A Study of different Text Line Extraction Techniques for Multi-font and Multi-size Printed Kannada Documents

R Prajna
Department of Information Science and Engineering,
P.E.S Institute of Technology
Bangalore, India
prajna.ramachandrabhat@gmail.com

Ramya V R
Department of Information Science and Engineering,
P.E.S Institute of Technology
Bangalore, India
vrramya3@gmail.com

Dr.Mamatha H.R
Associate professor
Department of Information Science and Engineering,
P.E.S Instituteof Technology
Bangalore, India
mamathahr@pes.edu

ABSTRACT
Line and word segmentation is one of the important step of OCR systems. For the identification of printed characters of non-Indian languages like English, Japanese, Chinese Optical Character Recognition (OCR) systems have been effectively developed. For Indian Languages, efforts are on the way for the development of efficient OCR systems, mainly for Kannada, one of the popular South Indian language .In this paper we have proposed a robust method for extraction of individual text lines for printed kannada documents based on the efficient segmentation methodologies such as morphology operations based projection profile ,projection profile and bounding box.

Keywords
Optical character recognition, Printed Characters,projection profile,morphology operations based on projection profile,bounding box.

1. INTRODUCTION
Optical Character Recognition (OCR) is one of the oldest sub fields of pattern recognition with a rich contribution for the recognition of printed documents.

Due to the affect and the advancements in the Information Technology, now a days in Karnataka more emphasis is given to use Kannada at all levels hence the use of Kannada in computer systems is also a necessity. So, efficient OCR systems for Kannada language are one of the most important present day requirements. Currently there are many efficient OCR systems available for handling printed English documents and also available for many European languages as well as some of the Asian languages such as Chinese, Japanese etc. However, there are not many recognized and reported efforts at developing OCR systems for Indian languages especially for a South Indian language like Kannada [1].

Segmentation of a document image into its basic entities namely text lines and words, is a critical stage towards printed document recognition. The difficulties that arise in printed documents make the segmentation procedure a challenging task. There are many problems encountered in the segmentation procedure. Text line detection is a major component in a document image analysis system, and also a preprocessing step for tasks such as character recognition,extraction of document structure provides information like character recognition, zone segmentation, skew correction .It includes difficulties like skew angle between the lines on the page ,adjacent text line touching . The difficulty in analysis of machine printed document lies in quality of image and complex layout structure. In this paper a methodology based on morphological operations based projection profile, bounding box, projection profile for segmentation of the printed Kannada script into lines is proposed.

The rest of the paper is organized as follows. Section 2 Explains the literature survey,Section 3 describes the characteristics of Kannada script, section 4 describes the challenges involved,section 5 discusses about the proposed methodology, and section 6 briefly discusses the experimental setup and the results obtained are discussed respectively. Finally in Sections 7 and 8, comparative study and conclusions are made.

2.LITERATURE SURVEY
Some of the schemes that are reported in the recent works for line and word segmentation approach for printeddocuments in Kannada and another Indian language Devnagari, Bangla, Teluguare as follows:

A robust method to extract individual text line has been proposed in [1].To extract the individual text lines modified histogram applied which was obtained from run length based

smearing. Foreground and background information is also used for accurate line segmentation. The contour points of the component are traced to take care of the problem of overlapping.

An effiecient approach to extract the text lines and skew correction of extracted text lines uses a new cost function which is mentioned in [2].This approaches consider the spacing between text lines and skew of each text line.The proposed approach normalizes the lower baseline to a horizontal line using a skating window approach,in order to correct the baseline skew The author claims that baseline correction approach highly improves the performance based on experimental results.

An morphological approach to extract textlines from palm script documents has been propsed in [3]. This paper explains an approach for extracting the line segments of a palm leaf script document image in an unsupervised way.Morphological operations and Connected Components Analysis (CCA) has been adopted to extract the lines from palm script document image written in Kannada.One of the morphological operation closing is used for connecting the characters in a line.Connected component analysis is used after the closing operation is performed in order to extract the connected components. The author claims that proposed method is computationally efficient for text line extraction and even addresses touching lines and curved lines.

A bounded box method for segmentation of document lines,words and characters has been proposed in [4].The method is based on pixel histogram obtained where horizontal histogram of an image is obtained,white pixels in each row is counted .With the help of histogram , the rows containing no white pixel is found and all such rows are replaced by 1,then the the image is inverted to make empty rows as 0 and text lines will have original pixels and the Bounding Box for text lines are marked.

An approach as been proposed in [5] to extract the text lines by vertically decomposing document into parallel pipe structures called stripes. Each row of a stripe is painted by a gray intensity, which is the average intensity value of gray values of all pixels present in that row-stripe.The painted stripes are then converted into two-tone and using some smoothing operations, the two-tone painted image is smoothed. A dilation operation is employed on the foreground portion of the smoothed image to obtain a single component for each text-line.

An automatic technique of separating the text lines using script characteristics and shape based features is presented in [6].

