IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-3, Issue-5 E-ISSN: 2347-2693

Text Line Extraction of Handwritten Kannada Documents Based on Bounding Box Technique

Chethana H T[1*] and Mamatha H R[2]

[1*,2]*Department of Information Science & Engineering ,*
P.E.S Institute of Technology, VTU,Bangalore,560085,India

www.ijcseonline.org

Received: May/02/2015 Revised: May/10//2015 Accepted: May/24/2015 Published: May/30/ 2015

Abstract— Optical Character Recognition is the process of transforming printed or handwritten text in to a form in which computer can understand and manipulate. An important task of any Optical Character Recognition(OCR)system is segmentation. Characters, words and lines are separated from image text documents by segmentation. Depending on the segmentation algorithm which is being used can affect the accuracy of OCR system. Segmentation of handwritten Kannada script poses challenges due to writing styles, skewed lines, overlapping lines, inter and intra word gaps. In this paper we have proposed method for segmentation of handwritten Kannada documents based on bounding box and morphological operations, an average segmentation rate of 92% for lines is obtained.

Keywords— Segmentation; Handwriting; Text lines; OCR; Bounding Box

I. INTRODUCTION

A document has a structure which provides additional information. Without document structure it is very difficult to index and retrieve correctly the information contained in a document. So document structure analysis is a crucial stage in any indexing and retrieval system such as optical character recognition. A method of dividing the document regions into text and non-text regions is known as document segmentation.While analysing the documents, segmentation plays a very important rôle. If the processing of documents like government files, books, newspapers, scientific journals etc is done automatically, then it reduces the time, money and effort. Extracting the region, identifying the type of region and processing of each region separately are the stages in the automation of document analysis[1].For example, OCR system process the text regions. It converts text region into machine understandable form and non-text regions are compressed and stored.

II. BACKGROUND

The process of extracting objects of interest from an OCR can be done in two ways – online, offline. Real time recognition of characters occurs in online character recognition, and handwriting is usually captured and stored in digital form by different means[12]. In offline character recognition, the handwritten text is typically scanned using high resolution scanner and this scanned document is made available to the recognition algorithm in the form of a binarised image. As there is no control over the medium and instrument used ,offline character recognition poses more challenging and difficult task than online character recognition.

For document structure extraction, text line extraction can be seen as a pre-processing step. By using modern editing tools, various techniques have been developed for segmentation of printed documents [5]. Printed and handwritten documents strongly differ from each other because layout formatting requirements are not well structured and thus their physical structure is harder to extract. Overlapping and touching components along with the narrow spaced lines are also included in handwritten pages. Depending on the writer movement, characters and words will have varying shapes. Full text recognition is not yet available, except for printed documents.

A. Characteristics and representation of text lines

Lines and blocks are immediately visible, when we look in to the physical structure of a document image from a certain distance. Columns, annotations in margins, stanzas, etc are present in these blocks. Blocks do not have rectangular shape in historical documents. So, structure of text line becomes the pre dominant physical structure.

In section II B. we will discuss some important definitions about components of text line and text line segmentation[5]. In section II C. and II D. we will describe the factors affecting the text line structure of the document and finally we will discuss how a text line can be represented.

B. Text line components

Baseline : In a text line all the lower part of the character bodies are connected by an imaginary line known as baseline as shown in Fig. 1.

Median line : In a text line all the upper part of the character bodies are connected by an imaginary line known as median line.

Upper line : In a text line all the top of ascenders are connected by an imaginary line known as upper line.

Lower line : In a text line all the bottom of descenders are connected by an imaginary line known as lower line.

Overlapping components: These components are present in the region of adjacent line which are descenders and ascenders.

Touching components : These components are present in the region of consecutive lines which are connected[5] as shown in Fig. 2.

There are two approaches to text line segmentation : one is searching for (fictitious) separating lines or paths, another one is searching for aligned physical units[5]. Depending on the text line structure complexity, segmentation technique is chosen.

