# Unconstrained Handwritten Text Line Segmentation for Kannada Language

**Shakunthala B. S,C S Pillai**

***Abstract:*** ***Segmentation is division of something into smaller parts and one of the Component of character recognition system. Separation of characters, words and lines are done in Segmentation from text documents. character recognition is a process which allows computers to recognize written or printed characters such as numbers or letters and to change them into a form that the computer can use. the accuracy of OCR system is done by taking the output of an OCR run for an image and comparing it to the original version of the same text. The main aim of this paper is to find out the various text line segmentations are Projection profiles, Weighted Bucket Method. Proposed method is horizontal projection profile and connected component method on Handwritten Kannada language. These methods are used for experimentation and finally comparing their accuracy and results.***

***Keywords : Projection profiles, Weighted Bucket Method. horizontal projection profile and connected component method, Segmentation, Preprocessing***

## I. INTRODUCTION

A document is considered as a structure it contains information. the document is not in a proper structure it is very difficult to get back the information. Document structure is a essential stage in character recognition. Handwritten Kannada documents main challenges are 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 using the following methods. Projection profiles, Weighted Bucket Method. Proposed method is horizontal projection profile and connected component method Then these methods are applied on Handwritten Kannada documents.

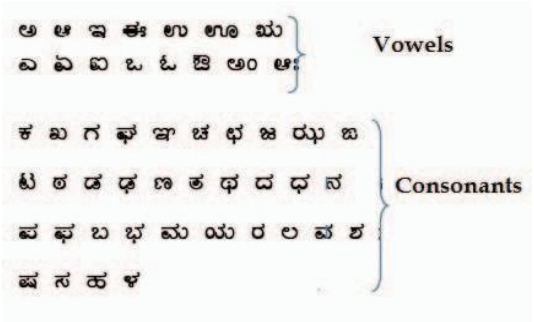


**Figure 1: Kannada language 49 phonemic letters**

## II. KANNADA SCRIPT

Kannada is a Dravidian language mainly used by the people of Karnataka, Andhra Pradesh, Tamil Nadu and Maharashtra. Kannada is spoken by about 44 million people. The language has 49 characters in its alphabet set (15 vowels and 34 consonants). This gives total of (544*34) +15=18511 distinct characters, samples of extra modifiers shown in the Figure 3.

ಕ ಕಾ ಕಿ ಕೀ ಕು ಕೂ ಕೃ ಕೄ ಕೆ ಕೇ ಕೈ ಕೊ ಕೋ ಕೌ ಕಂ ಕಃ

**Figure 2: sample of Kannada modifier glyphs(Diacritics)**

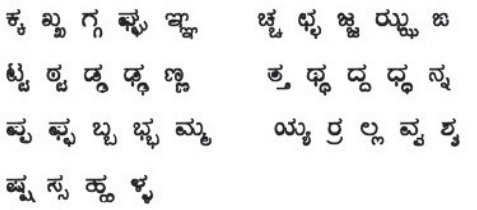

**Figure 3: Consonant conjuncts in Kannada (vattakshara)**

## III. DIFFICULTIES IN TEXT LINE SEGMENTATION

There are three difficulties in segmentation of text line such as text line components, influence of author style, influence of poor image quality.

### *A. Text line components*

**Baseline**: Imaginary lines are connected below the character bodies of Text consisting of a row of words written across a page as shown in figure 4.

**Revised Manuscript Received on July 22, 2019.**

***** Correspondence Author

**Shakunthala B S***, Assistant Professor, Department of Information Science & Engineering, Kalpataru Institute of Technology, Tiptur, Karnataka, India. Email: shakukit@gmail.com.

