WWJMRD 2017; 3(11): 263-267
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 Novel Approach to Line Segmentation for Handwritten Gurumukhi Scripts

Ravneet Kaur, Ashok Kumar Bathla

Abstract
OCR (Optical Character Recognition) is used to read the digitally scanned documents, and convert the documents into computer readable or understandable format. While segmentation, (using OCR for digitally scanned documents) lines, words and characters are segmented. For computer written documents, lines are equally spaced, so it is very easy to segment the lines with higher rate of success. This success rate drops when the document is handwritten. Handwritten document may have lines not equally spaced. So there will be some difficulty in segmenting the lines from the document with higher efficiency. This research paper has proposed a technique based on mid-point. Where lines will be segmented based on mid-point. While segmenting the lines, skewness is also removed. This skewness can occur while scanning the document. Using mid-point approach the success rate of 94% has been achieved. This proposed technique has been applied on dataset of 500 images.

Keywords: Mid-point, skewed, segmentation

Introduction

Historical documents contain a lot of information which is important for the society. These historical documents are very rare. These documents are found in every language and script. These document's conditions are deteriorating day by day. These documents need to be converted into digital documents using OCR so that their computer readable and understandable format can be prepared. The procedure to convert these documents to computer readable format using OCR is sometimes a difficult process because of the presence of various noises in the documents like skewness, overlapping, touching etc. These documents need to be carefully pre-processed to remove these kinds of noises. Once the noises are removed it will be easy to segment the lines, words or characters.

Historical documents of Gurmukhi script are handwritten. These documents have high order of skewness also they have calligraphy. Due to calligraphy there is high amount of overlapping, which leads to inefficiency in the segmentation of lines. Handwritten Gurmukhi documents have unequal sizes of characters; these characters are lower and upper zone characters.

Gurumukhi Language

Meaning of 'Gurmukhi' is 'straight from the mouth of Guru'. It is a script which is standardized by Second Guru Sri Guru Angad Dev Ji. Guru Granth Sahib is written in Gurumukhi script. Modern Gurumukhi language has 35 consonants, 10 vowel symbols, two symbols for nasal sounds and one sound for duplicate sound (Adak).

Segmentation Techniques

Segmentation for OCR for segmenting lines can be performed in three ways.

a. Segmentation based on base line
b. Segmentation based on header line.
c. Segmentation based on mid-point

Segmentation using base line is used for Arabic and Urdu languages. The space between two subsequent lines will be considered as base for segmenting the lines.

Segmentation using header line is used for languages like Gurmukhi, Devnagari etc. The line above the character having highest frequency level will be considered as identification for segmenting two lines.

Mid-point based technique is used when the mid-point is identified between two subsequent lines.

Proposed Methodology

Proposed research has put the emphasis on the segmentation of lines for handwritten Gurmukhi documents. In overall context, the skewed documents are considered. The skewness can be there either while scanning the handwritten document or while writing the documents. The document containing these problems will be having unsuccessful segmentation.

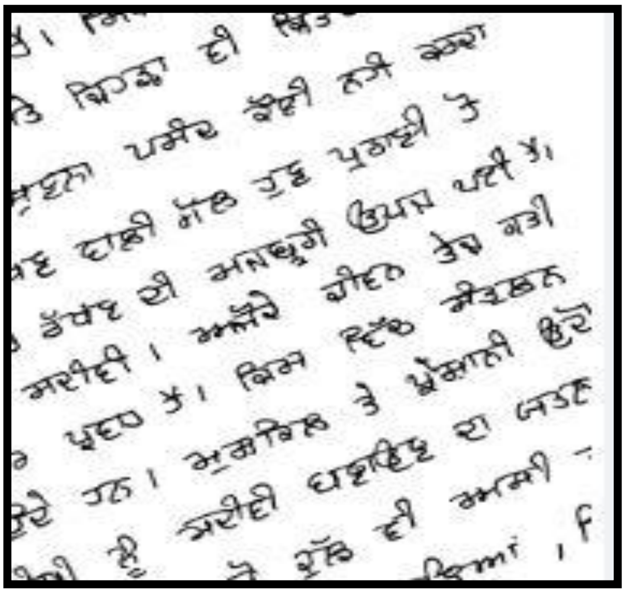


Fig. 1: Skewed Image

The proposed methodology is based on three basic steps:

Step1: In first step, the digitally scanned document of Gurmukhi text is feeded into the system. This digital document is created after scanning o the handwritten Gurmukhi text.