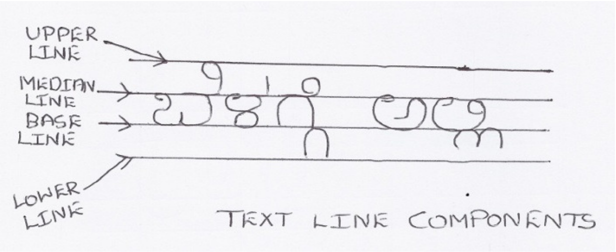


Fig. 1 Text Line Components

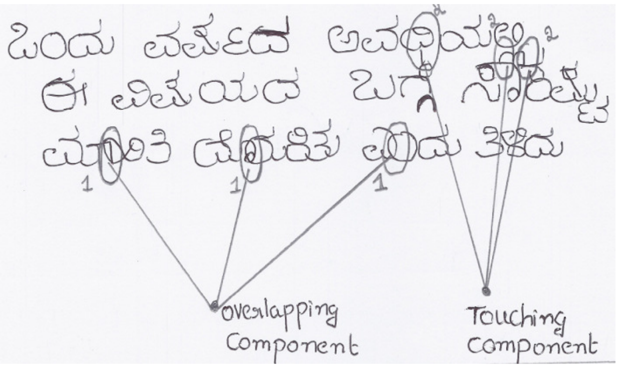


Fig. 2 Overlapping and touching components

C. Influence of author style

Baseline fluctuation : The baseline may be straight or curved. Depending on the writer movement, baselines will vary .

Line orientations : Lines will be oriented in different directions and at different angles.

Line spacing : Lines which are widely spaced are easy to find. When lower baseline of the first line touches with the upper baseline of the second line, text line extraction becomes more difficult.

D. Influence of poor image quality

Imperfect pre-processing : Smudges and seeping ink present in other side of the image in the document produce binarisation errors as shown in Fig.3.Smudges means each pixel which is present in the source image spreads on the periphery of the surrounding pixels as shown in Fig.4.

Stroke fragmentation and merging : Due to the presence of punctuation, dots and broken strokes makes the quality of the images to be low. Segmentation into the correct text line[5] becomes difficult when the components are broken as they are no longer linked to the median baseline of the handwriting.

