International Journal of Advance Research in Science and Engineering
Vol. No.4, Special Issue (01), November 2015
www.ijarse.com
IJARSE
ISSN 2319 - 8354

LINE AND WORD SEGMENTATION OF HANDWRITTEN TEXT DOCUMENTS WRITTEN IN GURMUKHI SCRIPT USING MID POINT DETECTION TECHNIQUE

Payal Jindal[1], Dr. Balkrishan Jindal[2]

[1]*Research Scholar, YCOE, Talwandi Sabo(India)*

[2]*Assistant Professor, C.E., YCoE, Punjabi University, Talwandi Sabo*

ABSTRACT

Text line segmentation of the handwritten documents is still one of the most complicated problems in developing a reliable OCR. The nature of handwriting makes the process of text line segmentation very challenging. Text characteristics can vary in font, size, shape, style, orientation, alignment, texture, color, and contrast and background information. These variations turn the process of word detection complex and difficult. A new technique to segment a handwritten document into distinct lines of text is presented. In this paper, the experiments are performed on various handwritten text images in Gurmukhi Script. The images with high skewness, less line gap, more gaps in words etc. are considered. The results of the proposed method are quite promised.

Keywords: Handwritten Character recognition, Line Segmentation, Mid-point Detection method, Word Segmentation.

I. INTRODUCTION

Optical Character Recognition, usually abbreviated as OCR, is the translation of handwritten or printed text into machine process able format. OCR is the field of pattern recognition and image processing. OCR bridges the gap between man and machine by providing a fast communication method. OCR involves activities like digitization, preprocessing, segmentation, feature extraction, classification and recognition. Segmentation is the most critical step and major challenge for document image processing. Segmentation is used to break the text into lines, words and characters. For the task of segmentation, an algorithm is used for finding segmentation points in handwritten script.

The challenge of a segmentation technique lies in the decision of best segmentation point for line, word and character isolation. Segmentation of handwritten text in Gurmukhi script is a challenging task because of the various writing styles. In the handwritten text, there are some problems which are uncommon in modern printed text. Among the most common are skewed lines, curvilinear lines, fluctuating lines, touching and overlapping components. Incorrect segmentation can lead to incorrect recognition.

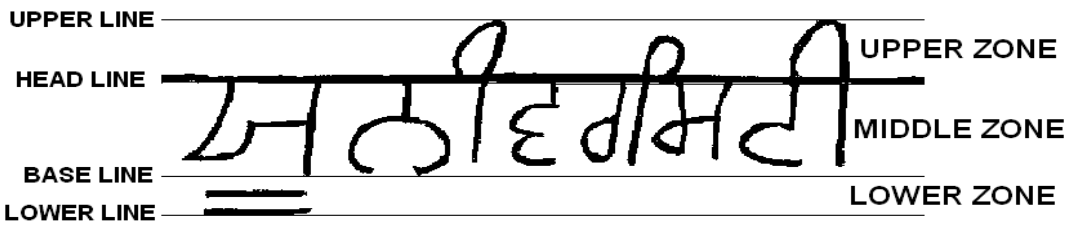

Fig 1.1 gurmukhi handwritten script word

Fig 1.1 describes that there are three zones by which text can be represented which are Upper zone, Middle zone and the Lower zones. Upper and lower zones contain some special characters line (Onkar, Dulankar, siari, Bihari) but middle zone contain the script alphabets.

II. RELATED WORK

Segmentation is a pre-processing phase of optical character recognition.OCR is a technique to encode the offline handwritten as well as printed documents. Results of OCR mostly depend upon effective line segmentation. Different properties of languages and variations in writing styles of different writers may complicate the process of segmentation.

Karmakar et al. [1] has explained the line and word segmentation of a document. The main objective of this paper is to recognize the spaces between two lines and words. Kaur and Himaniz [2] have introduced detection of skew in scanned document images. During scanning of a document, skew is automatically introduced in the image even after considering all the precautions well. Tang et al. [3] described a text line segmentation method based on matched filtering and top-down grouping for handwritten documents.

Garg and Kumar [4] discussed line segmentation in handwritten text based on projection profile technique. In this paper, if the text has sufficient gap between text lines and the document is properly scanned then the accuracy in line segmentation will be very high. Sharma and Sharma [5] have several techniques to segment handwritten text line have been proposed in the past. This paper seeks to provide a method to segment the skewed line of off-line handwritten characters. The main objective of the work was to segment the lines, words and to segment the character present in hand written document in Gurmukhi Script. We obtain the following table after putting the Handwritten Gurmukhi document for segmentation.

