ISSN:0975-9646

Comparative Study of Text Line Segmentation on Handwritten Kannada Documents

Chethana H T[#1], Mamatha H R[#2]

Department of Computer Science & Engineering ,VVCE,Mysuru,570002,India

Department of Information Science & Engineering , P.E.S Institute of Technology, VTU,Bangalore,560085,India

***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. The aim of this paper is to investigate different text line segmentation techniques like Projection profiles, Run Length Smearing method, Median segmentation and Bounding box method on Handwritten Kannada documents. The above said methods have been tested on two different datasets of varying complexity. A total number of 200 samples is used for experimentation. These methods are experimented and compare their accuracy and results.**

***Keywords*— Segmentation, Handwriting, Textlines, Handwritten Kannada Documents.**

I. INTRODUCTION

A document has a structure which contains information in it. If the document does not have proper structure then it is very difficult to get back the information contained in a document. Document structure analysis is a vital stage in optical character recognition. A method of dividing the document regions in to text and non-text regions is known as document segmentation. For document structure extraction, text line extraction can be seen as a pre-processing step. Handwritten Kannada documents poses challenges due to overlapping lines, touching lines, curved lines, additional modifiers, consonants, intra and inter word gaps.

The objective of this paper is to investigate different text line segmentation techniques like Projection profile, Run length smearing, Median segmentation, Bounding box method. These approaches are applied on Handwritten Kannada documents.

This paper is organized as follows: Section 2 describes complexity of text line segmentation, section 3 describes investigation of different text line segmentation methods, section 4 deals with results and discussions about text line segmentation and section 5 describes the conclusions of this paper.

II. TEXT LINE SEGMENTATION COMPLEXITY

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.

A. 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.

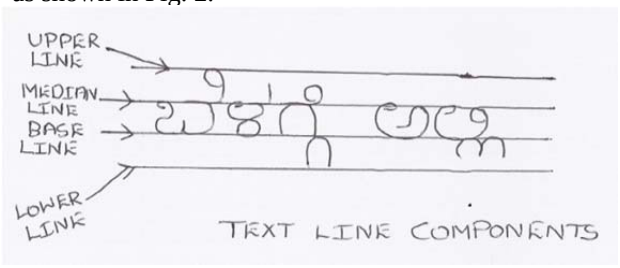


Fig. 1 Text Line Components

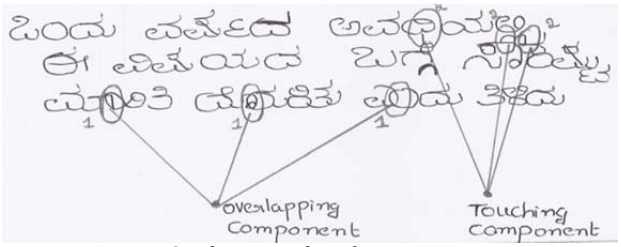


Fig. 2 Overlapping and touching components

B. 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.

C. 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.4.

Smudges means each pixel which is present in the source image spreads on the periphery of the surrounding pixels as shown in Fig.3.

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 as shown in Fig.5.