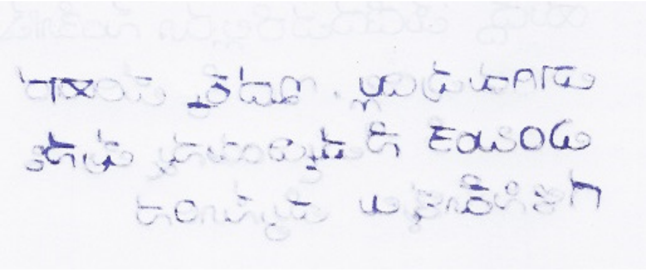


Fig. 3 Presence of seeping ink from other side of the document

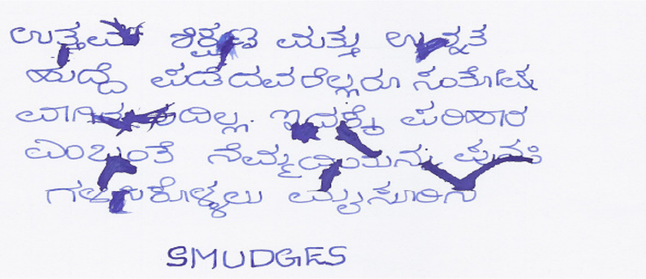


Fig. 4 Smudges

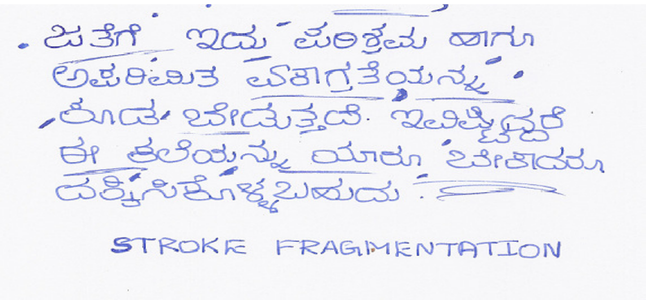


Fig. 5 Stroke Fragmentation

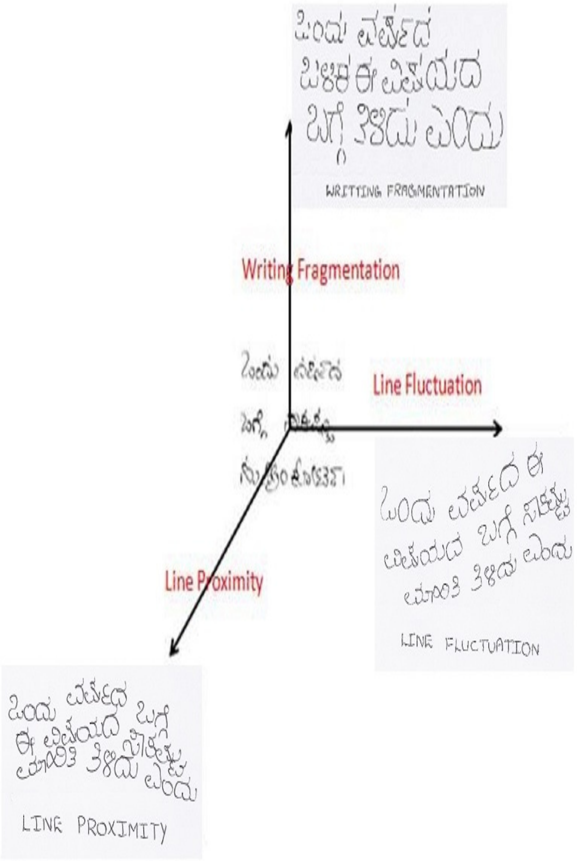


Fig. 6 The three main axis of document complexity for text line segmentation.

D. Zones in Kannada Word

A Kannada word can be divided in to different horizontal zones. Case (i) A word without subscript as shown in Fig. 7 (pronounced as ramanu).Case (ii) A word with subscript as shown in Fig. 8(pronounced as prashastavaagi).

Fig.7 shows the sample word which does not have subscript character, a word can be divided in to two zones - top and middle zones. Top line is an imaginary horizontal line that passes through the topmost pixel of the word. Baseline is an imaginary horizontal line that propagates through the base pixel of the word. In the profile, an horizontal line passing through the top most peak is head line. The portion between top and head line is top zone and the portion between the head line and the base line is middle zone[6].

Consider the sample word in Fig. 8 which have subscript character, a word is divided in to three zones - top, middle and bottom zones. A word without subscripts have top and middle zones which are similar to that of the word with subscripts. The portion between the baseline and the bottom line is bottom zone. An horizontal line passing through the bottommost pixel of the word is bottom line[6].

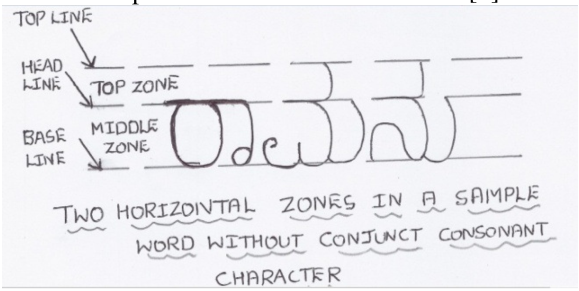

TOP LINE
HEAD LINE
TOP ZONE
BASE LINE
MIDDLE ZONE
TWO HORIZONTAL ZONES IN A SAMPLE WORD WITHOUT CONJUNCT CONSONANT CHARACTER

