International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 4– No.5,October 2012 – www.ijais.org

Morphological Operations and Projection Profiles based Segmentation of Handwritten Kannada Document

Mamatha H R
Department of Information Science and Engineering
P E S Institute of Technology
Bangalore, India

Srikantamurthy K
Department of Computer Science and Engineering
P E S Institute of Technology (South Campus)
Bangalore, India

ABSTRACT
Segmentation is an important task of any Optical Character Recognition (OCR) system. It separates the image text documents into lines, words and characters. The accuracy of OCR system mainly depends on the segmentation algorithm being used. Segmentation of handwritten text of some Indian languages like Kannada, Telugu, Assamese is difficult when compared with Latin based languages because of its structural complexity and increased character set. It contains vowels, consonants and compound characters. Some of the characters may overlap together. Despite several successful works in OCR all over the world, development of OCR tools in Indian languages is still an ongoing process. Character segmentation plays an important role in character recognition because incorrectly segmented characters are unlikely to be recognized correctly. In this paper, a segmentation scheme for segmenting handwritten Kannada scripts into lines, words and characters using morphological operations and projection profiles is proposed. The method was tested on totally unconstrained handwritten Kannada scripts, which pays more challenge and difficulty due to the complexity involved in the script. Usage of the morphology made extracting text lines efficient by an average extraction rate of 94.5% .Because of the varying inter and intra word gaps an average segmentation rate of 82.35% and 73.08% for words and characters respectively is obtained.

General Terms
OCR, Morphological operations, Projection Profiles, Segmentation.

Keywords
Handwritten Kannada Document.

1. INTRODUCTION
OCR refers to a process of generating a character input by optical means, like scanning, for recognition in subsequent stages by which a printed or handwritten text can be converted to a form which a computer can understand and manipulate. A generic character recognition system has different stages like preprocessing, segmentation, feature extraction and classification. Most of the Indian scripts are originated from Brahmi script through various transformations.

An Optical Character Recognition (OCR) system is the process of transforming human readable and optically sensed data to machine understandable codes. The high performance of any recognition system (OCR systems) depends on the detailed analysis of preprocessing and segmentation operations for removing noises and extracting character components respectively from the input document image. [1]

Segmentation is the process of extracting objects of interest from an image. The first step in segmentation is detecting lines. The subsequent steps are detecting the words in each line and the individual characters in each word. This is a crucial step of OCR systems as it extracts meaningful regions for analysis. This step attempts to decompose the image into classifiable units called character.

Segmentation of words into individual letters has been one of the major problems in handwriting recognition. Despite several successful works all over the work, development of such tools in specific languages is still an ongoing process especially in the Indian context. The complexity involved in the segmentation of characters in the uneven spacing between text lines and adjacent characters. The text lines can also be skewed in some cases.

In the recent past, the number of document images available for Indian languages has grown drastically with the establishment of Digital Library of India. The digital library documents originate from a variety of sources, and vary considerably in their structure, script, font, size, quality, etc. Text line extraction from unconstrained handwritten documents is a challenge because the text lines are often Skewed and the space between lines is not obvious. The complexity involved in the segmentation of the Handwritten Documents for Indian languages like Telugu, Tamil and Malayalam is very well explained in [2]. Curved and non-parallel text lines in handwritten documents also make the segmentation and recognition challenging.

Handwriting text line segmentation approaches can be categorized according to the different strategies used. These strategies are projection based, smearing, grouping, Hough-based, graph-based and Cut Text Minimization (CTM) approach[3].The projection-based algorithm proposed in [4] first obtains an initial set of candidate lines from the piece-wise projection profile of the document .The lines traverse around any obstructing handwritten connected component by associating it to the line above or below. The proposed method is robust to handle skewed documents and touching lines. In smearing based approach technique, consecutive black pixels along the horizontal direction are smeared. If the distance between the white space is within a predefined threshold, it is filled with black pixels. The bounding boxes of the connected components in the smeared image are considered as text lines. A new approach for text line

detection by adopting a state-of-the-art image segmentation technique is proposed in [5]. The authors first convert a binary image to gray scale using a Gaussian window, which enhances text line structures. Text lines are extracted by evolving an initial estimate using the level set method.

