Segmentation of Handwritten Documents Containing Kannada Script

Saleem Pasha
Assistant Professor
Department of Information Science & Engineering,
P.E.S. College of Engineering,
Mandya – 571401, Karnataka, India

M. C. Padma
Professor & Head
Department of Computer Science & Engineering,
P.E.S. College of Engineering,
Mandya – 571401, Karnataka, India

ABSTRACT
Segmentation is one of the important phases of Optical Character Recognition (OCR) system, which extracts objects of interest from an image. Feature extraction and classification phases of OCR will be more effective, if the techniques selected for segmentation is effective. This paper focuses on to develop a system for handwritten documents containing Kannada script and proposes suitable techniques to perform preprocessing and also segmentation such as line, word and character segmentation. Novelty is achieved by proposing a modified horizontal projection profile method for line segmentation, in which well separated lines and overlapping lines are detected. An average accuracy of 97.5% is achieved for line segmentation and word segmentation.

General Terms
Digital Image Processing.

Keywords
Segmentation, Optical Character Recognition (OCR), modified horizontal projection profile.

1. INTRODUCTION
Segmentation is the process of partitioning the preprocessed image into multiple segments. Using suitable preprocessing techniques, the input image is converted into preprocessed image. Segmentation techniques are applied to the preprocessed image to separate the text lines, to separate the words in each line and to separate the individual characters in each word. As preprocessing, segmentation, feature extraction and classification are the important phases in the development of OCR system, the results of the later phases (preprocessing, segmentation) will affect the performance of the subsequent phases (feature extraction, classification). Hence, in order to develop an OCR system, preprocessing and segmentation play a vital role.

Both preprocessing and segmentation techniques are applied on handwritten documents. The factors such as large character set, mood and variation in writing style of each individual, length of characters, pen quality, aged documents, paper quality, ink color and so on make the handwritten Kannada text more complex. Hence, the challenge is open for handwritten text compared to printed text. In this paper, suitable techniques are proposed to implement preprocessing. Also the paper proposes a three level segmentation applied on handwritten documents containing Kannada script. Novelty is achieved in line segmentation by proposing a modified horizontal projection profile method, which detects well separated lines and overlapping lines.

Further, the paper is divided into following sections. Survey relevant to segmentation is discussed in section 2. Section 3 introduces the Kannada script. In section 4, proposed model is presented. The paper is concluded in section 5.

2. SURVEY
As this paper concentrates on the preprocessing and segmentation phases of OCR system, survey relevant to preprocessing and segmentation of handwritten documents containing Kannada script is discussed.

Mamatha et al have developed a system to perform skew detection, correction and segmentation of handwritten Kannada documents [1]. In skew detection and correction, bounding box technique is used. Line segmentation is carried using Hough transform and word segmentation is carried using contour detection technique. The limitation of this paper is that character segmentation is not considered and achieved an average segmentation rate of 70% for words due to varying inter and intra word gaps. Mamatha et al have developed a system for segmentation of handwritten Kannada documents using morphological operations and projection profiles [2]. The limitations in [1] and [2] is overlapping lines is not considered. Thungamani et al has carried out a survey of methods and strategies for segmentation of handwritten Kannada characters [3]. In this paper, horizontal projection profile is used for line segmentation and vertical projection profile is used for word and character segmentation. The constraint of the projection profile method is that this method cannot be applied for overlapping lines and characters. The challenge of the character segmentation still remains open for compound characters. Ravi Kumar et al have proposed line segmentation of handwritten documents containing Kannada and English script using clustering method [4]. In this work, they are mainly concentrating on the document with single languages either English or Kannada, but the method does not work well for the document containing the text lines which are not separated well.