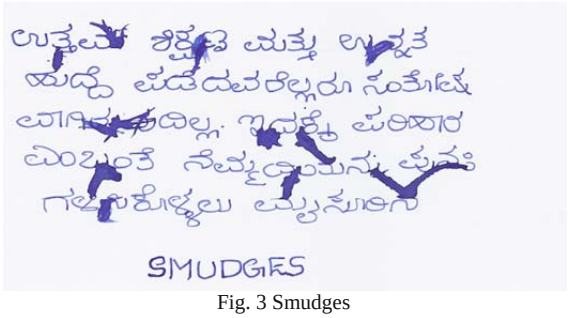


Fig. 3 Smudges

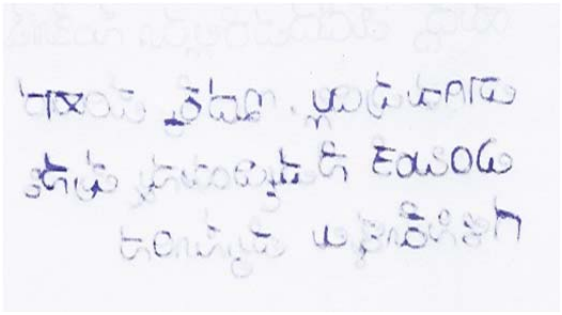


Fig. 4 Presence of seeping ink from other side of the document

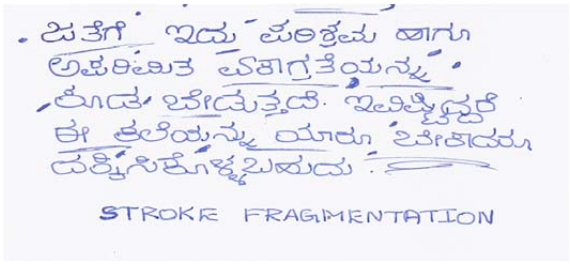


Fig. 5 Stroke fragmentation

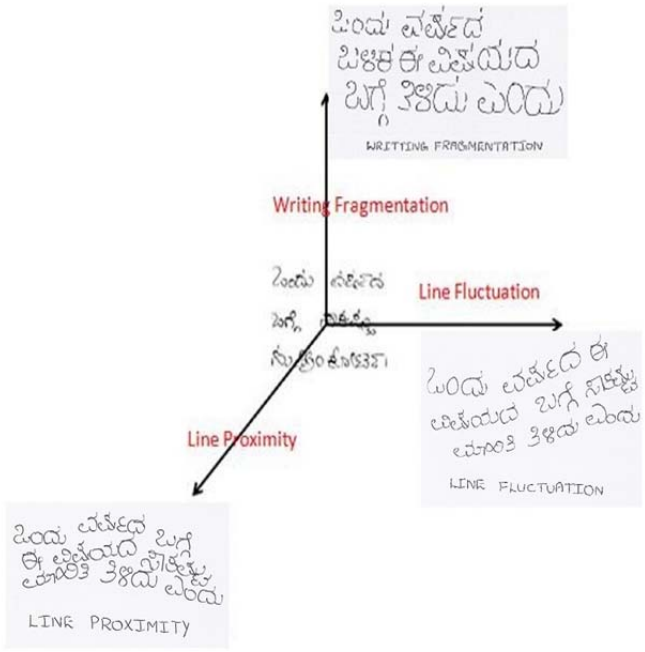


Fig. 6 The three main axes of document complexity for text line segmentation.

III. TEXT LINE SEGMENTATION METHODS

In this section various segmentation methodologies are proposed and discussed in detail.

A. An approach based on Projection Profile

A projection profile is a histogram giving the number of ON pixels accumulated along parallel lines. By looking for minima in horizontal projection profile of the page, we can separate the lines easily[4].Horizontal projection profile is used for text line segmentation. This approach comprises of two stages - pre processing followed by morphological operations and text line extraction.

Algorithm

Begin

Input: Handwritten Kannada text document

Output: Segmented lines

Step 1: Binarise the original image.

Step 2 :Morphological operations such as erosion followed by dilation operations are applied.

Step 3 : Find out the row number where each segment falls.

Step 4 : When there is a continuous white space and once a black pixel is found ,then that row number is marked as starting index of a line.

Step 5 : Ending index of a line is obtained by marking the row number where there is a continuous white space in a row.

Step 6 : Using the row numbers of starting and ending index of a line, height of the segment can be found easily.

Step 7 : Crop the image using height as a parameter with width of a segment constant and save it in a separate file.

End

B. An approach based on Run Length Smearing method

In this method, length of the white run is computed by finding the consecutive white pixels which appears in between two black pixels. We will fill up the white run length into black, when the length of white run is less than five times width of the stroke.

Algorithm

Input: Binarised Handwritten Kannada text document

Output: Segmented lines

Begin

Step 1 : Load the binarised image.

Step 2: Morphological operations such as erosion followed by dilation operations are applied to the binarised image.

Step 3:Consecutive black pixels along the horizontal direction are smeared. i.e. the white space between them is filled with black pixels if their distance is within a predefined threshold.

Step 4: If the distance exceeds the threshold then white space between them is not filled with black pixels.

Step 5: Starting and ending index of a line is found by marking the row numbers.

Step 6: By locating these row numbers height of a segment can be found easily.

Step 7: Crop the image using height as a parameter with width of a segment constant and save it in a separate file.

End

C. An approach based on Median Segmentation

The median segmentation method is used for segmenting the document in to individual text lines. This method has been applied to English documents. In this method, two basic assumptions are made.

1) Usually in all languages like Kannada, English, Tamil etc the distribution of pixels is more denser in the middle part rather than the upper and lower part of a line.

2) Pixel distribution is almost similar in both upper and lower part of a line.

Algorithm

Input: Binarised Handwritten Kannada text document

Output: Segmented lines

Begin

Step 1 : Load the binarised image.

Step 2 :Morphological operations are used for constructing bridge between the components. Dilation and erosion are two primitive morphological operations that can be applied to the binarised image[4].

Step 3:Histogram is constructed to find out the median line.

Step 4:Threshold is set for the histogram. Depending on the threshold, where the density of the black pixels are more, then that row number is marked as median line.