Neural Classifier based approach has been presented in [7] where the proposed method handles different font sizes and font types.Neural classifiers have been effectively used for classification of characters based on moment features.The Scheme of feature extraction is selected using moments and RBF neural networks as classifiers to identify and classify characters.The proposed method showed an encouraging recognition rate.

Schemes for skew detection and correction have been proposed in [8],where bounding box,hough transform,contour detection techniques have been used.An average recognition rate of 91% is obtained by using above mentioned techniques.

From the literature survey it is observed 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 printed Kannada documents.Segmentation of printed Kannada documents into lines,words and character is of great importance and much demanded by some specific applications.Segmentation of printed Kannada documents poses challenges due to additional modifier characters,writing styles,inter and intra word gaps.This motivated us to design effective schemes for text line extraction from printed Kannada documents which can then be used for word and character segmentation in turn this can be used in later stages of OCR so that the performance of subsequent steps in document image analysis would be more accurate.

.

3. THE CHARACTERISTICS OF KANNADA SCRIPT

Kannada is one of the four famous Dravidian languages of South India. Kannada is written horizontally from left to right. Lower and upper case is absent in Kannada. Kannada is a non-cursive script. That is, without joining the characters of the word .Kannada words is written. Within a word characters are isolated..

Kannada language consists 13 vowels and 34 consonants asthe basic alphabet of the language as shown in figures 3.1 and 3.2 respectively.

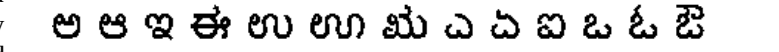


Figure 3.1. Vowels of Kannada Script

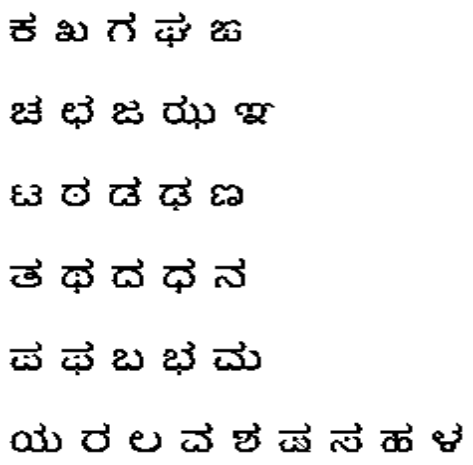


Figure 3.2. Consonants of Kannada Script

Each vowel has a vowel sign (modifier) and each consonant has a basic form (primitive).A basic form of consonant can combine with the vowel sign to form another set of 13 Consonant–Vowel (CV) composite characters called as 'gunithakshara'. In Kannada, all the 34 consonants have a Short/half form, refered as 'Vatthus', which can be usually called as subscripts or half consonants . Any half consonant can appear below any other consonant or a CV character as subscript character to form a conjunct-consonant character.
Some of the complex characters are listed below. The following figure shows the conjunct consonant (Vatthu)

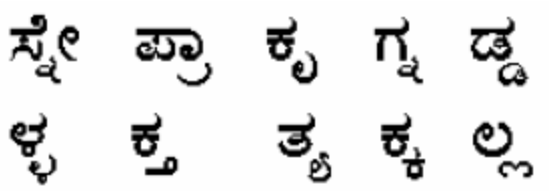


Figure3. 3. Shows the Conjunct Consonant (Subscript/Vatthu)

4. CHALLENGES INVOLVED

In this section categorizing the challenges involved in the segmentation of the printed text-lines. When dealing with printed text, line segmentation has to solve some obstacles, among the most predominant are:
1. Multi column documents.
2. Noisy documents.
3. Documents includes non-constant spaces between text lines, words and also with characters.
4. Documents consists marginal text.
5. Documents with various font sizes coexist.
6.Documents with graphical illustrations and ornamental characters.
7. Documents whose text is skewed and/or wrapped.

5.PROPOSED METHODOLOGY

In this section different method for segmentation of printed Kannada documentsinto lines is proposed.
Following are the proposed methods.

5.1 Horizontal Projection Profile

In order to extract individual text line, technique based on projection is used. A projection profile is a histogram which is giving the several number of ON pixels accumulated along parallel lines. Thus a horizontal projection profile is a one-dimensional [1D] array where each of the element denotes the number of ON pixels along a row in the image. Similarly a vertical projection profile gives the all column sums. It is easy to see that separating 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 one line. Such projection profile based methods are used for line, word and character segmentation.

To segment the document image into number of text lines, the valleys of the horizontal projection computed by a row-wise sum of black pixels are used,where the histogram height is least denotes the position between two consecutive horizontal projections can be determined as one boundary line. Document image is segmented into text lines using the obtained boundary lines.

5.2 Morphology

For extracting image components ,Mathematical morphology can be used as a tool. Image components are useful in the representation and description of region shape,such as skeletons, boundaries, and the convex hull. Dilation is a primitive morphological operation that grows or thickens objects in a binarized image.A shape which controls the extent of this thickening in a specific manner 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.