Alireza Alaei et al have proposed segmentation of handwritten documents using piece-wise painting algorithm [5]. They have performed line segmentation, but word segmentation and character segmentation is not considered. Sunanda et al have proposed standard error and weighted bucket algorithm to segment handwritten Kannada text document [6]. This method is proved to be very efficient compared to the traditional methods. In [6], line segmentation of handwritten Kannada text is performed, but overlapping text lines is not considered. Soumya et al have developed preprocessing of camera captured inscriptions and segmentation of handwritten Kannada text and ancient document images [7]. In this paper, after preprocessing of the input image, the preprocessed

image is passed through segmentation and segmentation of characters is achieved on the present handwritten Kannada scripts using connected component technique and bounding box method. The limitation in [7] is if the image contains compound characters, the single compound character containing two components are separated as two separate characters.

The limitations mentioned in the above works lead to scope for further research and improvement in case of preprocessing and segmentation of handwritten documents. Hence, suitable preprocessing and segmentation techniques are discussed in this paper. In order to perform line segmentation, a modified horizontal projection profile method is proposed, which detects well separated lines and overlapping lines. Also, word segmentation and character segmentation is performed.

3. KANNADA SCRIPT

The language spoken in Karnataka is Kannada. Aksharas in Kannada were developed from Kadamba and Chalukya scripts, descendents of Brahmi. Modern Kannada has 51 base characters called Varnamala. There are 16 vowels and 35 consonants [8]. Each vowel can modify a primary consonant to form a compound character. Therefore, a compound character consists of consonant–vowel and consonant–consonant–vowel combinations.

In this paper, preprocessing and segmentation of handwritten documents containing Kannada script is considered. A standard data set does not exist for Kannada script. So, data set is constructed. A sample handwritten document image is shown in the figure 1.

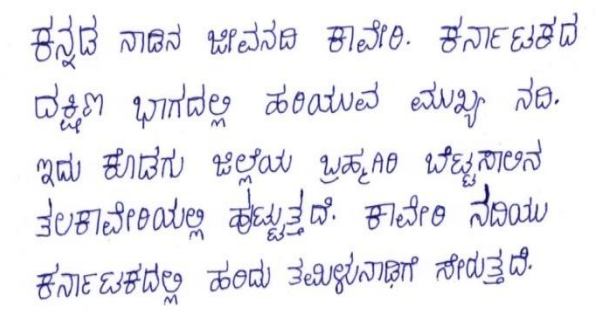

Fig 1: A Sample Handwritten Document Image

4. PROPOSED MODEL

4.1 Preprocessing

The main task of Preprocessing is to transform the input image into an image which is better suitable for feature extraction. Skew detection and correction, binarization, noise removal and morphological operations are used as preprocessing techniques. These techniques are explained below.

1. Skew detection and correction: Fourier transform technique is used to detect skews and correct skews in the image [9].
2. Binarization: Converting a gray scale image into binary image using a global thresholding approach is known as Binarization.
3. Noise removal: Median filtering is used for noise removal. This filtering technique is more effective when the goal is to simultaneously reduce noise and preserve edges.
4. Morphological operations: Morphology is a tool used for extracting the image components. This tool is useful in the representation and description of region shapes such as boundaries, skeletons and convex hull. Different morphological operations used are explained below.

1. Dilation: This operation grows or thickens the objects in an image. The dilation operation takes two pieces of data as inputs that is say A (which is to be dilated) and say B (structuring element). The main use of this structuring element is that it determines the precise effect of the dilation on the input image.
2. Erosion: This operation is the reverse of dilation operation. The objects in a binary image shrink or thin using erosion operation.
3. Close: Dilation followed by erosion is known as close operation.
4. Hit or Miss Transformation: This transformation is used as a basic tool for shape detection and is used to identify particular patterns of foreground and background pixels.

After applying the above preprocessing techniques on the input image, a preprocessed image is obtained which is given as an input to the segmentation phase of the OCR system.

4.2 Segmentation

Segmentation is a technique that is used to partition the image into multiple segments. A three level segmentation is performed using projection profile method and connected component method. Projection profile is analyzed as a data structure, which is used to store the number of non-background pixels, when the image is projected over the normal X-Y axis. Projection profile method is classified as horizontal projection and vertical projection. In line segmentation, a modified horizontal projection profile method is proposed as an attempt to detect well separated lines and overlapping lines. The three segmentations such as line segmentation, word segmentation and character segmentation are discussed below.