Step 5:The threshold is set for both upper and lower part of the line. By using the threshold values, height of the segment can be found .

Step 6 : Crop the image using height as a parameter with width of a segment constant and save it in a separate file.

End

D. An approach based on Bounding Box

A technique based on Bounding Box is used in order to extract individual text line. First the image is converted to gray scale and histogram of that image is plotted. Next find the row containing lesser number of white pixels and identify the measurements of centroids with the region props property. Finally with the help of measurements of centroids individual lines are cropped.

Algorithm

Input: Handwritten Kannada text document

Output: Segmented lines

Begin

Step 1 : Load the binarised image.

Step 2 : Morphological operations are used for constructing bridge between the components. Dilation and erosion are two primitive morphological operations that can be applied to the binarised image[4].

Step 3 : Histogram is constructed for the binarised image.

Step 4 : Using the histogram, find the rows containing lesser number of white pixels.

Step 5 : Find the centroids of the above rows using length and width as parameters.

Step 6 : Measurement of centroids is calculated by varying the threshold values.

Step 7 : Using these measurement, mark the bounding box for text lines.

Step 8 : Copy the pixels in bounding box and save in to separate file.

End

IV. RESULTS AND DISCUSSIONS

For experimental purpose we have considered 200 handwritten Kannada documents collected from the authors of [4] and [9].The datasets are named as DS1[4] and DS2[9].

DS1 data set Collection : 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. These documents are named based on sequence starting with the word “Kannada” followed by the symbol “_”, continued by a digit of 1, 2, 3 ,4(e g : "Kannada_1_001") and stored in "TIF" format. The first image file in the database is named as “Kannada_1_001.TIF”.

DS2 data set Collection : It contains 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. Author[4] has collected documents in such a way that documents with 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.

From the study it is seen that morphological operations with Projection Profile gives the best segmentation rate of 93.87% among all other proposed methods because this method works well for clearly separated lines and this method cannot divide the touching or overlapping lines and instead it will merge those lines.

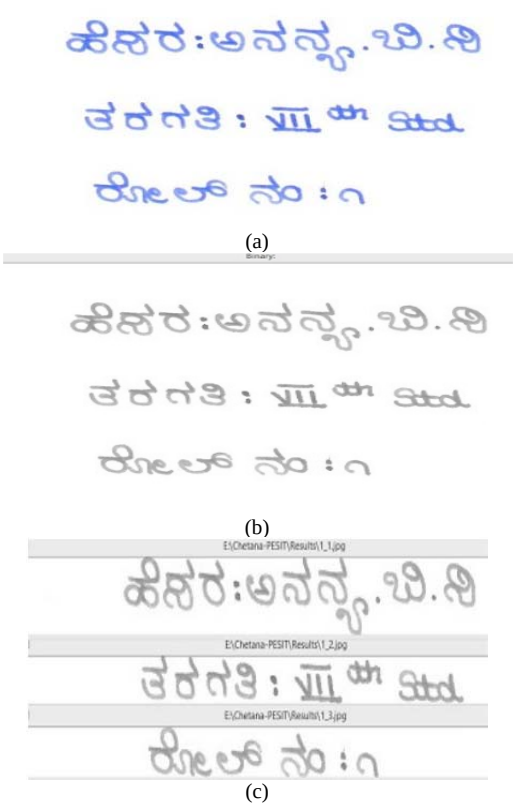
Table I shows the comparative study of different proposed methods on two different datasets. Table II shows the comparative study of our proposed method with the existing methods. To compare our proposed method with the existing work is very difficult as very few work exist in the line segmentation of handwritten Kannada script which is experimented on different datasets of complexity.

TABLE I Comparison of the results of different proposed methods for line segmentation on two different datasets

Segmentation Method	**Recognition Accuracy**	
	DS1[4]	**DS2[9]**
Projection Profile	92.87%	86.27%
Projection Profile with morphology	93.87%	87.56%
Median Segmentation	89.65%	87.27%
Median Segmentation with morphology	90.5%	87.77%
Run Length Smearing method	89.5%	88.2%
Run Length Smearing method with morphology	89.9%	88.6%
Bounding Box method	91.2%	89.6%
Bounding box with morphology	92%	90.4%