Fig. 7 A word without subscript

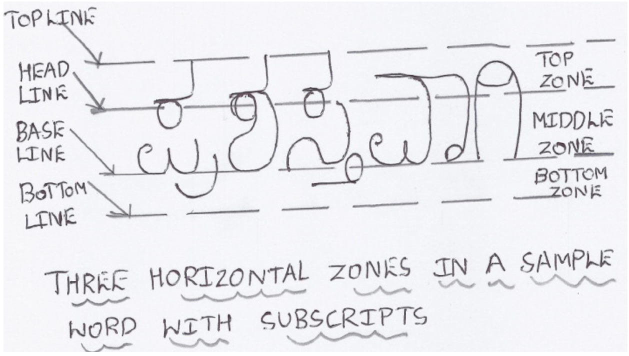


Fig. 8 A word with subscript

III. RELATED WORK

Many handwriting text line segmentation approaches have been developed for languages like English, Arabic, Telugu, Tamil, Chinese etc and they are mainly classified in to different strategies like Run Length Smearing, Grouping, Hough-based, Cut Text Minimization (CTM), Repulsive attractive network, Fringe Map, Median Segmentation, Connected component and Strip based Projection Profile approach respectively.

The Run Length Smearing method is proposed in[3] ,here consecutive black pixels along the horizontal direction are smeared and bounding boxes of the smeared image encloses the text lines. Grouping approach proposed in [6] involves black pixels or connected components are joined together to form alignments. This approach is not suitable for poorly degraded documents especially for manuscripts.

Hough based approach proposed in [10] involves locating straight lines in images using hough transform. This technique is able to detect text lines which are oriented in different directions and at different angles in handwritten documents. Cut Text Minimization (CTM) approach proposed in [4] finds a separating path between the text lines, thereby cutting segmentation line by minimizing the text line pixels.

An approach based on fringe map for line segmentation proposed in [7] generates segmenting paths between adjacent text lines. For a given input binary image they generate a fringe map. In order to locate potential regions, peak fringe numbers (PFN) are computed to find the paths which are separated.

Oztop et al have proposed repulsive attractive network method in Arabic which [5] works directly on gray level images and baseline units are constructed iteratively . In the image from top to bottom , baselines are constructed one by one. In the baseline, each pixels of the image acts as attractive forces and baselines which are extracted already acts as repulsive forces. Median segmentation proposed in [3] the white space between text lines is checked for binarised image. Detection of white space between text lines causes image to be segmented in to sets of paragraphs. Arivazhagan et al have proposed strip based projection profile[11] for line segmentation in Arabic, English where image is stored in the 2D-array and it is assumed height and width of an image can be calculated easily.

From the above literature survey, it is found that most of the work has been done for Chinese, Arabic, English etc. Few works are reported on Indian languages like Bangla, Devanagiri, Assamese and Telugu scripts[8].Very few works are reported on Handwritten Kannada Script and research is at an infant stage. This motivated us to work on text line segmentation of handwritten Kannada documents.

In this paper a methodology based on morphological operations and bounding box method for segmentation of the handwritten Kannada documents into lines is proposed. The rest of the paper is organised as follows. Section III A. describes the characteristics of Kannada script, section IV discusses about the proposed methodology, and section V briefly discusses the experimental setup and the results obtained are discussed respectively. Finally in sections VI and VII, comparative study and conclusions are made.

A. The Characteristics of Kannada Script

In this section, to point out the difficulties of segmentation we will describe some of the main characteristics of Kannada script briefly.

Kannada is a popular script and it is the official language of the Karnataka. Kannada is a Dravidian language [4] mainly used by the people of Andhra Pradesh, Maharashtra, Tamil Nadu and Karnataka. 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. 9, Fig. 10 and Fig. 11).