4.2.1 Line Segmentation

In this paper, a novel modified horizontal projection profile method is proposed to perform line segmentation. In line segmentation, an attempt is made to detect the well separated text lines and overlapped text lines in the image. Novelty is achieved in line segmentation, which consists of two parts. In the first part, an attempt is made to detect the well separated text lines and also a check is made to detect the presence of overlapping text lines in the text image. In the second part, if text lines are overlapping in the image, an attempt is also made to separate the overlapping text lines using the array concept. The line segmentation is explained in the form of steps as follows.

Steps in Line Segmentation Method:

1. In first part, Line segmentation function is applied to the preprocessed image (shown in figure 2) using novel modified horizontal projection profile method. As soon as white pixel is found, a line is drawn at that position. The position of this line is stored in an array, say ‘A’. The reason for drawing this line is to separate two text lines. This is applied for whole image, so that a line is available between

every two text lines and its position is stored in the array 'A'. To handle overlapping lines, the array 'A' is considered, which contains positions of the line drawn (shown in figure 3).

2. From the 'A', the difference between the position of second line and position of the first line is taken and stored in the first position of the array say 'B'. Again the difference between the position of third line and position of the second line of the array 'A' is taken and stored in the second position of the array 'B'. This step is repeated between every two positions in the array 'A' and result are stored in the array 'B' by incrementing its position.
3. All the values stored in the different positions of array B are added and the result is stored in a variable say 'sum'.
4. Average is calculated by taking the ratio of 'sum' by 'n'. Here 'n' is the number of positions in the array 'B'.
5. A Threshold is calculated by adding the average with 25% of the average.Threshold = average + (25% * average). From experimentation, the formula for threshold is decided.
6. The threshold obtained is compared with all the positions of the array 'B'. After comparison, one of the following decisions is selected. Decision 1: If the threshold is greater than the position of the array, it is concluded that segmented text line contains single text line (well separated lines). Decision 2: If the threshold is lesser than the position of the array, it is concluded that more than one text line is present in the result of the segmentation (overlapping lines).
7. After the execution of line segmentation function, if the result is Decision 1, each text line is cropped and saved in 'n' separate sub images. The 'n' sub images created is equal to number of text lines available (shown in figure 4).
8. After the execution of line segmentation function, if the result is Decision 2, the overlapping text line is cropped and saved in separate sub image. The sub image containing the overlapping text lines is passed through the overlapping line function (second part) to perform segmentation again.
9. In the second part of line segmentation function, the image containing overlapping lines (shown in figure 5) undergoes morphological operation such as Hit-or-Miss Transformation and close operation to convert the image into image containing lines (shown in figure 6). Lined image is passed through the morphological operation such as bwareaopen to remove the lines of pixels of size 7. A line is drawn at the last pixel of first text line, so that two text lines are segmented separately.
10. Finally, original text is extracted from the lined image along with the partitioned line in between (shown in figure 7). Each text line containing original text is cropped and saved in separate sub images (shown in figure 8).