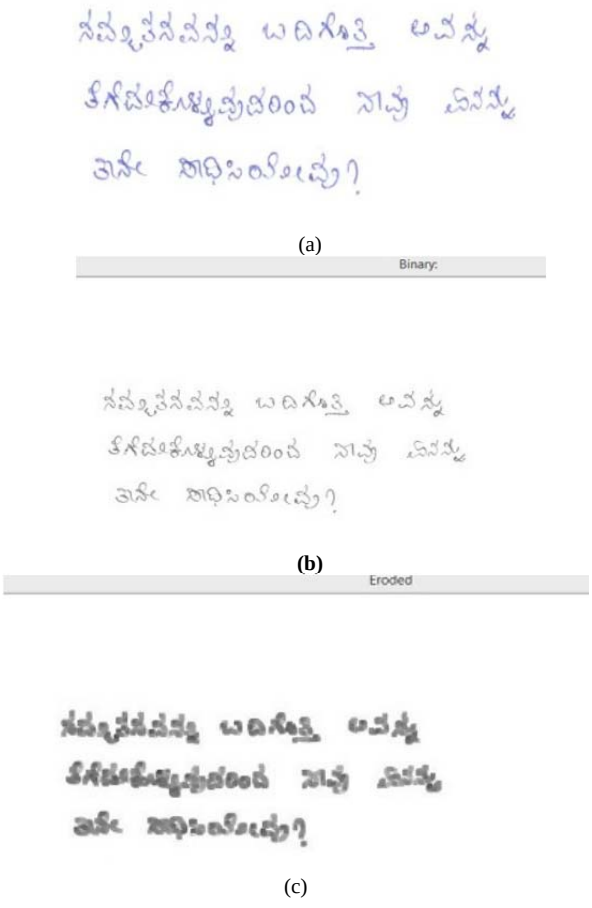
TABLE II Comparison of the results of proposed method with the existing methods for line segmentation.

Author	**Segmentation Method**	**Size of Dataset**	**Segmentation Rate**
Alaei et.al., [13]	Potential Piece-wise Separation technique	204	94.98%
Alaei et.al., [13]	Stripe based approach	204	95.32%
Aradya et.al., [14]	Component extension technique	250	Not specified
Proposed method on DS1[4]	Morphological operations and Projection profiles	100	93.87%
Proposed method on DS2[9]	Morphological operations and Projection profiles	100	87.56%

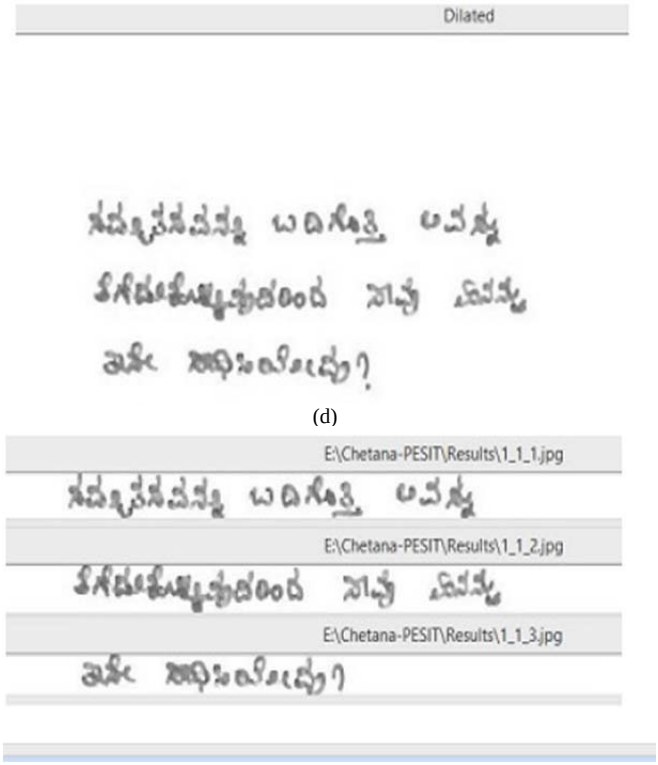


(a) (b) (c)

Fig. 6 Results of segmentation using Projection Profile (a) original document (b)binarised document (c)segmented document



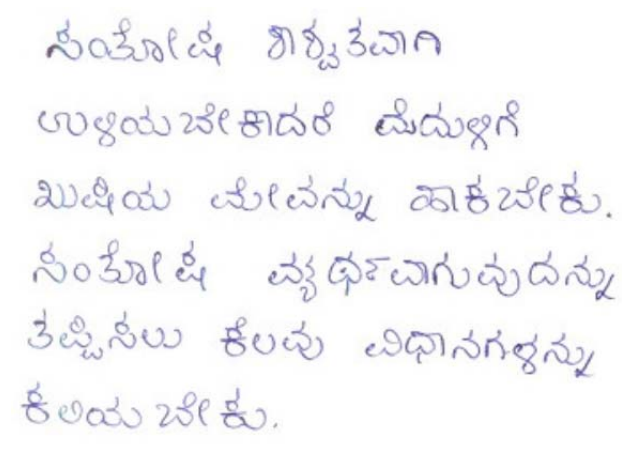
(a) **(b)** (c)