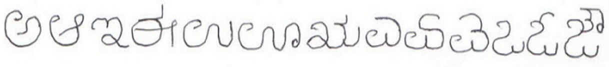


Fig. 9 Vowels of Kannada Script

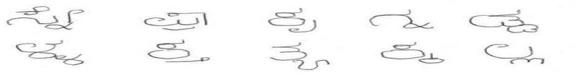


Fig. 10 Shows the conjunct consonant(Vatthu/Subscript)

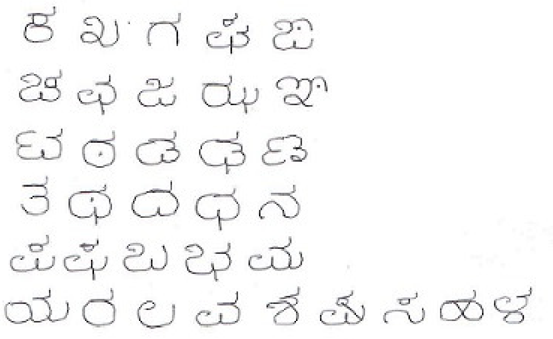


Fig. 11 Consonants of Kannada Script

IV. PROPOSED METHODOLOGY

In this section segmentation of handwritten Kannada documents in to lines is proposed. The proposed methodology comprises of two stages.

- Pre-processing
- Text line extraction.

A. Preprocessing

The pre-processing is a series of operations performed on the scanned input image. Binarisation process converts a gray scale image into a binary image using global thresholding technique [2].

B. Proposed method1 : Bounding Box method

After the completion of pre processing stage, next stage is to extract individual lines present in the given input image. We have used bounding box method for text line detection and extraction respectively.

```
Algorithm
Begin
Input: Handwritten Kannada text document
Output: Segmented lines
Step1: Binarise the original image.
Step2: Construct the horizontal histogram for the image.
Step3:Using the histogram, find the rows containing lesser number of white pixels.
Step4:Find the centroids of the above rows using length and width as parameters .
Step5:Measurement of centroids is calculated by varying the threshold values.
Step6:Using these measurement, mark the bounding box for text lines.
```

Step7:Copy the pixels in bounding box and save in to separate file.

C. *Proposed method2 :Morphololgical operations and Bounding Box method*

Algorithm
Begin
Input: Handwritten Kannada text document
Output: Segmented lines
Step1: Binarise the original image.
Step2: Morphological operations are used for constructing the bridge between the components. Dilation and erosion are two primitive morphological operations that can be applied to the binarised image[4].
Step3: Construct the horizontal histogram for the image.
Step4:Using the histogram, find the rows containing lesser number of white pixels.
Step5:Find the centroids of the above rows using length and width as parameters.
Step6:Measurement of centroids is calculated by varying the threshold values.
Step7:Using these measurement, mark the bounding box for text lines.
Step8:Copy the pixels in bounding box and save in to separate file.

V. EXPERIMENTAL RESULTS

This section presents the results of the experiments conducted to study the performance of the proposed method based on dataset collected by author of [4]. For experimental purpose, we have considered 200 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. Author has collected documents in such a way that documents which contains several adjacent text lines touching in several areas. Some of the documents have variable skew angles among text lines with different skew directions. The number of lines in each document varies from 02 to 20 lines. Segmentation accuracy of 200 text documents in this work is measured by the fraction percentage of number of lines correctly segmented to the total number of lines present in the document. An average segmentation rate of 92% using bounding box method is obtained.

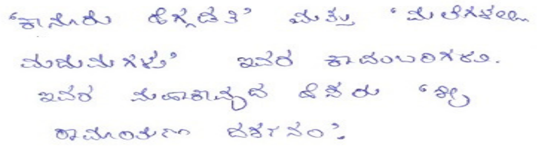
VI . COMPARATIVE STUDY

Table 1 shows the comparison of our proposed method for two different datasets for line segmentation. In order to analyze our proposed method on the standard dataset, we have collected the Kannada Handwritten Text Document (KHTD) Dataset from the author of [9].Four different text categories like four different text categories like movie, medical texts, sports news, stories and general news of Kannada were considered. The dataset is collected by author from different individuals belonging to different categories like age, educational background in a separate unruled A4 sheet without any restrictions. The participants are given to write text pages by different types of pens. Using a flat bed scanner with the resolution of 300 dpi, the documents which are collected from different individuals are then scanned in gray-scales. We have considered 100 documents from this dataset taking 25 documents from each category for experimentation.