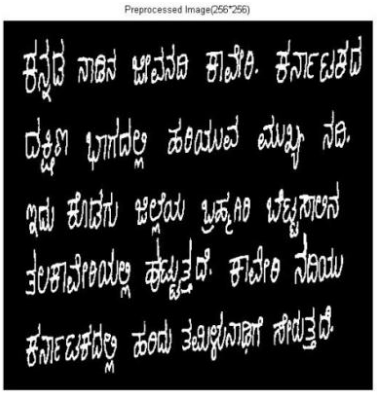

Fig 2: Preprocessed Image

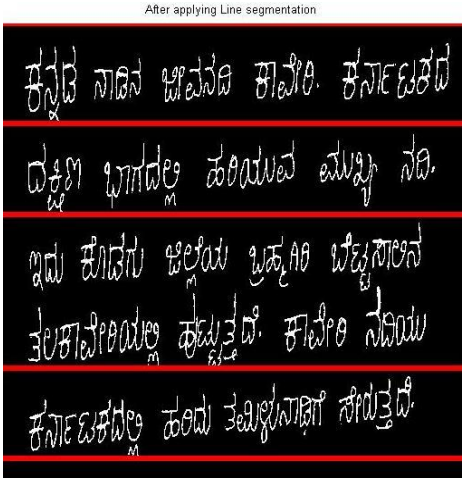

Fig 3: Results after Step 1

Fig 4: Results after Step 7

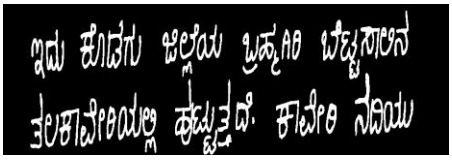

Fig 5: Results after Step 8

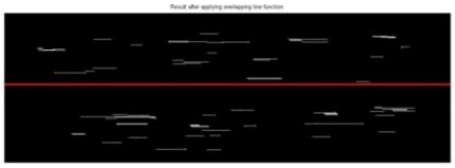

Fig 6: Results after Step 9

Fig 7: Results after Step 10

Sub image 1

Sub image 2

Fig 8: Final Results after segmentation of Fig. 6

The limitation in the overlapping line function is that after separating overlapping text lines, some pixels of the second text line may be moved to first line and vice versa as shown in the figure 7. After fine tuning, these additional pixels may be removed.

4.2.2 Word Segmentation

The result of line segmentation method is that lines are separated and stored in separate sub images. The result shown in figure 4 (one text line) is passed through word segmentation function to separate the words in the text line and these words are stored as separate sub images. Figure 9 shows the words are separated by red vertical line, after performing word segmentation.

Fig 9: Results of Word segmentation

4.2.3 Character Segmentation

The word segmentation results in separation of words and stored in separate sub images. So, this sub image contains single word and is passed through character segmentation function to separate the characters in the word. The concept of connected component method segments the characters. Whenever a character is recognized, a bounding box method is used to separate the character by applying a bounding box. A well separated character is the final result of the character segmentation. Bounding box shown in the figure 10 shows the result of character segmentation.

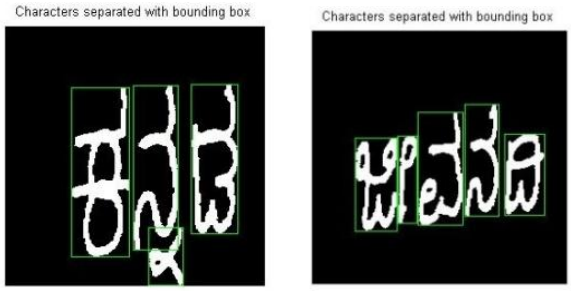

Fig 10: Results of Character segmentation

4.3 Flow Charts

The flow charts shown in figure 11, figure 12, figure 13 and figure 14 highlights the steps involved in performing complete segmentation process, modified line segmentation, separating overlapping lines and word segmentation, respectively.