Jain et al. [6] has introduced the word segmentation in OCR system. In this paper, segmentation is formulated in which textual area of image is estimated 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. Mehdi et al. [7] enhanced the efficiency of cursive handwriting based on word segmentation. Also the comparative analysis was taken in extensive research between bitmap and bitmap-data. The algorithm was tested on both type of images and results under different circumstances were compared. Jindal and Lehal [8] have described the historical documents are affected by problems of ageing and repeated use. 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.

Kumar and Jindal [9] have described a segmentation of handwritten document into distinct lines of text. They performed the experiments on various handwritten text images in Gurmukhi Script which are highly skewed, less gap between the lines, more gaps in words etc. Kumar et al. [10] has described a technique of Piece-wise projection along with contour tracing to segment a handwritten document into distinct lines of text. For experiments, we considered only single column document pages. By viewing the results on the computer's display, we calculate line segmentation accuracy manually by checking correctly segmented components.

Kumar and Singh [11] have described an algorithm which is used to segment the scanned document image as a lines, words and characters. Manohar et al. [12] has proposed a novel graph clustering based approach to combine the output of an ensemble of text line segmentation algorithms.

After literature review, it has been concluded that Line and Word Segmentation techniques have problem of accuracy. The accuracy of some methods is not according to the requirement. And also the Mid-Detection algorithm problem is that the segmented points generated are not giving the efficient results. To overcome these problems a new method of Line and Word segmentation from the database is proposed.

After literature review, it has been concluded that Line and Word Segmentation techniques have problem of accuracy. The accuracy of some methods is not according to the requirement. And also the Mid-Detection algorithm problem is that the segmented points generated are not giving the efficient results. To overcome these problems a new method of Line and Word segmentation from the database is proposed.

III. PROBLEMS IN LINE SEGMENTATION

Segmentation of a document image into text line is one of the important challenges in optical character recognition. Line segmentation of a handwritten document makes the process of segmentation more complicated.

Line segmentation of a handwritten or printed document is one of the major challenges in optical character recognition. There are various problems in segmentation of handwritten documents, for example, structural properties of the script, varying writing styles of different persons and uneven spaces between consecutive lines. Text line segmentation is a complex task because of irregularities in geometrical properties such as line height, width, and distance in between line.

The various problem arises in line segmentation are Skewed Text Lines, Overlapping Text Lines, Touching Text Lines, Connected Components.

Skewed Text Lines: Sometimes variations in handwriting of different persons cause the skewness that is slant position of header line. Skew text lines are categorized into three different types- Global Skew, Multiple Skew, and Non-uniform Skew.

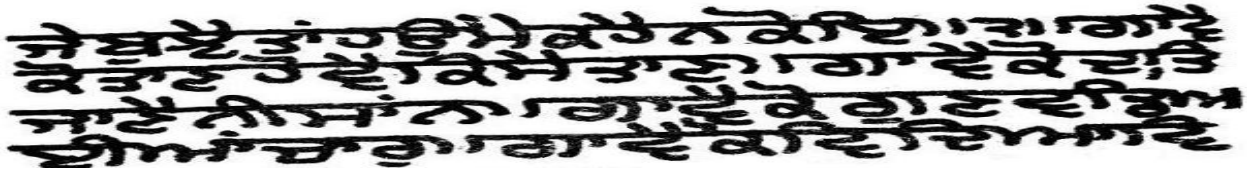

Fig 1.2 Scanned Image of Global Skew [9]

International Journal of Advance Research in Science and Engineering
Vol. No.4, Special Issue (01), November 2015
www.ijarse.com
IJARSE
ISSN 2319 - 8354

Multiple Skew arises when a document containing different orientation of different lines or blocks in different direction as shown in figure.