Grouping approach involves building alignments by aggregating units in a bottom-up approach. Units such as pixels, connected components, or blocks are then joined together to form alignments. An approach based on perceptual grouping of connected components of black pixels is proposed in [6]. Text lines are iteratively constructed by grouping neighboring connected components based on certain perceptual criteria such as similarity, continuity and proximity. According to the authors the proposed technique cannot be used on degraded or poorly structured documents, such as modern authorial manuscripts.

In Hough-based approach the Hough transform is used for locating straight lines in images. In [7] an iterative hypothesis validation strategy based on Hough transform was proposed. The skew orientation of handwritten text lines is acquired by applying the Hough transform to the center of gravity of each connected component in the document image. This technique is able to detect text line in handwritten documents which may contain lines oriented in different directions, erasures and annotations between main lines.

A graph cut based framework using a swap algorithm to segment document images containing complex scripts such as in Indian languages is presented in [2].In [8], the CTM method finds a path or cut line in between the text lines to be separated which minimizes the text line pixels cut by the segmentation line, especially descenders from the upper line and ascenders from the lower line. The bottom up approach of line segmentation from handwritten text is proposed in [9]. In [10], the authors have proposed morphology based handwritten line segmentation using foreground and back ground information. The morphological operation and run-length smearing algorithm (RLSA) is used. A method for line segmentation of handwritten Hindi text is reported in [11]. The method is based on header line detection, base line detection and contour technique. Text line extraction from multi-skewed handwritten documents is found in [12]. They assume that hypothetical water flows, form left and height sides of the image frame, face obstruction from characters of text lines. The stripes of areas left unwetted on the image frame are finally labeled for extraction of text lines.

For word segmentation there exist two distinct tendencies. In the first, after taking as input a text line image, the connected components are calculated. The distances between adjacent connected components are measured using a metric such as the Euclidean distance, the bounding box distance or the convex hull metric. Finally, a threshold is defined which is used to classify the calculated distances as either inter-word or inter-characters gaps [13]. In [14], the word segmentation problem is considered as a text line recognition task, adapted to the characteristics of segmentation. That is, at a certain position of a text line, it has to be decided whether the considered position belongs to a letter of a word, or to a space between two words. For this purpose, three different recognizers based on Hidden Markov Models are designed, and results of writer- dependent as well as writer-independent experiments are reported in the paper.

An approach based on fringe maps to generate segmenting paths between adjacent text lines is proposed in [15]. First they generate a fringe map for the input binary image; next the authors compute peak fringe numbers (PFN) to locate potential regions to find a separating path. PFNs between lines are used to generate a segmenting path to separate adjacent lines. In [16] method for Line Segmentation of Handwritten Hindi text is presented. The method is based on header line detection, base line detection and contour following technique. No preprocessing like skew correction, thinning or noise removal has been done on the data. The authors claim that this method is suitable for fluctuating lines or variable skew lines of text. Also, they confirm that this method is invariant of non uniform skew between words in a line (non uniform text line skew) and the contour following after header line detection correctly separates some of the overlapped lines of text.

An approach to segment the scanned document image is presented in [17]. Here the whole image is considered as one large window. Then this large window is broken into less large windows giving lines, once the lines are identified then each window consisting of a line is used to find a word present in that line and finally to characters. The authors have used the concept of variable sized window, that is, the window whose size can be adjusted according to needs. In [18], the authors have proposed a novel two stage evaluation methodology for word segmentation Techniques. They have proposed a robust evaluation methodology that treats the distance computation and the gap classification stages independently. A survey of methods and strategies in character segmentation is presented in [19].

From the above literature survey it is clear that most of the work has been done for English, Chinese and Arabic etc. Few works are reported on Indian languages like Bangla, Devanagari, Assamese and Telugu scripts. Very few works are reported on text line extraction on Handwritten Kannada Script. To our best knowledge there has been no work in the word and character segmentation of the handwritten Kannada script. Segmentation of handwritten Kannada script into lines, words and character is of great importance and much demanded by some specific applications. Segmentation of handwritten Kannada script poses challenges due to additional modifier characters, writing styles, skewed lines, inter and intra word gaps.