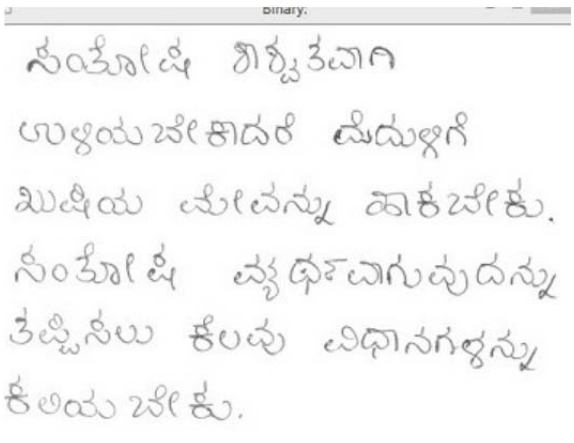
(d)

(e)

Fig. 7 Results of segmentation using morphology with Projection Profile (a)original document (b) binarised document (c) Eroded image (d) Dilated image (e)segmented lines



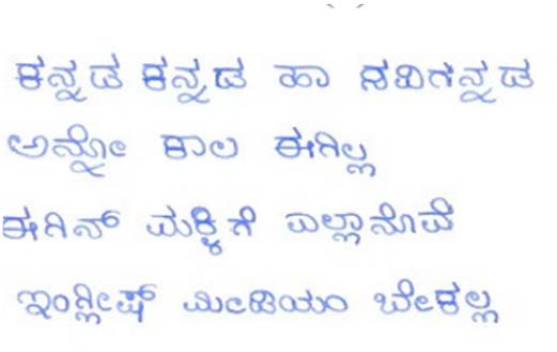
(a)



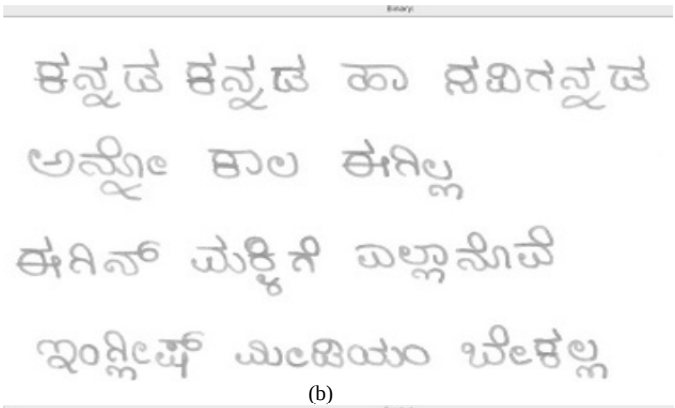
(b)

(c)

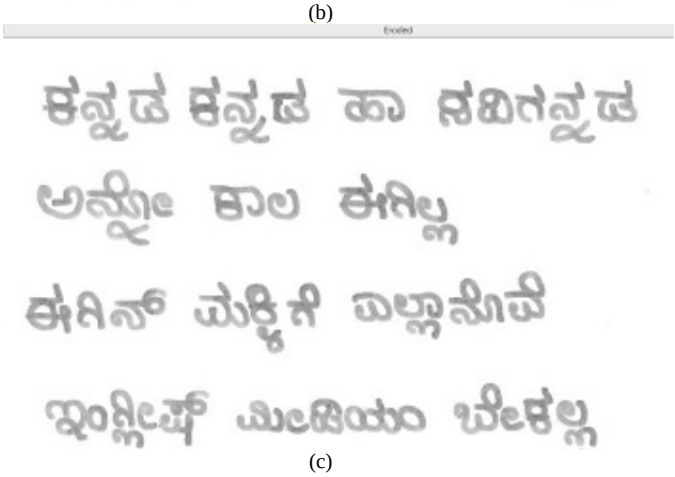
Fig. 8 Results of segmentation using Median Segmentation (a) original document(b) binarised document (c) segmented lines



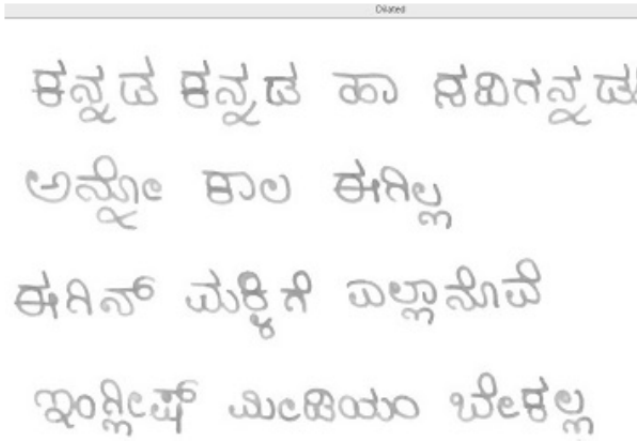
(a)