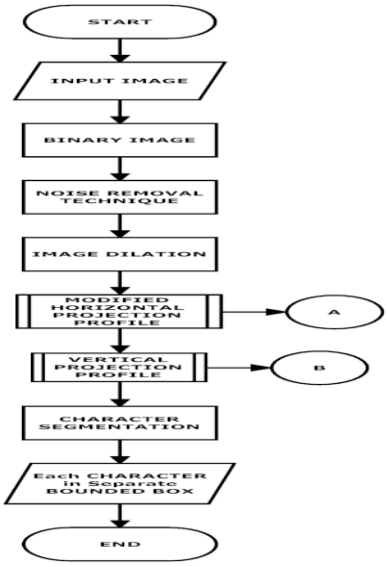

Fig 11: Flow chart of complete segmentation process

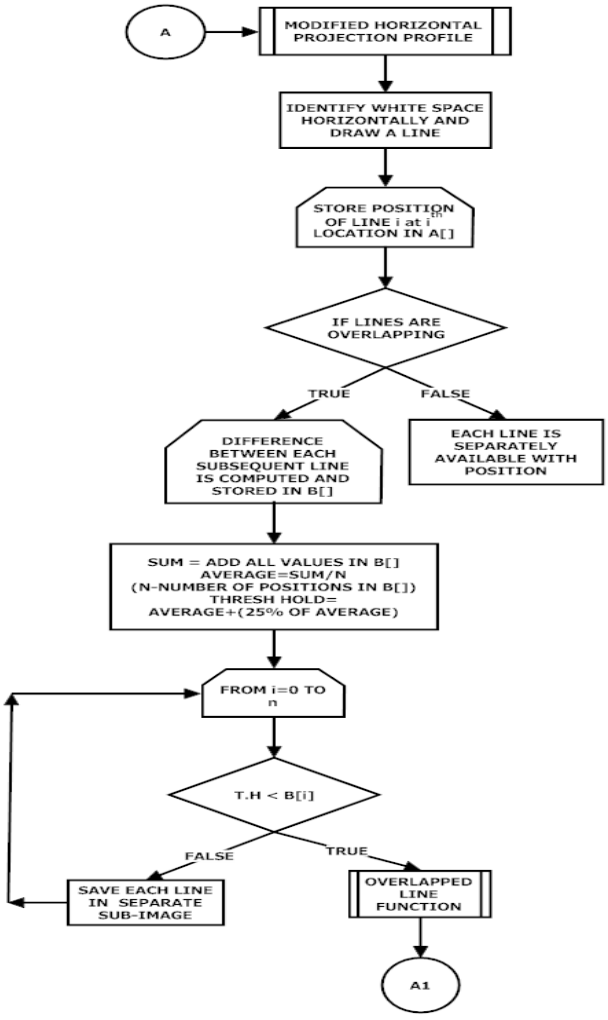

Fig 12: Flow chart showing modified horizontal projection profile

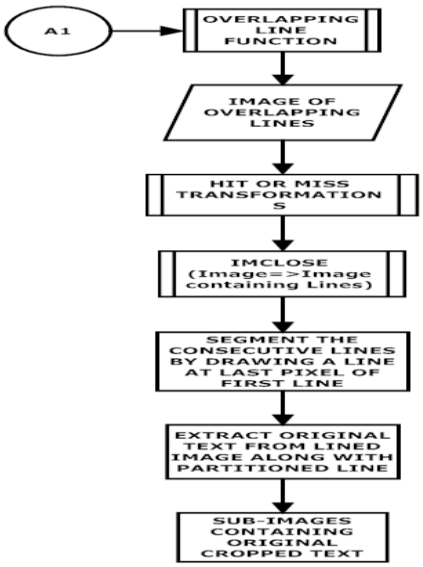

Fig 13: Flow chart showing separation of overlapping text lines

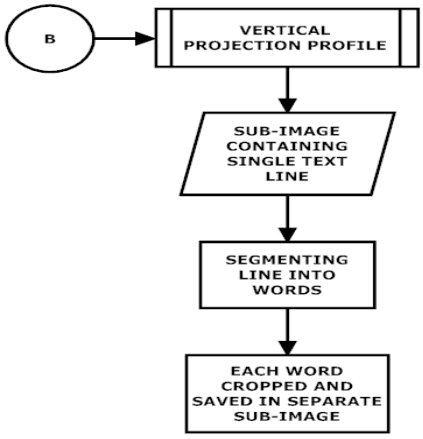

Fig 14: Flow chart showing vertical projection profile method

4.4 Experimental Results and Discussions

The active topic in OCR application and pattern classification is handwritten character recognition. The handwritten documents containing Kannada script are distinct due to different font size and style. These handwritten documents are created with no restriction on the pen, paper, ink color, ink flow, size, etc. The data set has been created for the experimentation, since a standard data set does not exist. At present, neatly handwritten documents containing Kannada script is considered. A total of 50 handwritten documents are considered for the experimentation purpose. The proposed model is implemented using Matlab in Windows7 platform. Samples of handwritten documents containing Kannada script is collected from different writers belonging to different age groups. All the sampled images were scanned with a resolution of 300dpi.