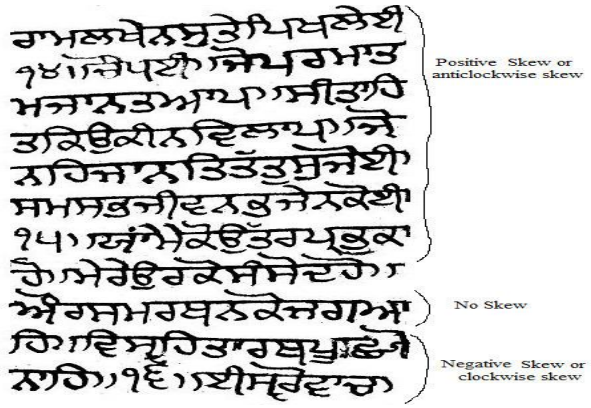

Fig 1.3 Scanned Image of Multiple Skew [9]

Non-uniform Skew present in that case when lines have different slope of header lines of different words containing in same line as shown in figure.

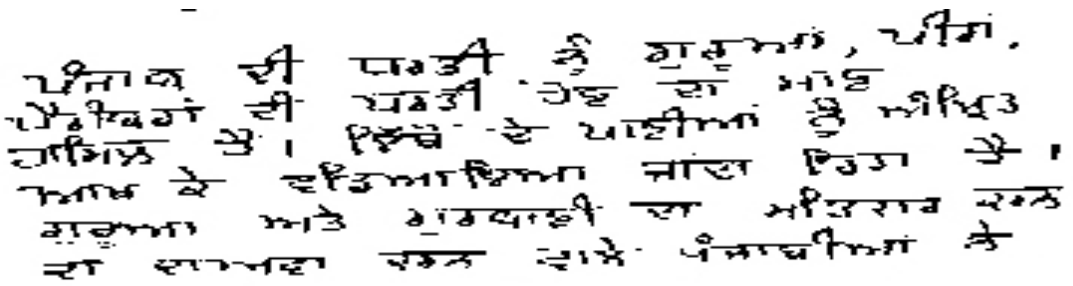

Fig 1.4 scanned image of non-uniform skew [5]

Touching Text Lines: When more than one character of two consecutive lines are touching with each other due to writing style. In this case, characters usually touch the base line and other part of the text line also.

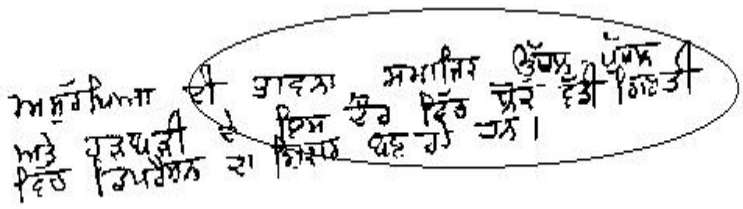

Fig 1.5 scanned images of text lines with touching characters [6]

3.1 Proposed Method

The proposed algorithm segments the lines of a text document written in script.

Algorithm for Line Segmentation

Step 1:-Input the text document written in Gurmukhi script.

Step 2:-Binarize the input and store it into a matrix.

Step 3:-Find the Average Height of the Line in the document.

Step 4:- Divide the document into Vertical strips of size equal to 100 pixels.

Step 5:- Using Horizontal Profile Projection, find the White spaces between the two adjacent lines.

Step 6:- Find the midpoint of the white spaces detected in the step 5.

Step 7:- Calculate the difference between adjacent midpoints.

Step 8:- If the difference is greater than Height of the line then it is assumed that lines have touching components or overlapping with each other.

Step 9:-Find the no. of Lines in between the midpoints.

Step 10:-Extract the midpoints between two consecutive lines found in step 9.

Step 11:-Mark the points obtained in step 10 as segmentation points.

Step 12:-Segment the lines from the extracted segmentation points.

Step 13:-Repeat steps 5 to 12 for each strip obtains in the text document.

Step 14:- Save the matrix into image.

Step 15:- Display the output.

Step 16:- End.

Algorithm for Word Segmentation

Step 1:- Input the Handwritten text Line written in Gurmukhi Script.

Step 2:- Binarize the input and store it into a matrix.

Step 3:- Find White spaces between the Words using Vertical Profile Projection technique.

Step 4:- Find the midpoints of these white spaces and mark these points as the segmentation points.

Step 5:-Segment the Line into Words from the points obtained in the step 4.

Step 6:- Save the matrix into an image.

Step 7:- Display the image as an output.

Step 8:- End.

IV. RESULTS

In this section, the results with the proposed method are discussed. The proposed method is tested on scanned handwritten documents written in script by different writers. Different documents are tested within four main categories as: Simple, Overlapping and Connected Components. A single algorithm is developed for segmenting these types of documents and 94% of overall efficiency has been achieved. Scanned input images are used as input images.