In terms of set operations,Mathematically dilation is defined.The dilation of A by B denoted$A \oplus B$, is defined as in equation 1,

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

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

In a binary image ,erosion "thins" or "shrinks" objects.Here also, as in dilation structuring elementcontrols the manner and extent of shrinking.

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

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

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 that component is removed. 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. 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.

5.3 Bounding Box

In order to extract individual text line, technique based on bounding box is used. First the image is converted to gray scale and histogram of that image is plotted.Next find the white spaces and identify the measurements of centroids with the regionprops property which calculates centroid forconnected components in the image.Regionprops computes various properties of the individual objects in the binary image. The method result is a structure array including an entry per property per object.Finally with the help of measurements of centroids individual lines are cropped.

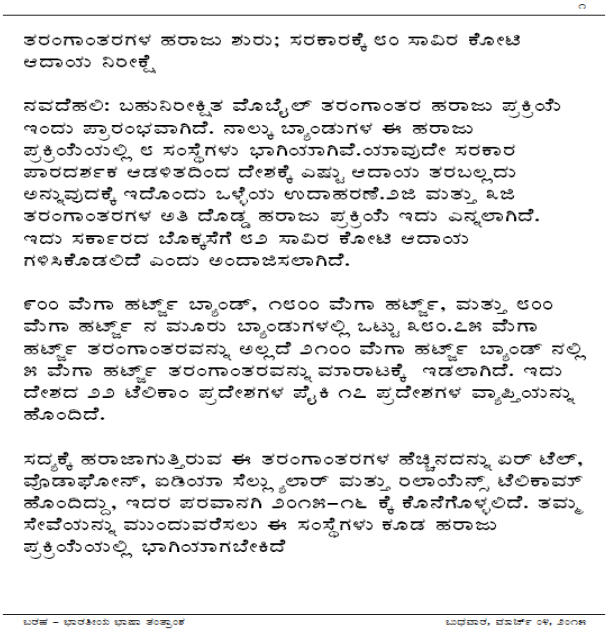
6.EXPERIMENTAL RESULTS

This section presents the results of the experiments conducted to study the performance of the proposed method on document dataset.For experimental purpose, we have considered 35 printed kannada document images collected from baraha software.The data set contains varieties of font styles such as BRH Kannada,BRH Amerikannada,BRH Srigandha,BRH Kailasam and font sizes.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. For the experimentationSingle column document images is considered. The number of lines in each document varies from 14 to 24 lines. For each document image, the corresponding ground truth information like the number of lines, words and characters is manually created .The total number of text lines is 433respectively.

Segmentation accuracy of 35 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. The proposed methodology for line segmentation of printed Kannada text document using the method based on Morphological operations and projection profiles gave an average line segmentation rate of 86% and was more effective than the method based on projection profiles which gave an average accuracy rate of 84% ,and the method based on bounding box gave an average accuracy rate of 80%.

Accuracy obtained from the proposed method is reduced because we have considered different documents with different font sizes(25,28,30 etc) and font styles.The accuracy for the documents with same font size and font styles would have been much more higher than what we have obtained.Some of the problems that were proposed 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 is present in the document.

6.1 Experimental Results for Morphology based Projection Profile



ತರಂಗಾಂತರಗಳ ಹರಾಜು ಶುರು; ಸರಕಾರಕ್ಕೆ ೮೦ ಸಾವಿರ ಕೋಟಿ ಆದಾಯ ನಿರೀಕ್ಷೆ

ನವದೆಹಲಿ: ಬಹುನಿರೀಕ್ಷಿತ ಮೊಬೈಲ್ ತರಂಗಾಂತರ ಹರಾಜು ಪ್ರಕ್ರಿಯೆ ಇಂದು ಪ್ರಾರಂಭವಾಗಿದೆ. ನಾಲ್ಕು ಬ್ಯಾಂಡುಗಳ ಈ ಹರಾಜು ಪ್ರಕ್ರಿಯೆಯಲ್ಲಿ ೮ ಸಂಸ್ಥೆಗಳು ಭಾಗಿಯಾಗಿವೆ.ಯಾವುದೇ ಸರಕಾರ ಪಾರದರ್ಶಕ ಆಡಳಿತದಿಂದ ದೇಶಕ್ಕೆ ಎಷ್ಟು ಆದಾಯ ತರಬಲ್ಲದು ಅನ್ನುವುದಕ್ಕೆ ಇದೊಂದು ಒಳ್ಳೆಯ ಉದಾಹರಣೆ.೨ಜಿ ಮತ್ತು ೩ಜಿ ತರಂಗಾಂತರಗಳ ಅತಿ ದೊಡ್ಡ ಹರಾಜು ಪ್ರಕ್ರಿಯೆ ಇದು ಎನ್ನಲಾಗಿದೆ. ಇದು ಸರ್ಕಾರದ ಬೊಕ್ಕಸಕ್ಕೆ ೮೨ ಸಾವಿರ ಕೋಟಿ ಆದಾಯ ಗಳಿಸಿಕೊಡಲಿದೆ ಎಂದು ಅಂದಾಜಿಸಲಾಗಿದೆ.