In this paper a methodology based on morphological operations and projection profile for segmentation of the handwritten Kannada script into lines, words and characters is proposed.

The rest of the paper is organized as follows. Section 2 describes the characteristics of Kannada script, section 3 discusses about the proposed methodology, and section 4 briefly discusses the experimental setup and the results obtained are discussed respectively. Finally in Sections 6 and 7, comparative study and conclusions are made.

2. THE CHARACTERISTICS OF KANNADA SCRIPT

In this section, some of the main characteristics of Kannada script is described briefly to point out the main difficulties for segmenting.

Kannada is a popular script and it is the official language of the southern Indian state, Karnataka. Kannada is a Dravidian language mainly used by the people of Karnataka, Andhra Pradesh, Tamil Nadu and Maharashtra. Kannada is spoken by about 44 million people. The language has 47 characters in its

alphabet set (13 vowels and 34 consonants are as shown in Fig.1 and Fig.2).

A Character can be one of the following,

- A stand alone vowel or a consonant
- A consonant modified by a vowel.
- A consonant modified by one or more consonants and a vowel.

Some of the complex characters are listed below to show the complication of the segmentation. The Fig.3 shows the conjunct consonant (Subscript/Vatthu).

ಅ ಆ ಇ ಈ ಉ ಊ ಋ ಎ ಏ ಐ ಒ ಓ ಔ

Fig 1: Vowels of Kannada Script

ಕ ಖ ಗ ಘ ಙ

ಚ ಛ ಜ ಝ ಞ

ಟ ಠ ಡ ಢ ಣ

ತ ಥ ದ ಧ ನ

ಪ ಫ ಬ ಭ ಮ

ಯ ರ ಲ ವ ಶ ಷ ಸ ಹ ಳ

Fig 2: Consonants of Kannada Script

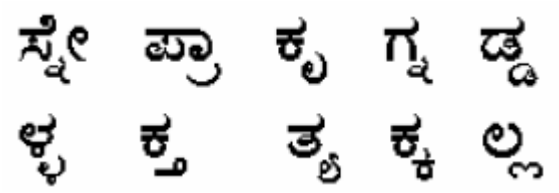

Fig 3: Shows the conjunct consonant (Subscript/Vatthu)

3. PROPOSED METHODOLOGY

In this section segmentation of unconstrained handwritten Kannada script into lines, words and characters is proposed. The proposed method consists of two stages. In the first stage, mathematical morphology technique is used for constructing bridge between the components .In the next stage, projection technique is proposed for the segmentation of the text into line, words and characters.

3.1 Morphology

Mathematical morphology is a tool for extracting image components that are useful in the representation and description of region shape, such as boundaries, skeletons and the convex hull. Dilation is a primitive morphological operation that grows or thickens objects in a binary image .The specific manner and extent of this thickening is controlled by a shape referred to as a structuring element. Structuring elements are small sets or sub images used to probe an image under study for properties of interest.

Mathematically, dilation is defined in terms of set operations. The dilation of A by B denoted A⊕B, is defined as in (1),

$$A \oplus B = \left\{ z/\left(\hat{B}\right)_z \cap A \neq \phi \right\} \quad (1)$$

Where A and B are sets in 2D integer space z2, Φ is the empty set and B is the structuring element and z is the set of all displacements.

Erosion "shrinks" or "thins" objects in a binary image. As in dilation, the manner and extent of shrinking is controlled by a structuring element [20].

Mathematically, erosion of A by B, denoted $A \ominus B$ is defined as in (2),

$$A \ominus B = \{ z/(B)_z \cap A^{|} \neq \phi \} \quad (2)$$

Initially, all the connected components in a document image are detected and removed from the binary image using connected component analysis algorithm. For a component, if the number of on pixels is very small compared to a preset threshold then we remove that component. After this process, the proposed method uses morphology operation that is by using appropriate size of structure element, erosion and dilation will be applied to the binary image. In erosion the last zero value pixel present at the boundary of the image is converted into 1 and in dilation last one value pixel present at the boundary is converted to zero. In this experiment, the unwanted pixels/dots present in the scanned image are removed by applying erosion and the disconnected components are connected using dilation. After dilation, the dilated image is inverted and then the content present in the image is cropped by identifying the rows. The rows are identified by finding the minimum and maximum positions of the zero valued pixels. The line structural element is used for the segmentation of text into lines and rectangular structural element for the segmentation of the lines into words and characters.