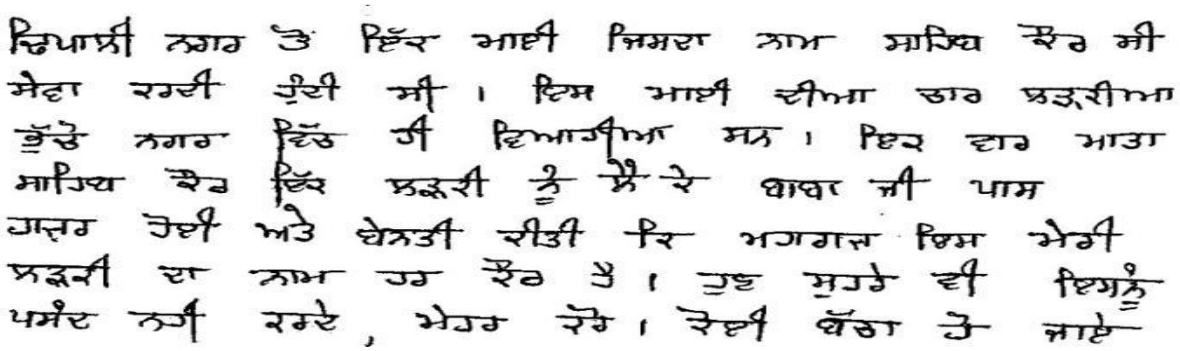

Fig 1.6 Handwritten Scanned Input Image 1[6]

International Journal of Advance Research in Science and Engineering
Vol. No.4, Special Issue (01), November 2015
www.ijarse.com
IJARSE
ISSN 2319 - 8354

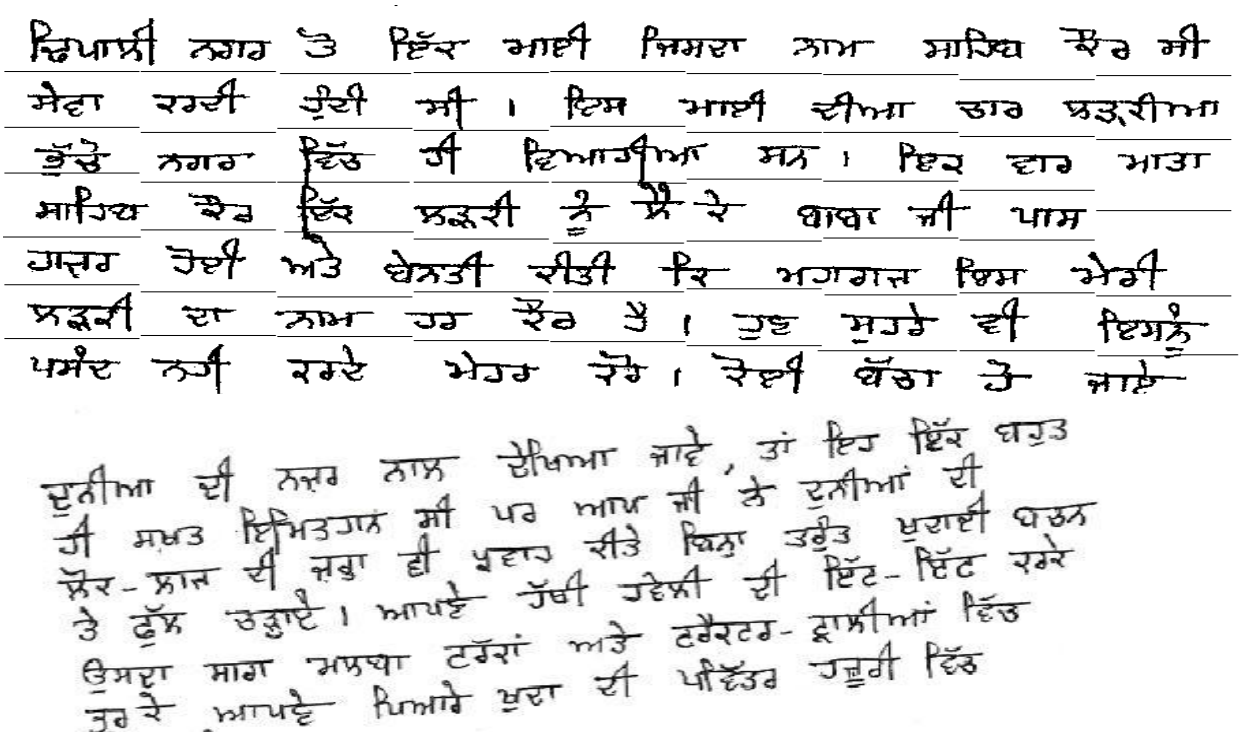

Fig 1.8 handwritten scanned input image 2[6]