(b)



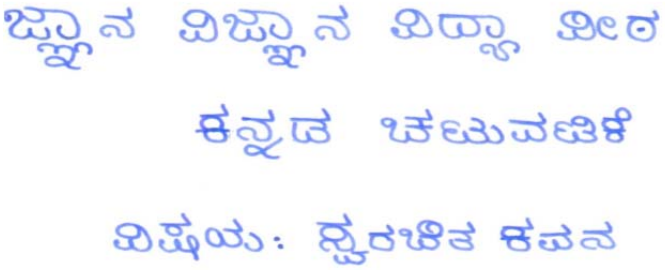
(c)



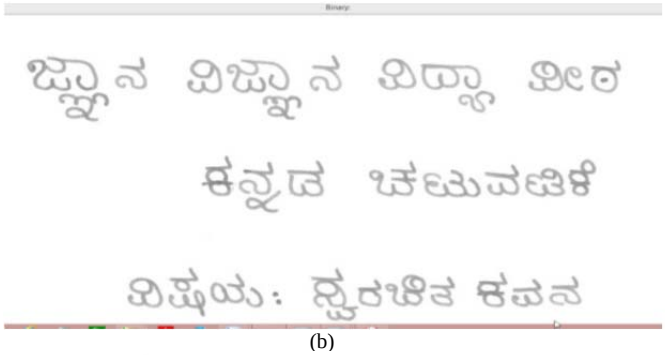
(d)

(e)

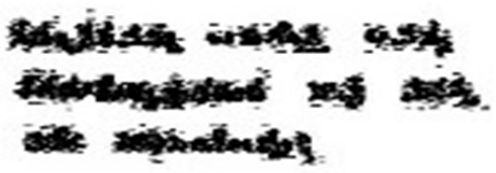
Fig. 9 Results of segmentation using morphology with Median Segmentation (a)original document (b) binarised document (c) Eroded image (d) Dilated image (e)segmented lines



(a)



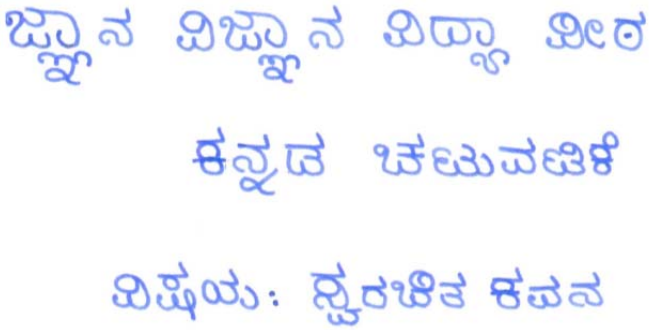
(b)



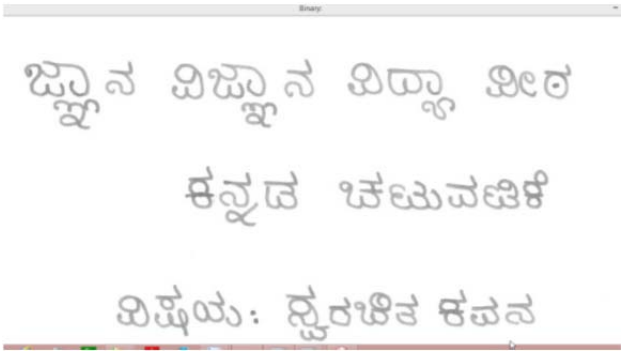
(c)

(d)

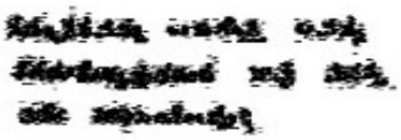
Fig. 10 Results of segmentation using morphology with Median Segmentation (a)original document (b) binarised document (c) Eroded image (d) Dilated image (e)segmented lines



(a)



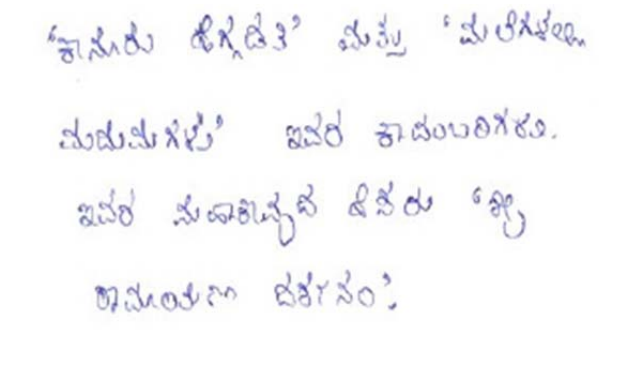
(b)