೯೦೦ ಮೆಗಾ ಹರ್ಟ್ಜ್ ಬ್ಯಾಂಡ್, ೧೮೦೦ ಮೆಗಾ ಹರ್ಟ್ಜ್, ಮತ್ತು ೮೦೦ ಮೆಗಾ ಹರ್ಟ್ಜ್ ನ ಮೂರು ಬ್ಯಾಂಡುಗಳಲ್ಲಿ ಒಟ್ಟು ೩೮೦.೭೫ ಮೆಗಾ ಹರ್ಟ್ಜ್ ತರಂಗಾಂತರವನ್ನು ಅಲ್ಲದೆ ೨೧೦೦ ಮೆಗಾ ಹರ್ಟ್ಜ್ ಬ್ಯಾಂಡ್ ನಲ್ಲಿ ೫ ಮೆಗಾ ಹರ್ಟ್ಜ್ ತರಂಗಾಂತರವನ್ನು ಮಾರಾಟಕ್ಕೆ ಇಡಲಾಗಿದೆ. ಇದು ದೇಶದ ೨೨ ಟೆಲಿಕಾಂ ಪ್ರದೇಶಗಳ ಪೈಕಿ ೧೭ ಪ್ರದೇಶಗಳ ವ್ಯಾಪ್ತಿಯನ್ನು ಹೊಂದಿದೆ.

ಸದ್ಯಕ್ಕೆ ಹರಾಜಾಗುತ್ತಿರುವ ಈ ತರಂಗಾಂತರಗಳ ಹೆಚ್ಚಿನದನ್ನು ಭಾರ್ ಟೆಲ್, ವೊಡಾಫೋನ್, ಐಡಿಯಾ ಸೆಲ್ಯುಲಾರ್ ಮತ್ತು ರಿಲಾಯನ್ಸ್ ಟೆಲಿಕಾಮ್ ಹೊಂದಿದ್ದು, ಇದರ ಪರವಾನಗಿ ೨೦೧೫–೧೬ ಕ್ಕೆ ಕೊನೆಗೊಳ್ಳಲಿದೆ. ತಮ್ಮ ಸೇವೆಯನ್ನು ಮುಂದುವರೆಸಲು ಈ ಸಂಸ್ಥೆಗಳು ಕೂಡ ಹರಾಜು ಪ್ರಕ್ರಿಯೆಯಲ್ಲಿ ಭಾಗಿಯಾಗಬೇಕಿದೆ