3.2 Projection technique

After the completion of first stage, the next stage is to extract individual text lines present in the document. In order to extract individual text line, a technique based on projection is used. A projection profile is a histogram giving the number of ON pixels accumulated along parallel lines. Thus a horizontal projection profile is a one-dimensional array where each element denotes the number of ON pixels along a row in the image. Similarly a vertical projection profile gives the column sums. It is easy to see that one can separate lines by looking for minima in horizontal projection profile of the page and then one can separate words by looking at minima in vertical projection profile of a single line. We have used such projection profile based methods for line, word and character segmentation.

To segment the document image into several text lines, we use the valleys of the horizontal projection computed by a row-wise sum of black pixels. The position between two consecutive horizontal projections where the histogram height is least denotes one boundary line. Using these boundary lines, document image is segmented into several text lines. Similarly, to segment each text line into several text words, we use the valleys of the vertical projection of each text line obtained by computing the column-wise sum of black pixels. The position between two consecutive vertical projections where the histogram height is least denotes one boundary line. Using these boundary lines, every text line is segmented into several text words.

4. EXPERIMENTAL RESULTS

This section presents the results of the experiments conducted to study the performance of the proposed method. The method has been implemented in MATLAB 7.8 on Dual Core 3 GHz with 2 GB RAM. For experimental purpose, we have considered 100 handwritten document pages collected from different individuals of various professions like school children, undergraduate and postgraduate students, house wives, office employees etc., from different cities and villages. The data set contains varieties of writing styles. Non-text elements are not included in the documents and almost all the documents have two or more adjacent text lines touching in several areas. Some of the documents have variable skew angles among text lines with different skew directions. single column document pages is considered for the experimentation. The number of lines in each document varies from 03 to 28 lines. For each document image, the corresponding ground truth information like the number of lines, words and characters is created manually. The total number of text lines was 714 while the corresponding number of words and characters was 3921 and 13734 respectively. Segmentation accuracy of 100 text documents in this work is measured by the fraction percentage of number of lines/words/characters correctly segmented to the total number of lines/words/characters present in the document. Our proposed methodology gave an average segmentation rate of 94.5%, 82.35% and 73.08% for lines, words and characters respectively.

Some of the problems that were posed during the line segmentation were due to the fact that the consonant conjuncts which appear below the base consonant which results in a false white space in the horizontal projection. Also overlapping of the consonant conjuncts of one line with the vowel modifiers which appear towards the top of the next line can mask some of the minima that should have been seen in horizontal projection. Most of the errors encountered in the word and character segmentation phase are due to the non-uniform spacing between characters of the same word and between adjacent words. Different stages from input handwritten Kannada document image to the segmentation at respective levels are shown from Fig 4 to Fig 15.