Table 1. Comparison of the results of the proposed method with the existing methods for line segmentation for two different datasets.

Author	**Segmentation Method**	**Size of Dataset**	**Segmentation rate**
Alaei et al.,[13]	Potential Piece-wise Separation Line technique	204	94.98%
Alaei et al.,[13]	Stripe based approach	204	95.32%
Aradhya et al.,[14]	Component extension technique	250	Not specified
Proposed method 1 on our dataset	Bounding Box	100	91.2%
Proposed method 1 on Nagabhushan P et.al[9]	Bounding Box	100	89.6%
Proposed method 2 on our dataset	Morphological operations and bounding box	100	92%
Proposed method 2 on Nagabhushan P et.al[9]	Morphological operations and bounding box	100	90.4%

Different stages from input handwritten Kannada document image to the segmentation using bounding box method at respective levels are shown from Fig.12 to Fig. 14 .
Different stages from input handwritten Kannada document image to the segmentation at respective levels after applying morphological operations are shown from Fig. 15 to Fig. 19.



'ಕಾನೂರು ಹೆಗ್ಗಡಿತಿ' ಮತ್ತು 'ಮಲೆಗಳಲ್ಲಿ
ಮದುಮಗಳು' ಇವರ ಕಾದಂಬರಿಗಳು.
ಇವರ ಮಹಾಕಾವ್ಯದ ಹೆಸರು 'ಶ್ರೀ
ರಾಮಾಯಣ ದರ್ಶನಂ'.

Fig. 12 Input image

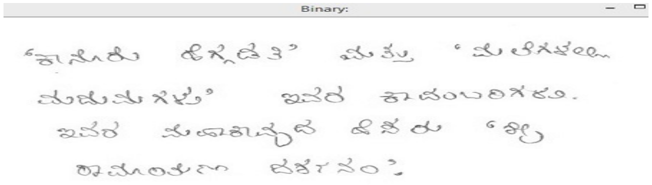


Fig. 13 Binarised image

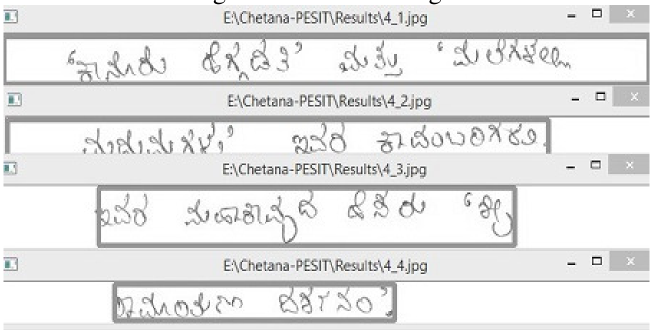

E:\Chetana-PESIT\Results\4_1.jpg
E:\Chetana-PESIT\Results\4_2.jpg
E:\Chetana-PESIT\Results\4_3.jpg
E:\Chetana-PESIT\Results\4_4.jpg