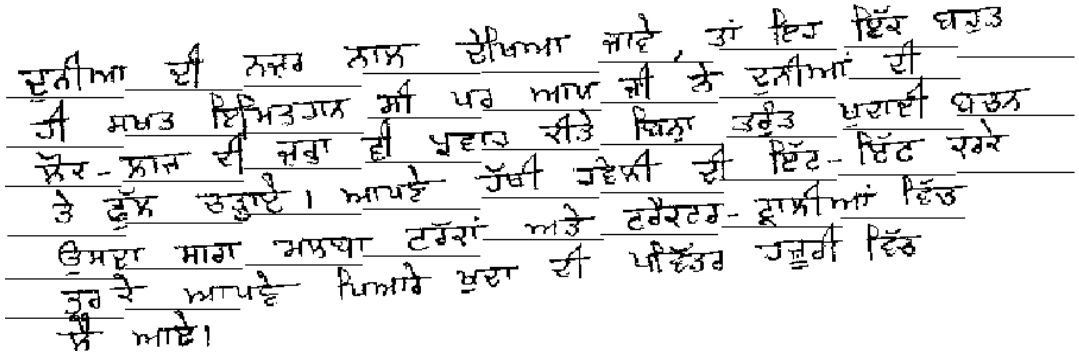

Fig 1.9 output image using proposed method

Word Segmentation Results:-

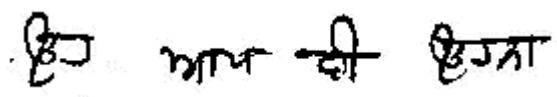

Fig. 1.10 handwritten scanned image 1

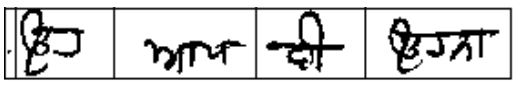

Fig. 1.11 output image using proposed method

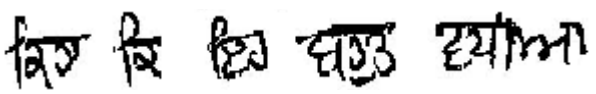

Fig. 1.12 handwritten scanned image 2

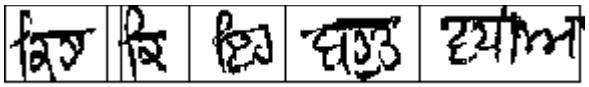

Fig. 1.13 output image using proposed method

Table 4.1 Results of proposed method for Word Segmentation in terms of accuracy

Handwritten Scanned Image	No. of Words	Correctly Segmented	Accuracy
Image 1	4	4	100%

The following table demonstrates the testing of developed system by giving various numbers of input documents written in script:

Table 4.2 Results of Sharma and Sharma method for Line Segmentation[5]

Handwritten Scanned Image	No. of Lines	Correctly Segmented	Accuracy
Image 1	17	15	89%

Table 4.3 Results of proposed method for Line Segmentation in terms of accuracy

Handwritten Scanned Image	No. of Lines	Correctly Segmented	Accuracy
Image 1	7	7	100%

The result of the proposed method is shown in Table 4.3 in terms of accuracy. Some images are Analyzed and listed in this table. Due to space problem only result of some images are presented. But, experiments are performed on 20 different images. Results of proposed method and are shown in Table 4.1 to Table 4.3. From these tables, it is concluded that the proposed method is better than the existing methods [5].