ಪ್ರಭ – ಭಾರತೀಯ ಭಾಷಾ ತಂತ್ರಾಂಶ

ಬುಧವಾರ, ಡಿಸೆಂಬರ್ ೨೪, ೨೦೧೪

Fig 6.1.1 Input Image

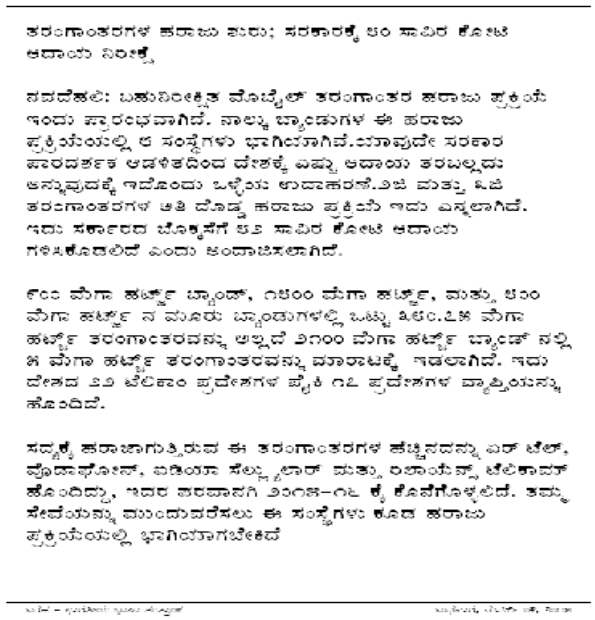


Fig 6.1.2 Gray Scale Image

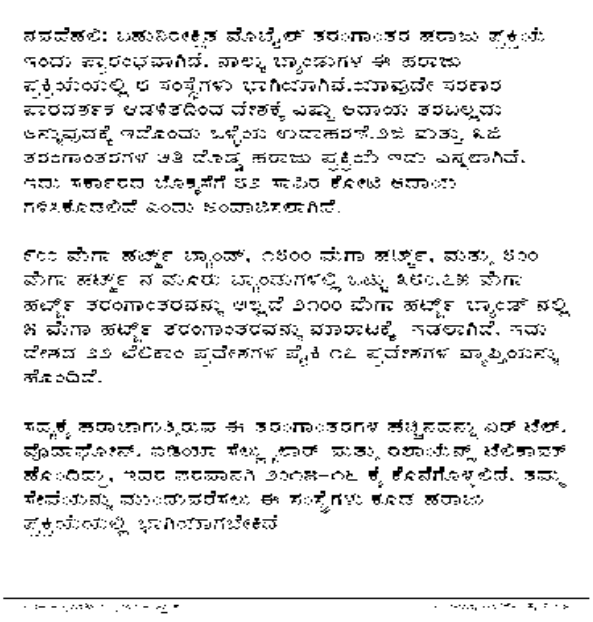


Fig 6.1.3 Binary Image

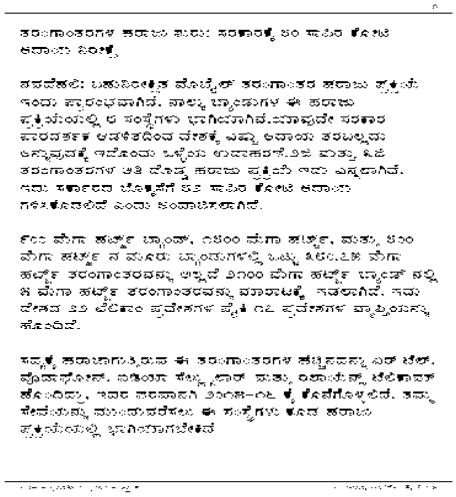


Fig 6.1.4 Image after Erosion

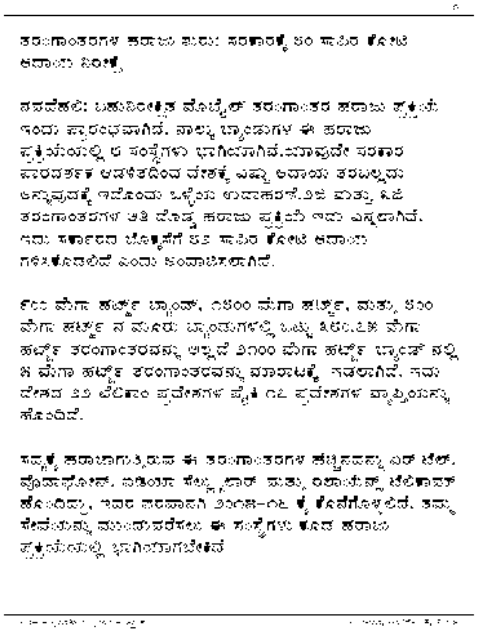