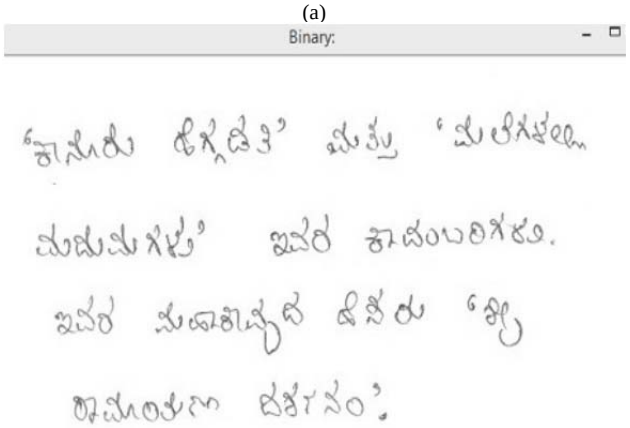
(c)

(d)

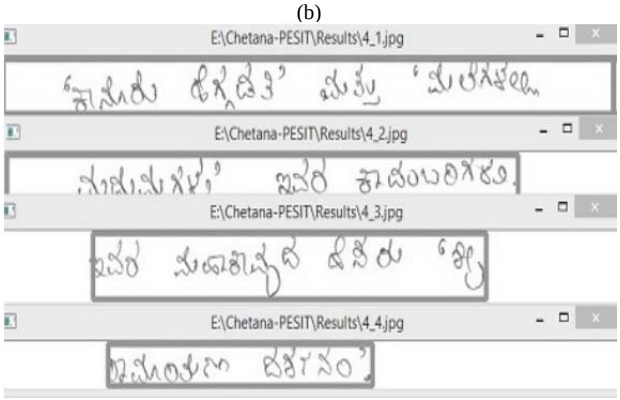
Fig. 11 Results of segmentation using Run length Smearing method (a) original document (b) binarised document (c) Smeared image (d) segmented lines



(a)

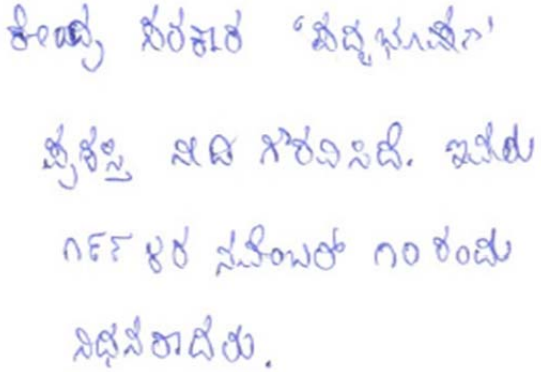


(b)

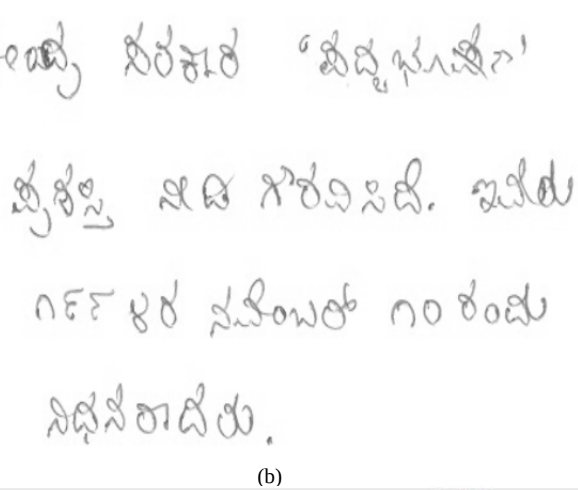


(c)

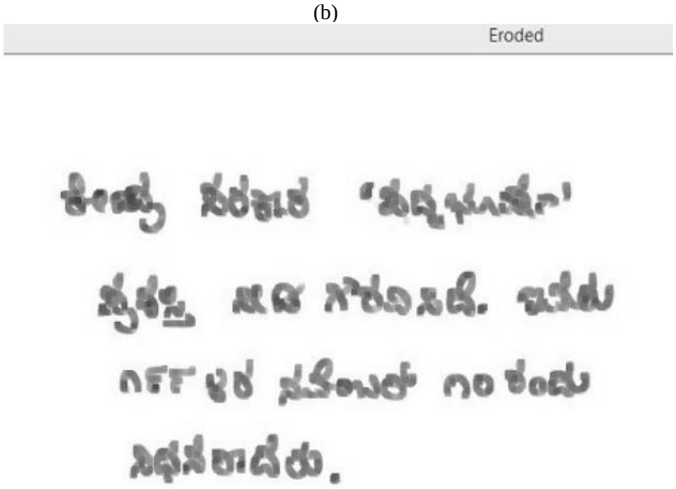
Fig. 12 Results of segmentation using Bounding Box method(a) original document(b) binarised document (c) segmented lines