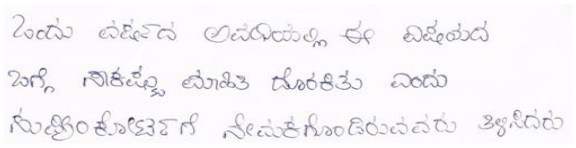

Fig 4: Input image

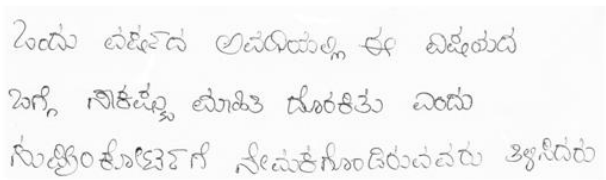

Fig 5: Gray scaled image

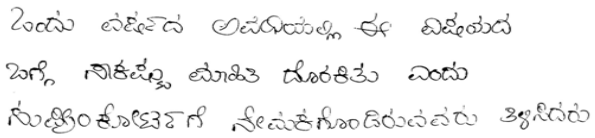

Fig 6: Binarised Image

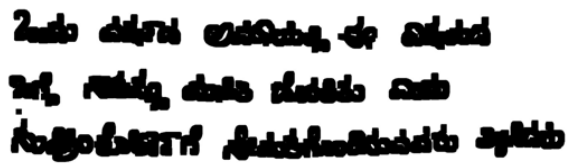

Fig 7: Image after Erosion

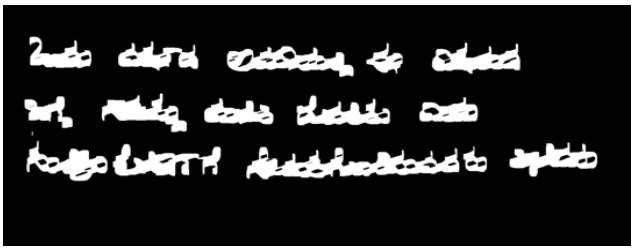

Fig 8:Image after Dilation

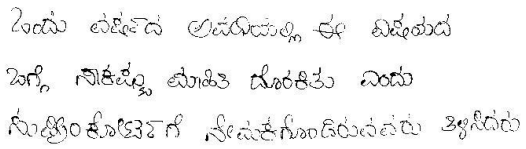

Fig 9: Cropped image

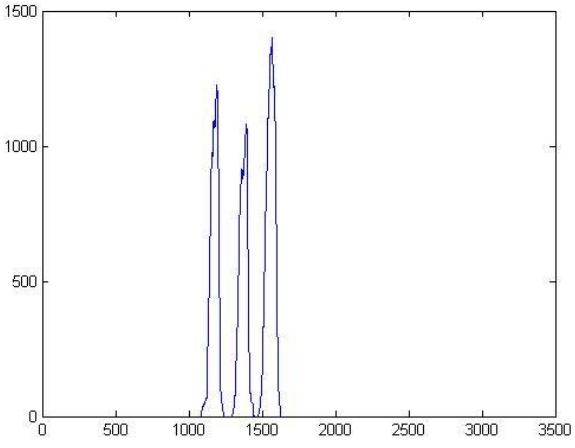

Fig 10: Horizontal projection of the cropped image

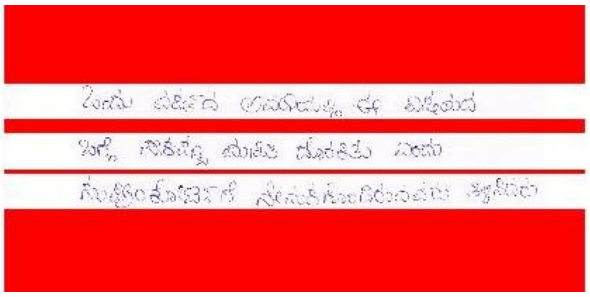

Fig 11: Result of Line segmentation

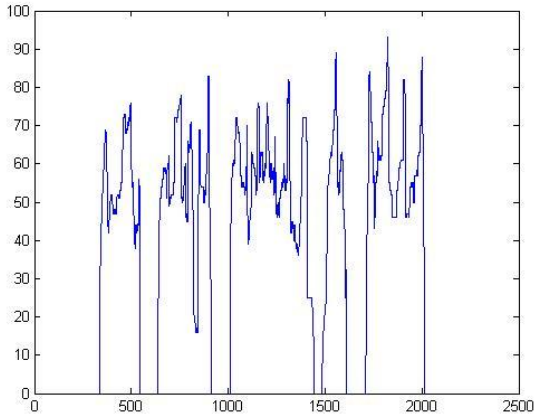

Fig 12: Vertical projection on the segmented line for word segmentation

(a)

(b)

(c)

Fig 13: Result after word segmentation (a, b, c)