Fig 6.1.5 Image after Dilation

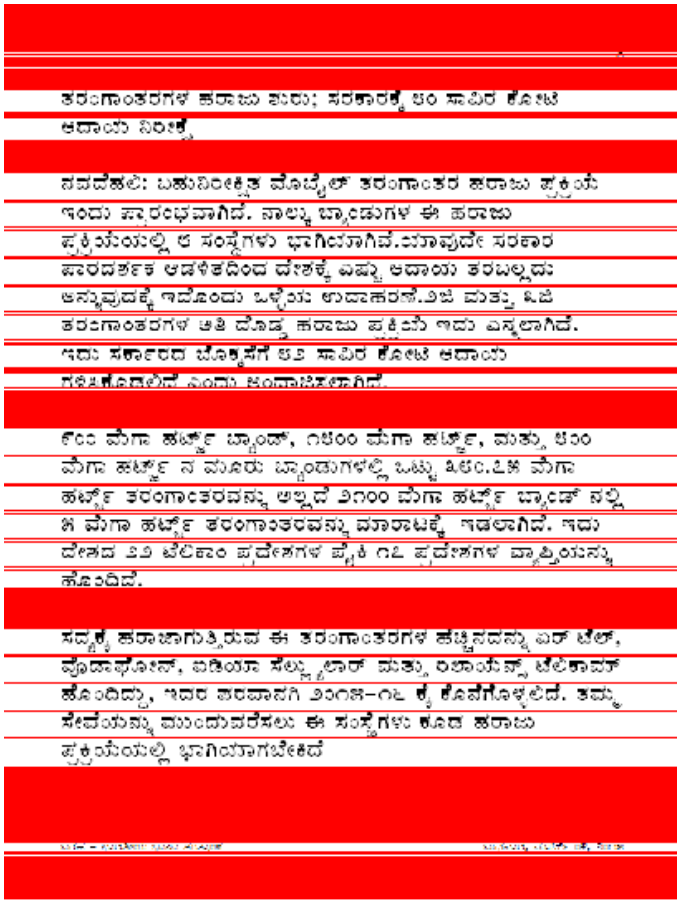


Fig 6.1.5 Detected Lines

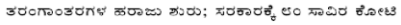


Fig 6.1.6 Extracted Line

6.2 Experimental Results for Horizontal Projection Profile

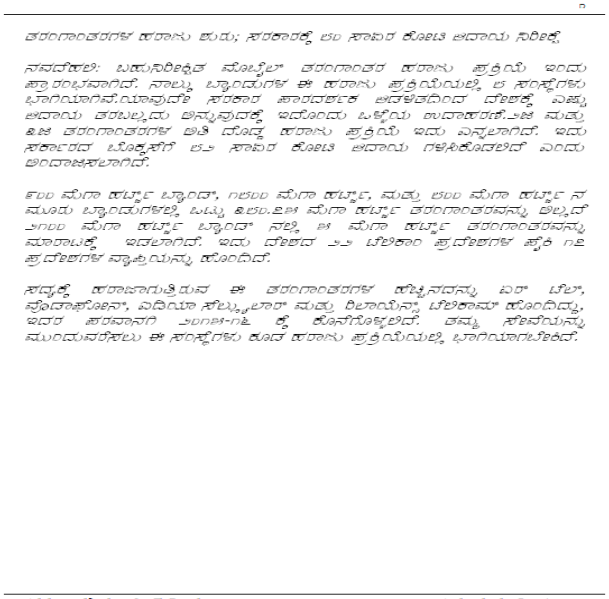


Fig 6.2.1 Input Image

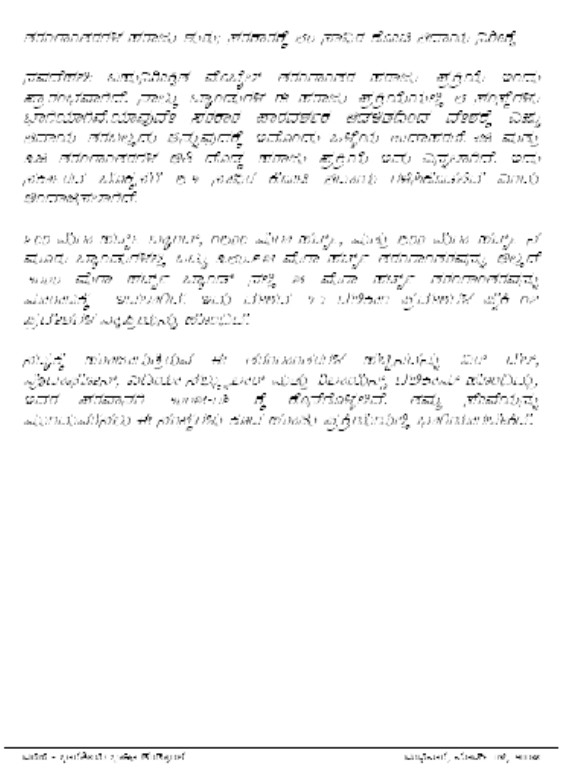


Fig 6.2.2 Gray Scale Image