Fig. 14 Result of line segmentation after applying bounding box

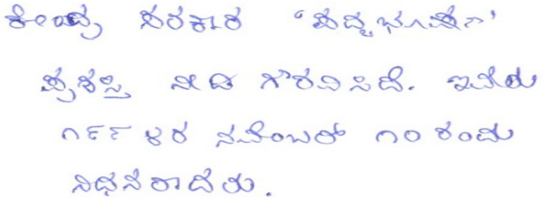


Fig. 15 Input image

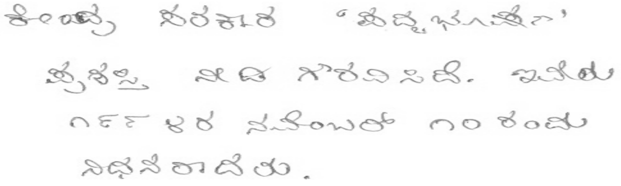


Fig. 16 Binarised image

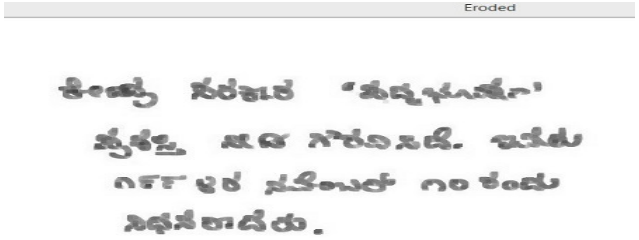


Fig. 17 Result of input image after applying erosion

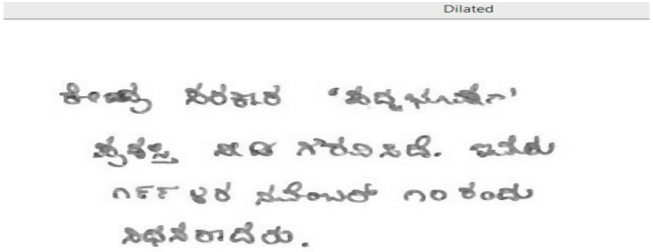


Fig. 18 Result of input image after applying dilation

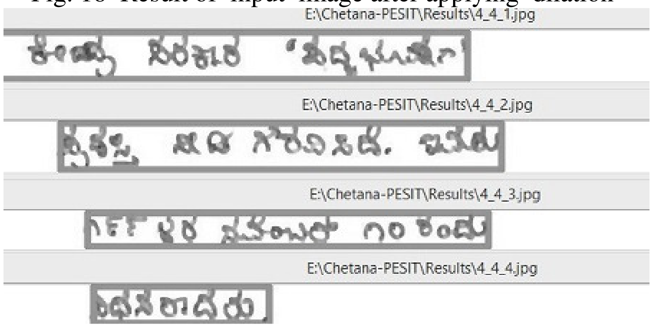

E:\Chetana-PESIT\Results\4_4_1.jpg
E:\Chetana-PESIT\Results\4_4_2.jpg
E:\Chetana-PESIT\Results\4_4_3.jpg
E:\Chetana-PESIT\Results\4_4_4.jpg

Fig. 19 Result of line segmentation after applying bounding box with morphological operations

VII. CONCLUSION

In this paper, we have proposed a bounding box method for segmentation of Kannada handwritten documents in to lines. The method was tested on totally unconstrained handwritten Kannada documents. An average segmentation rate of 91.2% and 89.6% for two different datasets is obtained respectively. Usage of morphology made extracting the lines efficiently by an average segmentation rate of 92% and 90.4% respectively. Both the approaches work for simple handwritten text with less skewness in the text lines. These methods cannot be applied for text with more skew in the lines. An average segmentation rate of 0.8% is improved for both the datasets.

ACKNOWLEDGMENTS

We would like to thank Nagabhushan P and Alireza Alaei authors of [9] for providing us with thier dataset for our experimentation.

REFERENCES

[1] Priyadharshini N and Vijaya MS , "Genetic Programming for Document Segmentation and Region Classification using Discipulus Perceptron", (IJARAI) International Journal of Advanced Research in Artificial Intelligence ,Vol.2 ,No.2, **2013**

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

[3] Pulagam Soujanya, Vijaya Kumar Koppula , Kishore Gaddam and P. Sruthi , "Comparative Study of Text Line Segmentation Algorithms on Low Quality Documents", Special Issue of International Journal of Computer Science & Informatics *(IJCSI)* ,ISSN (PRINT): 22315292 , Vol .II , Issue1 , 2.