Step2: Pre-processing of the digitally scanned document is done. In pre-processing skewness is detected. If the skewness is found then the skewness is removed from the document.

Step3: In pre-processed document, based on mid-point approach of segmentation, lines are segmented.

Related Work

Casey et al. (1996) [12]: This paper is based on study of various character segmentation techniques. In this, each text recognition technique provides the varied amount of results. Also different types of constraints lie in 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. In back propagation the weight is adjusted based on error correction rule. The synaptic weights are adjusted to make actual response to the any inputted image. This technique is applied on large image set.

Wong et al. (2000) [16]: This paper has proposed a technique to extract the text from running videos. 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. This algorithm has been applied on the dataset of broadcast television. While working on this dataset this algorithm has generated a high success rate. They have proposed an algorithm based on following steps- **(i)** Identify the expected line segment in the video frame. (ii) Expected text block detection. (iii) Filtering the text block found from the videos. (iv) Boundary adjustment by aligning the text line along the axis of the frame. (v) Clustering based on two colors. One is black color and other is white. (vi) Last step is to filter the segmented image.

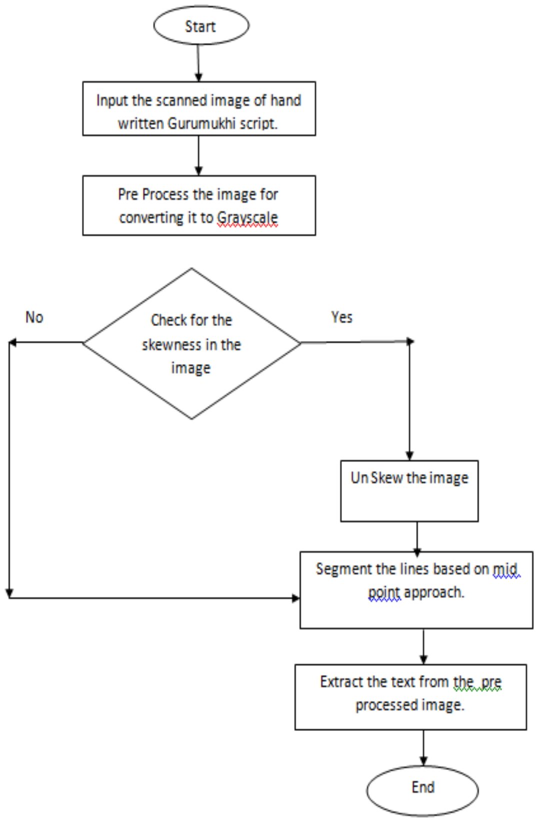
Gllavata et al. (2003) [6]: Text Detection in images or videos is an important step. In this, text recognisation using an algorithm is done automatically. That means detect, localize and extract the horizontally aligned text from the image. The paper has used color reduction technique, method of edge 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. 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 the OCR for reading the character lying in the box.

Gatos et al. (2005) [2]: This paper is based on segmentation-free approach for keyword search in historical typewritten documents. Paper proposes 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 resource saving. This paper has the aim of identifying the keywords typed by the user in a large collection of digitized typewritten historical documents. This paper has proposed an algorithm based on various steps like binarization of the image, creation of synthetic image of keywords, and word segmentation using dynamic parameters.