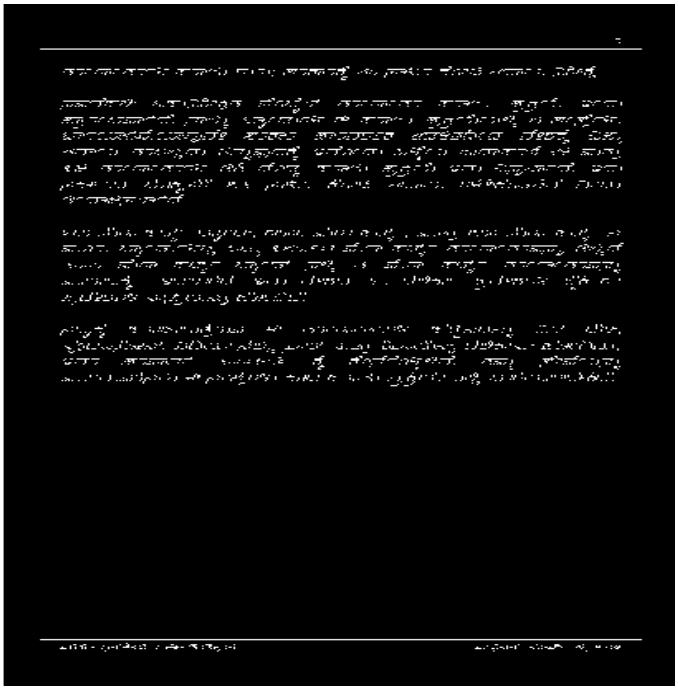


Fig 6.2.3 Binary Image

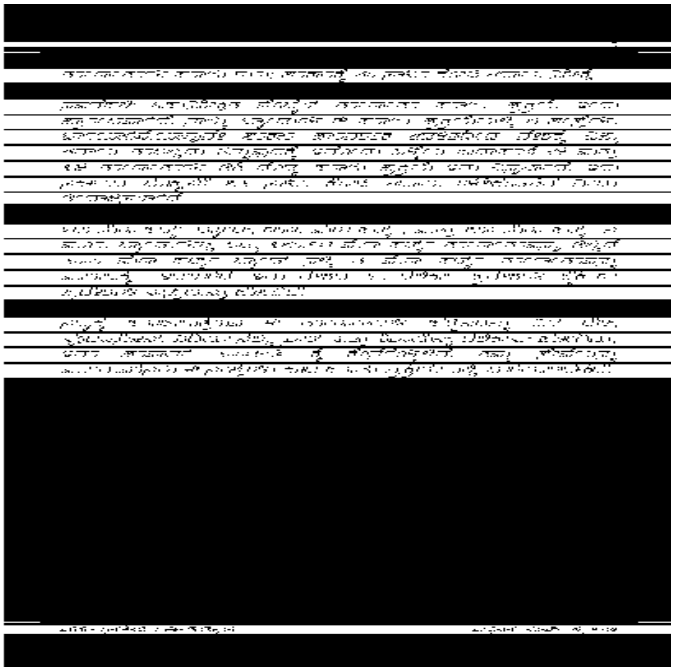


Fig 6.2.5 Detected Lines

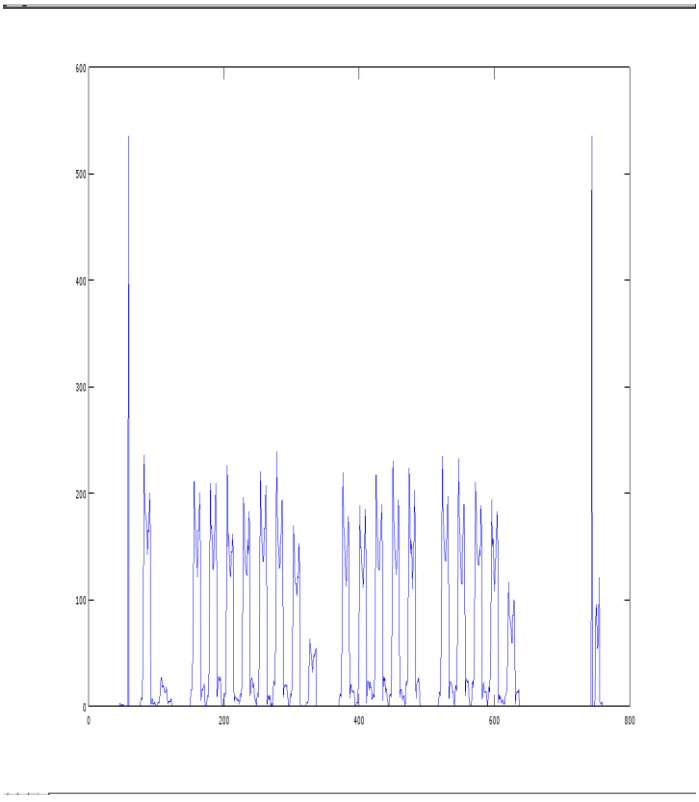


Fig 6.2.4 Horizontal Projection Profile

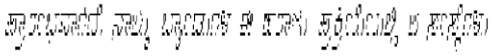
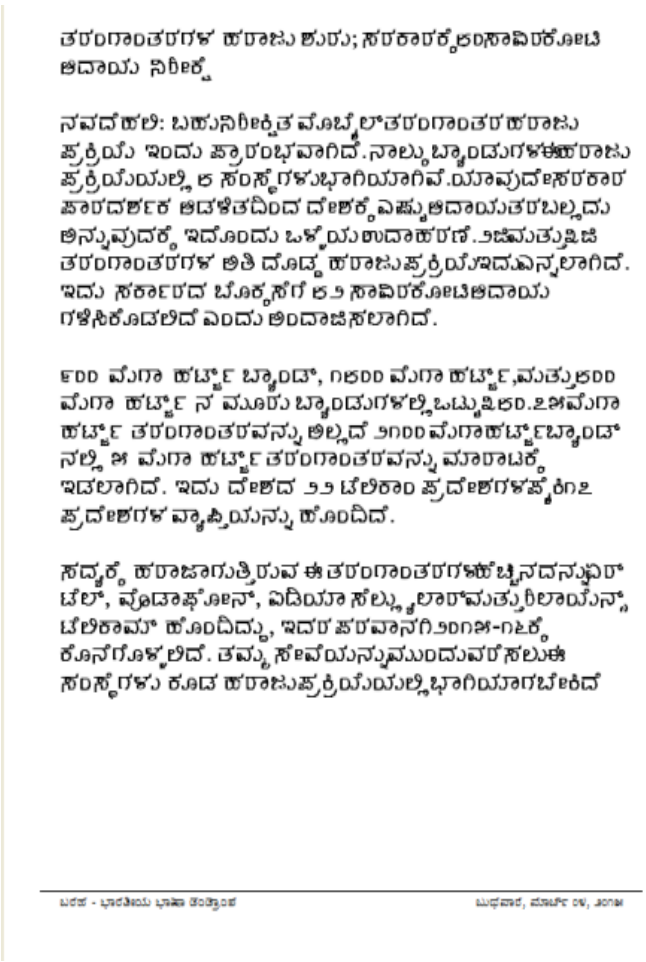
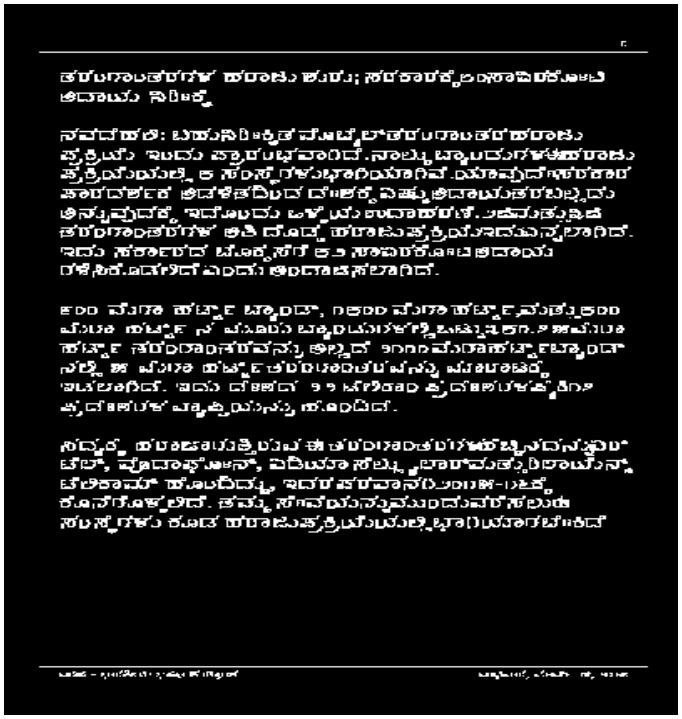