[4] Mamatha HR and Srikantamurthy K , "Morphological Operations and Projection Profiles based Segmentation of Handwritten Kannada Document", International Journal of Applied Information Systems (IJAIS)– ISSN:2249-0868 Foundation of Computer Science FCS,**2012.**

[5] Laurence Likforman- Sulem , Abderrazak Zahour and Bruno Taconet,"Text line segmentation of historical documents: a survey", IJDAR9:123–138 DOI 10.1007/s10032-006-0023-z . M , **2007**.

[6] Munish Kumar, R.K. Sharma and M.K. Jindal , "Segmentation of Lines and Words in Handwritten Gurumukhi Script Documents", Indian Institute of Information Technology Allahabad, India.

[7] Vijaya Kumar Koppula and Atul Negi , "Using Fringe Maps for Text Line Segmentation in Printed or Handwritten Document Images", **2010** ,pp.**8388**.

[8] Mamatha H R and Srikantamurthy K ,"Skew Detection, Correction and Segmentation of Handwritten Kannada Document", International Journal of Advanced Science and Technology ,Vol. 48, November,**2012**.

[9] Nagabhushan P, Alireza Alaei and Umapada pal, "A Benchmark Kannada Handwritten Document Dataset and its Segmentation", International Conference on Document Analysis and Recognition,**2011.**

[10] Laurence Likforman- Sulem and Ana hid Hanimyan , "A Hough Based Algorithm for Extracting Text Lines in Handwritten Documents", Claudie Faure Ecole Nationale SupCrieure des T&communications, CNRS-URA 82046 rue Barrault,**1995.**

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

[12] A.V. Aho, J.E. Hopcroft and J.D. Ullman , "Data Structures and Algorithms", Addison- Wesley, **1983**.

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

[14] V. N. Manjunath Aradhya and C Naveena ,"Text Line Segmentation of Unconstrained Handwritten Kannada Script", In the proceedings of ICCCS'11,**pp.**231-23, **2011.**

[15] M.K Jindal, R. K. Sharma & G.S. Lehal , "Segmentation of Horizontally Overlapping Lines in Printed Indian Scripts", International Journal of Computational Intelligence Research, ISSN 0973-1873 Vol.3, No.4, **pp.** 277–286,**2007.**

[16] G. Louloudis, B. Gatos, I. Pratikakis & K.Halatsis, "A Block-Based Hough Transform Mapping for Text Line Detection in Handwritten Documents", Proceedings of the Tenth International Workshop on Frontiers in Handwriting Recognition, La Baule, Oct**. 2006.**

[17] B.M.Sagar, Dr.Shobha G and Dr. Ramakanth kumar P, "OCR for printed kannada text to Machine editable format using Database approach", 9th WSEAS International Conference on AUTOMATION and INFORMATION (ICAI'08) , Bucharest , Romania , June24-26 , **2008.**

AUTHORS PROFILE

Dr. Mamatha H. R.
She received her B E degree in Computer Science and Engineering from the Kuvempu University in 1998 and the M.Tech degree in Computer Networks and Engineering from the Visvesvaraya Technological University in 2006. She obtained her Doctoral Degree from Visvesvaraya Technological University. She has total 17+ years of teaching experience. Her current research interests include Pattern Recognition and Image Processing. She has published 25+ international papers. She is a life member of Indian Society for Technical Education, MIR Labs and IACSIT. She is a reviewer for various international conferences and journals. She has mentored students for various competitions at international level including the Windows Embedded Students Challenge Competition-2006 held at Microsoft Campus, Redmond, Seattle, USA. Currently she is working as Professor in the Department of Information Science and Engineering, P E S Institute of Technology.

Chethana H.T.
She received her B E degree in Computer Science and Engineering from the VTU University in 2011. She is currently pursuing M.Tech degree in Software Engineering branch from P E S Institute of technology under Visvesvaraya Technological University. Her current research interests include Pattern Recognition and Image Processing. She is currently working under the project titled "Segmentation algorithms for Handwritten Kannada Documents" in M.Tech under Image Processing domain.