorithm based on various steps like binarization of the image, second is creation of synthetic image of keywords, and third step is word segmentation using dynamic parameters.

**T. Konidaris et.al (2005) [15]:** this paper has proposed novel segmentation-free approach for keyword search in historical typewritten documents combining image preprocessing, synthetic data creation, word spotting and user's feedback technologies. This paper has taken the data set of type written document related to old historical documents. The proposed algorithm is based on application of color reduction technique. Binarized the image. Later on edge detection technique is followed. Later on text is recognized using projection profile and geometric properties.

**K. Ntzios et.al (2005) [9]:** this paper has proposed a algorithm to detect the line segmentation for old Greek scripts. These are the precious scripts written in lower case letters. In this paper they have proposed a scheme for extracting the text and then open and lose the cavities for better text extraction. Once the text is extracted the text is digitized and kept in blocks. This paper has proposed an efficient segmentation free approach. It is based on open and closed cavity character representation; they have proposed a segmentation free approach. This approach is based on various steps like input, preprocessing, Skeletonization. In this procedure it has to recognize the open and closed cavities. Open cavities are lying between two subsequent words. And close cavities are lying between two characters. Later on based on features estimation the words and lines are recognized.

**Comparative Analysis**

Various research papers are being reviewed based on different techniques of segmentation. So that traditional documents can be read with higher amount of success.

| Paper Name | Author | Technique | Constraints | Accuracy |
|---|---|---|---|---|
| Segmentation of Isolated and Touching Characters in Offline Handwritten Gurumukhi Script Recog-nition | Munish Kumar[4] | This paper has been focused onto the Gurumukhi character recognition. They are focused onto the water reservoir technique | Further work can be done to enhance the recognition rate of overlapping characters. | 92% |
| A Human Interactive Proof Algorithm Using | Amalia Rusu[6] | They have given a new algorithm named as HIP. Which distinguish between hand written | Further work can be enhanced to understand | 91% |

| | | | | |
|---|---|---|---|---|
| Handwriting Recognition | | and human written text. | more complex cased where there is more noise. | |
| Segmentation of Touching Characters in Printed Devnagari and Bangla Scripts Using Fuzzy Multifactorial Analysis | Utpal Garain[7] | This paper has given focus onto the Identification of devnagari and bangle script. They have focused on touching characters. They have used the technique named as fuzzy multifactorial analysis. During the research the speed of the recognition has been enhanced. | According to the paper this study can be put on to the other types of fonts as well. | 94% |
| Text line segmentation of historical documents: a survey | Laurence Likforman-Sulem[3] | This paper has surveyed different techniques used while segmented the lines on to the traditional written rare scripts. | Large amount of work still need to be done for understanding the global technique for understanding the documents. | N.A. |
| Text line and word segmentation of handwritten documents | G. Louloudis[2] | Segmentation methodology of handwritten documents in their distinct entities, namely, text lines and words. Text line segmentation is achieved by applying Hough transform on a subset of the document image connected components. | Further work need to be done to recognize the punctuations present in the words in upper and lower layer. | 92% |
| An old greek handwritten OCR system based on an efficient segmentation-free approach | K. Ntzios | This paper on focused on identifying the noval approach of text recognition. This approach has been tested on large number of greek documents. | Overlapping and touching character still is a problem in the current document. | 91% |
| Segmentation of Horizontally Overlapping Lines in Printed Indian Scripts | M.K. Jindal | have proposed a solution for segmenting horizontally overlapping lines and solved the problem of other strips in eight most widely used Printed Indian scripts. | Further this solution need to be applied on to the traditional documents. | 93% |
| A Survey of Methods and Strategies in Character Segmentation | Richard G. Casey | They have surveyed different segmentation technique of character recognition. This technique has been studied as provided different percentage of truthness. | According to this paper further study of hybrid technique can be followed. | N.A. |

**Conclusion**

Study of various research papers focused on segmenting the lines or words from the rare scripts is performed. It is much clear that there exist multifold problems in these kind of documents. No paper has given a complete solution to the problem. Each technique has large scope for work to be improved. In old script there are many fold problems like overlapping, missing letters, skewness, touching characters, unequal width and different types of fonts etc. few problems need to be taken so that better technique can be developed which can cover up these problems.

**References**


1. A. Rodriguez, S. Kim, J. H. Ki, Y. B. Fernandez, “English To Spanish Translation Of Signboard Images From Mobile Phone Camera”, IEEE South east con, vol. 09 Pp. 356–361,2011.
2. B. Gatos, T. Konidaris, K. Ntzios, I. Pratikakis and S.J. Perantonis,” A segmentation-free approach for keyword search in historical typewritten documents”, IEEE,vol. 4,pp.67-78,2005
3. Charif, Omar, HichemOmrani, and Philippe Trigano, ”OMBP: Optic Modified Back Propagation training algorithm for fast convergence of Feed forward Neural Network”, ACM, vol. 6,pp.56-67,2011.
4. Durai, S. Anna, and E. Anna Saro,” Image compression with back propagation neural network using cumulative distribution function”, World Academy of Science, Engineering and Technology vol. 17, pp: 60-64, 2006.
5. G. Louloudis, B.Gatos, I.Pratikakis, C.Halatsis,” Text line and word segmentation of handwritten documents”, 2009, Pattern Recognition42, 3169—3183
6. Gllavata,Julinda, R.Ewerth, B Freisleben,” A robust algorithm for text detection in images”, Image and Signal Processing and Analysis, ISPA, Proceedings of the 3rd International Symposium on. IEEE, vol. 2, pp.90-100, 2005.
7. G. Lauloudis, B, gatos, I. Pratikakis, and C. Halatsis, ”Text lines detection in handwritten documents”, Pattern Recognition, vol. 41, pp. 3758-3772,2014.
8. K. Ntzios, B. Gatos, I. Pratikakis, T. Konidaris, S. J. Perantonis, ”An old Greek handwritten OCR system based on an efficient segmentation-free approach”, Springer-Verlag, vol. 4,pp: 345-355,2007
9. K. Ntzios, I. Pratikakis and S.J. Perantonis,”ICDAR2009 Handwriting Segmentation Contest”, IEEE, vol. 4,pp:78-88,2005
10. Manish Kumar Jindal,” On segmentation of touching character and overlapping lines in degraded printed Gurmukhi Script”, International Journal of Image and Graphics, Vol. 9, pp:3321–353,2009.
11. N. Stamatopoulos1 and G. Louloudis1,"ICDAR2009 Handwriting Segmentation Contest", IEEE, vol. 4, pp: 450-460, 2009.
12. Richard G. Casey and Eric Lecolinet, "A survey of methods and strategies in character recognition", vol. 18, pp: 345-355, 1996.
13. R. Manmatha and J,L. Rothfeder, ”A scale space approach for automatically segmenting words from historical handwritten documents”, pattern analysis and machine intelligence, IEEE Transaction on, vol. 27, pp. 1212-1215,2015.

14. [14] Simpel Jindal, Gurpreet Singh Lehal," Line Segmentation of Handwritten Gurmukhi Manuscripts", ACM, vol. 3, pp: 789-799, 2012.
15. T. Konidaris, K. Ntzios, I. Pratikakis and S.J. Perantonis,"A Segmentation-free Approach for Keyword Search in Historical Typewritten Documents", IEEE, vol. 4,pp:786-799,2005
16. Wong, E. K. and M. Chen, "A robust algorithm for text extraction in color video", Proceedings of IEEE Int. Conference on Multimedia and Expo (ICME), vol. 4, pp: 456-466, 2000.