Fig 6.2.6 Extracted Line

6.3 Experimental results for Bounding Box



6.3.1 Input Image



6.3.2-Grayscale Image

6.3.3 –Extracted Line

7.COMPARATIVE STUDY

The Table 7.1 shows the comparison of proposed method for line segmentation. In order to analyze our method on the standard dataset, we collected the Kannada Printed Text Document from baraha Software.We have considered text categories like poems and general texts of Kannada.

For the experimentation 35 documents are considered. The proposed method based on Morphological operations and projection profiles is tested on this dataset as it was more effective than the other proposed method based on bounding box, horizontal projection profile for line segmentation and obtained an accuracy of 86%.

Table 7.1 Comparison of proposed method

sl.no	Segmentation Method	Segmentation rate(%)
1	Morphology Based	86
2	Horizontal Projection Profile	84
3	BoundingBox Based	80

8 CONCLUSIONS

Developing an OCR for printed kannada documents is quite challenging and prone to errors due to structural complexity and increased character set of Kannada language. An attempt is made in this direction and extraction of lines is done considering documents with different font sizes and font styles. But the accuracy obtained from the proposed method is reduced because we have considered different documents with different font sizes and font styles.The accuracy for the documents with same font size and font styles would have been much more higher than what we have obtained. Better noise removal techniques can be used to enhance the Preprocessing and Segmentation phases.Efficient extraction and classification methods are used to get good performance and accuracy of results. This can be further enhanced to recognize word,characters and convert the recognized characters to electronic form.

9.REFERENCES

[1] Nallapareddy Priyanka, Srikanta Pal, Ranju Mandal "Line and Word Segmentation Approach for Printed Documents",IJCA Special Issue on "Recent Trends in Image Processing and Pattern Recognition"RTIPPR, 2010.

[2]Sunanda dixit,Suresh Hosahalli Narayana,Mahesh Belur "Kannada text line extraction based on energy minimization and skew correction".

[3] B. Gangamma, Srikanta Murthy K, Riddhi J. Shah, Swati D V "Text Line Extraction from Palm Script Documents Using Morphological Approach"

[4]Vikas J Dongre , Vijay H Mankar "Devnagari document segmentation using histogram approach".International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.1, No.3, August 2011

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

[6] U. Pal and B. B. Chaudhuri **"**Script Line Separation From Indian Multi-Script Documents**".**In Proc. 4thICDAR.

[7]R. Sanjeev Kunte, R. D. Sudhaker Samuel"An OCR system for printed Kannada text using Two-stage Multi-network classification approach employing Wavelet features",International Conference on Computational Intelligence and Multimedia Applications 2007.

[8] Mamatha Hosalli Ramappa and Srikantamurthy Krishnamurthy "Skew Detection, Correction and Segmentation of Handwritten Kannada Document",International Journal of Advanced Science and Technology Vol. 48, November, 2012.

[9] M.RaviKumar, R.Pradeep,B.S.PuneethKumar,Prasad Babu"A Simple Text-line segmentation Method for Handwritten Documents", IJCA Proceedings on National Conference on Advanced Computing and Communications 2012NCACC(1):46-61, August 2012.

[10] G. Louloudi1, B. Gatos, I. Pratikakis, C. Halatsis. "Line And Word Segmentation of HandwrittenDocuments",Proceedings of the 1st International Conference on Frontiers in Handwriting Recognition (ICFHR), 247-252.

[11]M.Ravi Kumar, R. Pradeep, B.S.PuneethKumar,PrasadBabu"A Simple Text-line segmentation Method for Handwritten Documents", IJCA Proceedings on National Conferecne on Advanced Computing and Communications 2012NCACC(1):46-61, August 2012.

[12] Laurence Likforman-Sulem,Abderrazak Zahour,Bruno Taconet "Text line segmentation of historical documents: a survey**",**IJDAR (2007) 9:123–138

[13] Mamatha H R, 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, New York, USA Volume 4– No.5,October 2012 – www.ijais.org