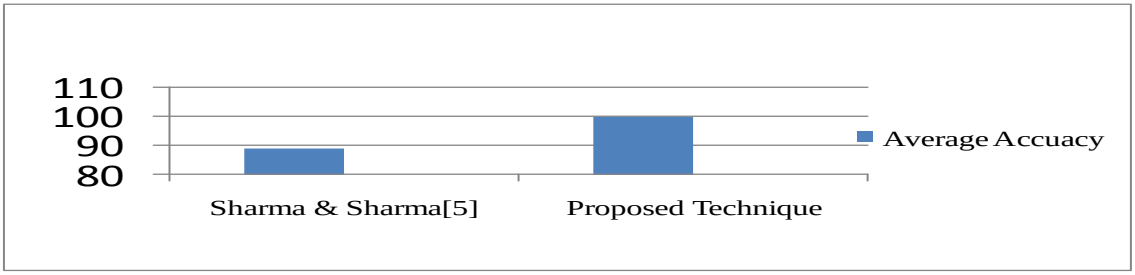

Fig 1.15 Comparison of the proposed method with Existing methods in terms of accuracy

Fig 1.15 shows the comparison of the proposed method with existing methods [5]. The average accuracy of Sharma and Sharma's method [9] in Line Segmentation is 89%%, but results of the proposed method for line segmentation is 100%. From Fig 1.18, it is concluded that the proposed method is better than the existing methods [5].

Table 4.4 Comparison of Proposed method with existing techniques

Sr. No.	Author	Segmentation Type	Doc. Language	Accuracy
1	Sonam Jain	Word	English	99%
2	Mehdi et al.	Word	English	85%
3	Nallapareddy Priyanka	Word	Multiscript	99.5%
4	Nallapareddy Priyanka	Line	Multiscript	99.5%
5	Munish Kumar	Word	Gurmukhi	98.2%
6	Proposed Method	Line & Word	Gurmukhi	100%

Table 4.4 shows the performance of the proposed method is compared with the existing methods in terms of accuracy, where average of each individual category is calculated. From Table 4.3 it is concluded that the proposed method is better than others in term of accuracy in segmentation of Gurmukhi handwritten scripts which Suffers from the problems of connected components, overlapping and Skew Lines & Words.

V. CONCLUSION

In this paper, the proposed method presented a simple line and word segmentation technique which is very different from conventional methods that are being used currently like histogram based approach, projection based approach or thinning approach. The midpoint detection based approach proposed here is simply based on recognition of spaces that separates two lines or two words. 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.

REFERENCES

[1] Karmakar, P., Nayak, B. and Bhoi, N. "Line and Word Segmentation of a Printed Text Document", International Journal of Computer Science and Information Technologies, vol. 5, No. 1, pp.157-160, 2014.

[2] Kaur, N. and Himani. "A Review of Different Skew Detection Techniques", International Journal of Emerging Trends in Engineering and Development, vol.2, No.4, pp. 108-115, 2014.

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

[4] Garg, R. and Kumar, N. "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, 2014.

[5] Sharma, A. and Sharma, A. "Line Segmentation of Gurmukhi Text on Chunk Based Projection Profiles", International Journal of Computer Science And Technology, vol. 4, No.1, pp. 92-94, 2013.

[6] Sneh and Kumar, M. "Segmentation of Connected Components and Overlapping Lines in Handwritten Documents", International Journal of Emerging Trends in Engineering and Development, vol. 4, No.5, pp. 114-118, 2014.

[7] Jain, S. and Singh, H. "A Novel Approach for Word Segmentation in Correlation based OCR System", International Journal of Computer Applications, vol. 99, No.18, pp. 12-20, 2014

[8] Mehdi, M. and Riaz, A. "Optimized Word Segmentation for the Word Based Cursive Handwriting Recognition", Institute of Electrical and Electronics Engineers, pp. 299-304, 2013.

[9] Jindal, S. and Lehal, G. "Line Segmentation of Handwritten Gurmukhi Manuscripts", Proc. of the 3rd International on Advance Computing Conference, Institute of Electrical and Electronics Engineers, , Mumbai, pp. 1797-1801, 2012.

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

[11] Kumar, A., Jindal, S. and Singla, G. "Line Segmentation Using Contour Tracing", Journal of Global Research in Computer Science, vol.3, No.1, pp.50-54,2012.

[12] Kumar, R. and Singh, A. "Algorithm to Detect and Segment Gurmukhi Handwritten Text into Lines, Words and Characters", International Journal of Engineering and Technology, vol.3, No.4, 2011.

[13] Manohar, V., Vitaladevuni, S., Cao, H., Prasad, R. and Natarajan, P. "Graph Clustering-based Ensemble Method for Handwritten Text Line Segmentation", Document Analysis and Recognition, International Conference, Beijing, pp. 574-578, 2011.