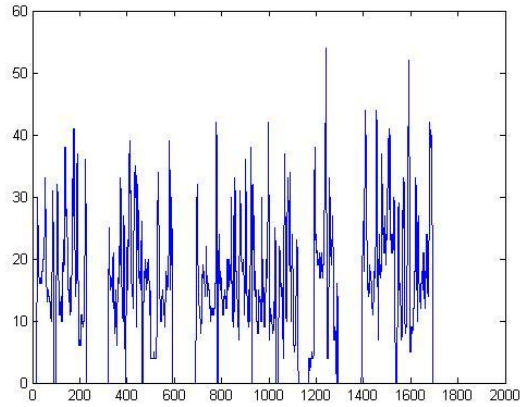

Fig 14. Vertical projection on the segmented line for Character segmentation

(a)

(b)

(c)

(d)

Fig 15: Result after Character segmentation (a, b, c, d)

5. COMPARATIVE STUDY

The Table 1 shows the comparison of existing methods with proposed method. To compare our proposed method with the existing work is very difficult as very few works exist in the line segmentation of handwritten Kannada script which is experimented on different datasets of complexity. To the best of our knowledge there is no work found in the word and character segmentation of the handwritten Kannada script. In [21] the authors have experimented with the methods specified in [22] and [23].

Table 1 . Comparison of proposed method with the existing methods for line segmentation

Author	Segmentation Method	Size of dataset	Segmentation rate
Alireza Alaei et al.,[21]	Potential Piece-wise Separation Line technique	204	94.98%
Alireza Alaei et al.,[21]	Stripe based approach	204	95.32%
Manjunath Aradhya et al.,[24]	Component extension technique	250	Not specified
Proposed	Morphological operations and projection profile based approach	100	94.5%

6. CONCLUSION

In this paper, a segmentation scheme for handwritten Kannada scripts is proposed. The proposed method consists of two stages. In the first stage, mathematical morphology technique is used for removing disconnected components and constructing bridge between the components. In the next stage the projection profile technique is used for segmentation of text into lines, words and characters. The method was tested on totally unconstrained handwritten Kannada scripts, which pays more challenge and difficulty due to the complexity involved in the script. Usage of the morphology made extracting text lines efficiently by an average extraction rate of 94.5% .Because of the varying inter and intra word gaps we could get an average segmentation rate of 82.35% and 73.08% for words and characters respectively.

7. ACKNOWLEDGMENTS

Authors would like to thank Ms Manasa Bhat and Ms Deepa R of the Department of Information Science and Engineering, P E S Institute of Technology, Bangalore, India who has helped us for collecting the documents and preparing the ground truths. The authors would also like to thank all writers who contributed for this dataset.

8. REFERENCES

[1] K. Srikanta Murthy, G. Hemantha Kumar, P. Shivakumar and P.R. Ranganath. 2004. Nearest Neighbour Clustering approach for line and character segmentation in epigraphical scripts. In the proceedings of International Conference on Cognitive Systems (ICCS-2004), New Delhi, December 14-15, 2004.

[2] K.S. Sesh Kumar, A.M. Namboodiri, and C.V. Jawahar.2006. Learning Segmentation of Documents with Complex Scripts. In the proceedings of ICVGIP 2006, LNCS 4338, pp. 749–760, 2006

[3] Zaidi Razak, Khansa Zulkiflee , Mohd Yamani Idna Idris, Emran Mohd Tamil, Mohd Noorzaily ,Mohamed Noor, Rosli Salleh, Mohd Yaakob ,Zulkifli Mohd Yusof and Mashkuri Yaacob,"Off-line Handwriting Text Line Segmentation : A Review" ,IJCSNS International Journal of Computer Science and Network Security, vol.8 No.7, July 2008 pp 12-20

[4] M. Arivazhagan, H. Srinivasan, S. N. Srihari.2007. A Statistical Approach to Handwritten Line Segmentation. In Proceedings of SPIE Document Recognition and Retrieval XIV , San Jose, CA, February 2007

[5] Y. Li, Y. Zheng, D. Doermann, and S. Jaeger.2006. A new algorithm for detecting text line in handwritten documents. In International Workshop on Frontiers in Handwriting Recognition, 2006, pp. 35–40

[6] L. Likforman-Sulem and C. Faure.1994.Extracting text lines in handwritten documents by perceptual grouping. Advances in handwriting and drawing : a multidisciplinary approach,C. Faure, P. Keuss, G. Lorette and A. Winter Eds, Europia,Paris, 1994, pp. 117-135

[7] L. Likforman-Sulem, A. Hanimyan and C. Faure.1995. A Hough based algorithm for extracting text lines in handwritten documents.in the proceedings of Third International Conference on Document Analysis and Recognition, Vol. 2, August 1995, pp. 774-777.