Algorithm

Step1: Input the scanned image of handwritten Gurumukhi text.
Step2: Pre-process the image for removing the noise and skewness in the image.
Step3: Check the skewness in the image if skewness exists then goes to step 4 else go to step 5.
Step4: Un-skew the image containing hand written text.
Step5: Extract the text from the scanned image.
Step6: Extract the lines based on midpoint approach.
Step5: End

Flowchart


Start
Input the scanned image of hand written Gurumukhi script.
Pre Process the image for converting it to Grayscale
Check for the skewness in the image
No
Yes
Un Skew the image
Segment the lines based on mid point approach.
Extract the text from the pre processed image.
End

Results Analysis

a. Sample Dataset

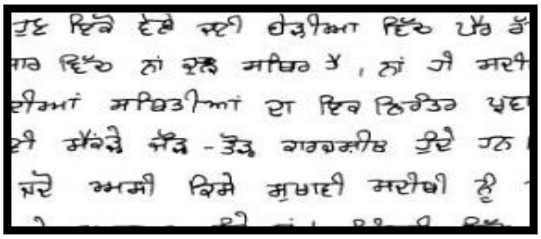


Fig. 2: Sample 1

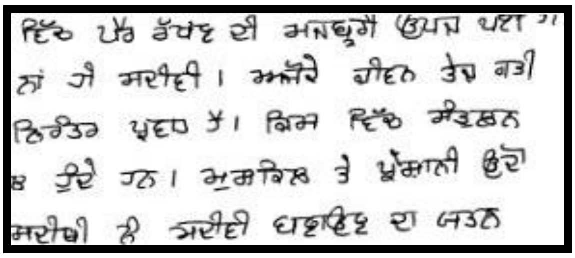


Fig. 3: Sample 2

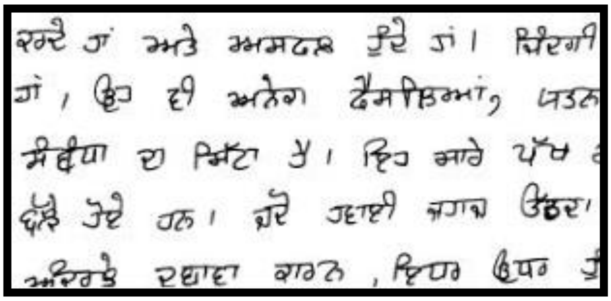


Fig. 4: Sample 3

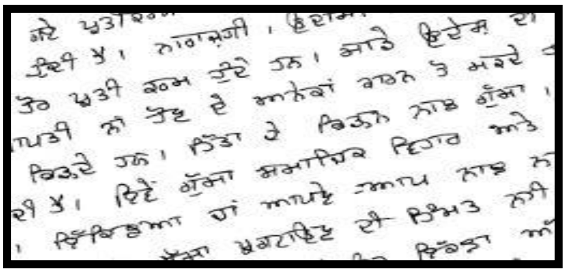


Fig. 5: Sample 4

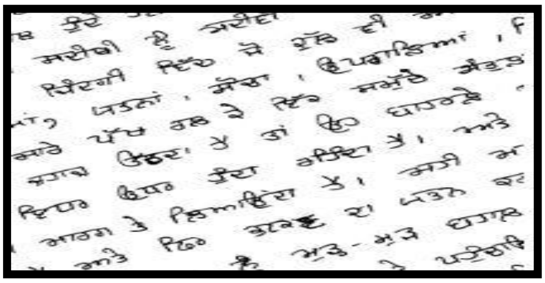


Fig. 6: Sample 5

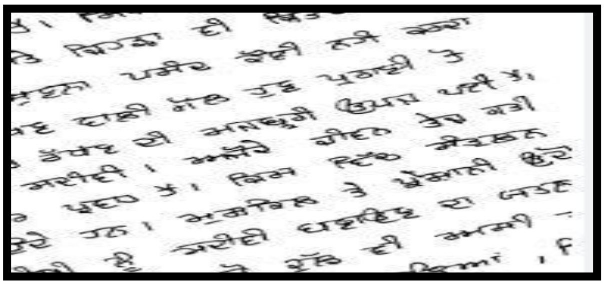


Fig. 7: Sample 6

b. Segmentation Steps.