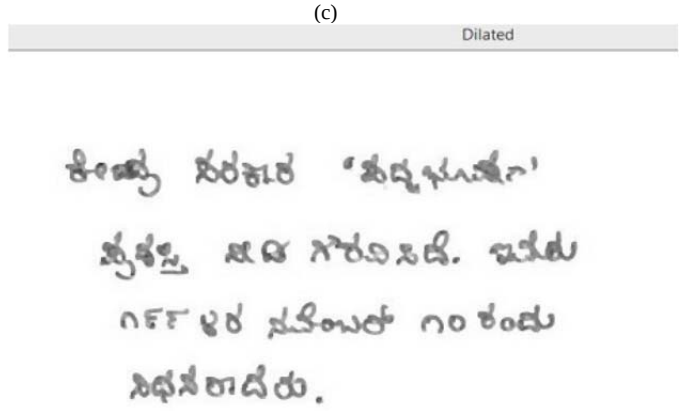
(a)



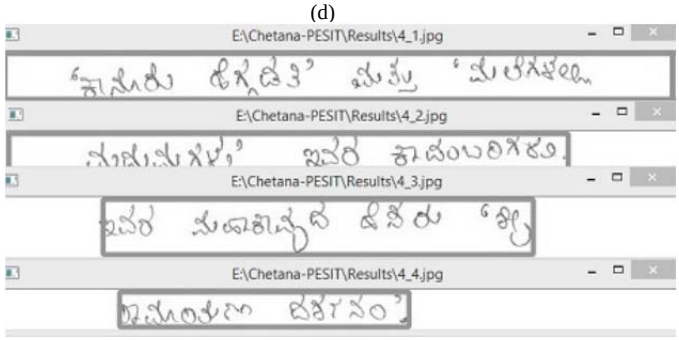
(b)



(c)



(d)



(e)

Fig. 13 Results of segmentation using morphology with Bounding Box method(a)original document (b) binarised document (c) Eroded image (d) Dilated image (e)segmented lines

V. CONCLUSIONS

Developing an OCR for handwritten Kannada documents is quite challenging and prone to errors due to its structural complexity and increased character set in Kannada language. An attempt is made in this direction and extraction of lines is done considering documents with different handwriting styles. But the accuracy obtained from the proposed methods are reduced because we have considered different documents with different handwriting styles. The accuracy for the documents with good handwriting style with less skew would have been much more higher than what we have obtained.

Four different methods like Projection profiles, Run Length smearing method, Median segmentation and Bounding box methods are proposed for text line extraction of Handwritten Kannada Documents. These proposed methods are experimented on two different datasets named as DS1 and DS2 collected from the authors of [4] and [9].

Morphological operations with projection profile gives the best segmentation rate of 93.87% among all other proposed methods because this method works well for clearly separated lines and this method cannot divide the touching or overlapping lines and instead it will merge those lines.

ACKNOWLEDGMENT

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 & Vijaya MS, *Genetic Programming for Document Segmentation and Region Classification using Discipulus Perceptron,(IJARAI)* International Journal of Advanced Research in ArtificialIntelligence,Vol.2,No.2,2013.

[2] Rafael C. Gonzalez, Richard E. Woods and Steven L.Eddins ,*Digital Image Processing using MATLA*,. 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, IJDAR(2007)* 9:123–138 DOI 10.1007/s10032-006-0023-z.

[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*,In the proceedings of 2010 Second Vaagdevi International Conference on Information Technology for Real World Problems,2010,pp8388.

[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*,2011 International Conference on Document Analysis and Recognition.

[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 and H. Srinivasan, S. N. Srihari, *A Statistical Approach to Handwritten Line Segmentation*,In Proceedings of SPIE Document Recognition and Retrieval XIV,SanJose,CA,February2007.

[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),2011,pp.917–928.

[14] V. N. Manjunath Aradhya and C Naveena ,*Text Line Segmentation of Unconstrained Handwritten Kannada Script*, In the proceedings of ICCCS'11,2011,pp231-234.

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

[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, June 24-26,2008.