[8] C. Weliwitage, A. L. Harvey and A. B. Jennings.2005.Handwritten Document Offline Text Line Segmentation. In Proceedings of Digital Imaging Computing: Techniques and Applications, 2005, pp. 184-187

[9] D.Sarkar and R.Ghose. A bottom up approach of line segmentation from handwritten text.

[10] U.pal ,P.P.Roy and J.Liados. 2010.Morphology based handwritten line segmentation using foreground and backgroud information. In the proceeding of Intl conference on Frontiers in Handwriting Recognation,2010

[11] L. Kaur N.K. Garg and M.K. Jindal. 2010.A new method for line segmentation of handwritten Hindi text. In the proceeding of 7th Intl conference on Information technology, pages 392–397, 2010.

[12] M. kundu M. Nasipuri S. Basu, C. Chaudhuri and D.K. Basu.2007. "Text line extraction from multi-skewed handwritten documents". Patten Recognition, 40:1825–1839, 2007.

[13] G. Louloudis , B. Gatos , I. Pratikakis and C. Halatsis , 2009."Line And Word Segmentation of Handwritten Documents " ,Journal Pattern Recognition archive Volume 42 Issue 12, December, 2009, pp 3169-3183

[14] F. Luthy, T. Varga, H. Bunke.2007. "Using Hidden Markov Models as a Tool for Handwritten Text Line Segmentation", Ninth International Conference on Document Analysis and Recognition, Curitiba, Brazil, 2007, pp. 8-12

[15] Vijaya Kumar Koppula , Atul Negi .2010.Using Fringe Maps for Text Line Segmentation in Printed or Handwritten Document Images.In the proceedings of 2010 Second Vaagdevi International Conference on Information Technology for Real World Problems,2010,pp 83-88

[16] Naresh Kumar Garg, Lakhwinder Kaur and M. K. Jindal.2010.A New Method for Line Segmentation of Handwritten Hindi Text.In the proceedings 2010 Seventh

International Conference on Information Technology,2010, pp 392-397

[17] Rajiv Kumar and Amardeep Singh, 2010.Detection and Segmentation of Lines and Words in Gurmukhi Handwritten Text. In the proceedings of IEEE 2nd International Advance Computing Conference,2010,pp 353-356

[18] G. Louloudis , N. Stamatopoulos , B. Gatos .2009.A Novel Two Stage Evaluation Methodology for Word Segmentation Techniques. In the proceedings of 10th International Conference on Document Analysis and Recognition,2009, pp 686-690

[19] Richard G. Casey and Eric Lecolinet.1996."A survey of Methods and Strategies in Character Segmentation", IEEE Transactions On Pattern Analysis And Machine Intelligence, vol. 18, no. 7, July 1996 pp 690-706

[20] Rafael C.Gonzalez ,Richard E.Woods and Steven L.Eddins ,Digital Image Processing using MATLAB, Indian Edition,2009,pp 348-361.

[21] Alireza Alaei, P. Nagabhushan and Umapada Pal.2011.A Benchmark Kannada Handwritten Document Dataset and its Segmentation. In the proceedings of International Conference on Document Analysis and Recognition, 2011,pp 141-145.

[22] B. Gatos, N. Stamatopoulos and G. Louloudis.2009."ICDAR 2009 Handwriting Segmentation Contest," Proc. of 10th ICDAR, 2009, pp. 1393–1397.

[23] A. Alaei, U. Pal and P. Nagabhushan.2011."A new scheme for unconstrained handwritten text-line segmentation", Pattern Recognition, 44 (4), 2011, pp. 917–928.

[24] V. N. Manjunath Aradhya and C. Naveena.2011.Text Line Segmentation of Unconstrained Handwritten Kannada Script.In the proceedings of ICCCS'11, 2011, pp 231-234.