Table 1 shows the accuracy achieved for segmentation phase. An average accuracy of 97.5% is achieved for line segmentation and word segmentation. But accuracy cannot be calculated at character level because number of characters segmented is greater than the total number of characters, that is total number of characters available in 50 documents is 4300* and number of characters segmented is 4700*. The reason for this is Kannada script is composed of consonant modifiers that are combined with one of the characters to form compound characters and this compound characters occur quiet frequently in this language. Segmentation of such compound characters are difficult and need a different perspective to handle compound characters.

Table 2 shows the comparison of the proposed method with the existing methods. In method 1, only line and word segmentation is considered. Method 2 also concentrates only on the line and word segmentation. Method 3 is limited only to line segmentation. In the proposed method, we have developed a three level segmentation such as line, word and character segmentation. In line segmentation, an attempt is also made to detect well separated lines and overlapping lines.

Table 1. Accuracy Achieved in Segmentation

Number of Documents	50
Total number of lines available in 50 documents	250
Total number of lines segmented	245
Accuracy in percentage for Line segmentation	98%
Total number of Words Available in 50 documents	1150
Number of Words Segmented	1120
Accuracy of Word segmentation	97%
Total number of Characters Available in 50 documents	4300 *
Number of Characters Segmented	4700 *
Average Accuracy of Line and Word Segmentation	97.5%

Table 2. Comparison of the Proposed method with the Existing methods

Methods	**Segmentation Technique**	**Size of Data set**	**Accuracy in Percentage**
Method 1 [2]	Morphological operations, projection profile	100	94.5%
Method 2 [5]	Potential Piece-wise	204	94.98%
Method 3 [6]	Weighted bucket algorithm	80	98.12%
Proposed Method	Modified projection profile, connected component.	50	97.5%

5. CONCLUSION

Segmentation is one of the important phases of Optical Character Recognition (OCR) system. Segmentation partitions the preprocessed image into multiple segments. In this paper, suitable techniques are used to perform preprocessing and also an attempt is made to perform a three level segmentation such as line, word and character segmentation. A modified horizontal projection profile method is used for line segmentation, which detects well separated lines and overlapping lines. An average accuracy of 97.5% is achieved for line segmentation and word segmentation. Complete accuracy is not achieved at the character level.

6. REFERENCES

[1] Mamatha Hosalli Ramappa and Srikantamurthy Krishnamurthy. 2012. Skew Detection, Correction and Segmentation of Handwritten Kannada Document. International Journal of Advanced Science and Technology, 72 – 88.

[2] Mamatha H R and Srikantamurthy K. 2012. Morphological Operations and Projection Profiles based Segmentation of Handwritten Kannada Document. International Journal of Applied Information Systems, 13–19.

[3] M. Thungamani and P. Ramakhanth Kumar. 2012. A Survey of Methods and Strategies in Handwritten Kannada Character Segmentation. International Journal of Science and Research, 18–23.

[4] M.Ravi Kumar, Nayana N Shetty and B.P.Pragathi. 2012. Text Line Segmentation of Handwritten Documents using Clustering Method based on Thresholding Approach. International Journal of Computer Applications, 9 – 12.

[5] Alireza Alaei, P. Nagabhushan and Umapada Pal. 2011. A Benchmark Kannada Handwritten Document Dataset and its Segmentation. International Conference on Document Analysis and Recognition, 141 – 145.

[6] A. Sunanda Dixit, B. S.Ranjitha and Dr .H. N. Suresh. Segmentation of handwritten Kannada text document through computation of standard error and weighted bucket algorithm. International Journal of Advanced Computer Technology, 55 – 62.

[7] Soumya A and G Hemantha Kumar. 2014. Preprocessing of Camera Captured Inscriptions and Segmentation of Handwritten Kannada text. International Journal of Advanced Research in Computer and Communication Engineering, 6794 – 6803.

[8] Indira K and Selvi S. S. 2009. Kannada character recognition system: a review. InterJRI Sci Technol, 31–42.

[9] Postl W. 1986. Detection of liner oblique structure and skew scan in digitized documents. In: proceeding of international conference on pattern recognition, 687–689.