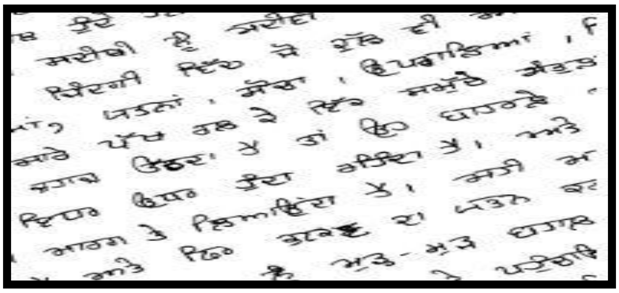


Fig. 8: Inputted skewed Image

.

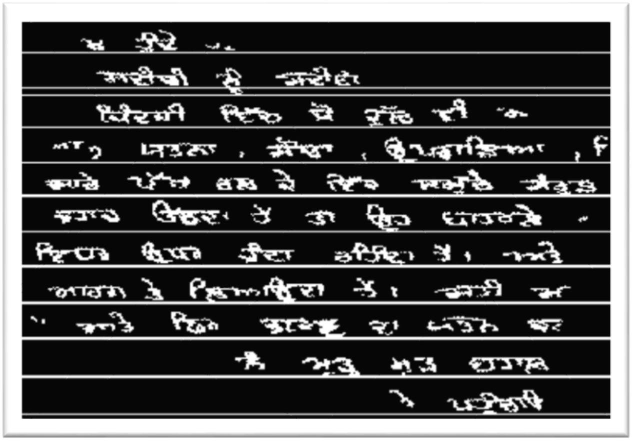


Fig. 9: un-skewed and segmented image

Discussions

First, using skew identification procedure, while pre-processing the image, the angle of the skewness is identified. Then skew angle is removed. This sample output shows the segmented output of the hand written document. This document is bit noisy in terms of unequal space in between to two lines. This sample shows the output of both original image and segmented image. In first output the original image which is inputted has been shown. In second segment white line is segmenting two text lines. This technique has efficient results. Small noise in terms of overlapping also has been removed while showing the result.

From the above processing it is clear that the images used to process for text extraction in the form of line extraction has shown the satisfactory results. That means two types of sample dataset are being taken first was for line having bit overlapping and second dataset images include skewness. Both type of Handwritten Gurumukhi script documents are segmented well.

Accuracy

The algorithm has been performed on 500 image set. This image set is basically digitally scanned documents. These images are black and white images. Each image is passed through same segmentation procedure. That is preprocessing, segmentation etc. In pre-processing un-skewing of the image takes place so that the segmentation is done to the optimal level. Out of 500 images 470 images are doing fine. 30 images are such where results are not getting fine. So the result is around 94%. According to the rest of the researches this is the best result.

Comparison

Author	Technique followed	Accuracy
M. K. Jindal	Water Reservoir	92%
Naresh Kumar Garg	Vertical projection profile	90%
Parika Mangla	End Detection	89%
PROPOSED	Mid-point based	94%

Conclusion

OCR is optical character reader used for reading the lines, words and characters from the digitally scanned documents. Various samples of handwritten documents which are digitally scanned are being taken. These documents are having skewed documents as well. The midpoint approach has been performed for the text extraction. This Technique has shown the satisfactory results. In previous research only overlapping was considered. But they have not considered the skewness. In current research skewness has been considered and also resolved successfully. The skewness is identified while preprocessing the documents. So that success rate can be increased.

Future Work

In current research large database has been taken. This database includes two types of documents. These documents are of straight lines with overlapping and the skewed lines have been taken. The technique has been applied on to the hand written documents. In later research this research with same technique can be extended by taking more overlapped database. This will complete the approach.

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, Hichem Omrani, 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. 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. 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. manpreet Kaur,Ashok Kumar Bathla,"A Review On Different Techniques for Character Segmentation for HINDI Hand Written Words",wwjmrd,vol.:3(10),pp:143-147,2017.
17. Manas Kaur,Ashok Kumar Bathla,"A Review on Different Character Segmentation Techniques for Handwritten Gurmukhi Scripts",wwjmrd,vol.:3(10),pp:162-166,2017.
18. Ashok Kumar Bathla,"Challenges in recognition of Devanagari Scripts due to segmentation of handwritten text", IEEE,6-18 March 2016