**Dr. C S Pillai**, Professor, Department of Computer Science & Engineering, ACS college of engineering, Bangalore, Karnataka, India. Email: pillai.cs5@gmail.com

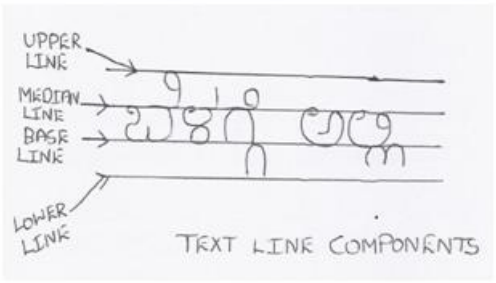


**Fig. 4 Text Line Components**

**Median line**: Imaginary lines are connected upper the character bodies of a text line.

**Upper line** : ascenders are connected by an above the imaginary line.

**Lower line** : descenders are connected by below the imaginary line.

**Overlapping letterss**: Overlapping letters are present in the region of next line may be decreasers and incresers.

**Touching components**: The touching components are comes in the region of one after the other lines.

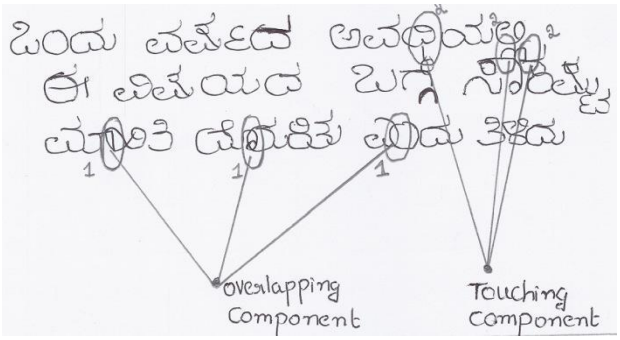


**Fig. 5 Overlapping and touching components**

***B. The effect of author style***

**Prominent line variation** : The Prominent line is straight or curved. Prominent line will be varied depending on the writer movement.

**Line orientations** : lines will be positioned in various directions and angles.

**Line spacing :** Lines spacing is easy to find. The lower well-known line of the first line touches with the upper well-known line of the second line.

***C. The effect of poor image quality***

Smudges and seeping ink present in other side of the image in the document produce binarisation errors as shown in Fig.6 & Fig.7. Stroke fragmentation and merging as shown in Fig 8.

Due to the presence of punctuation, dots and broken strokes makes the quality of the images to be low.

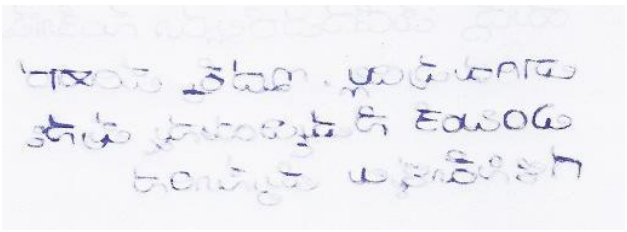

**Fig. 6 Presence of seeping ink from other side of the Document**

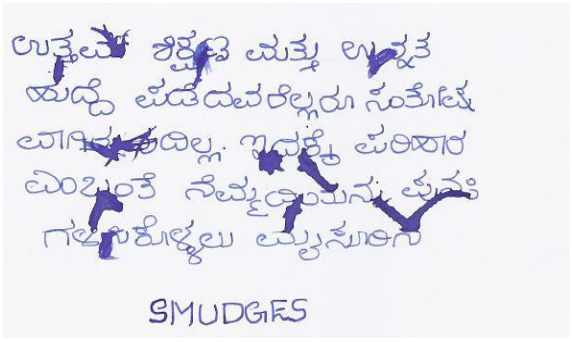


**Fig. 7 Smudges**

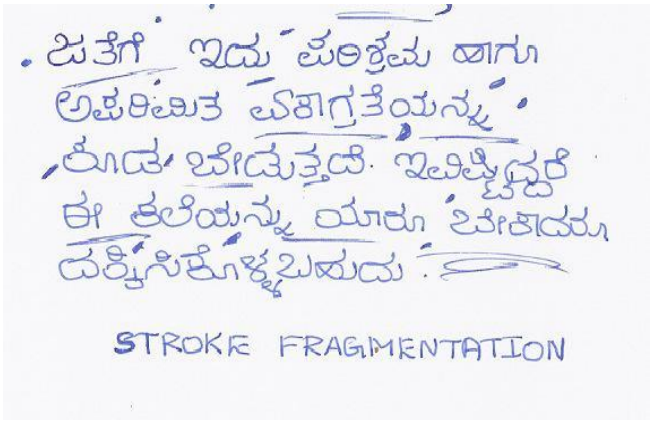


**Fig. 8 Stroke Fragmentation**

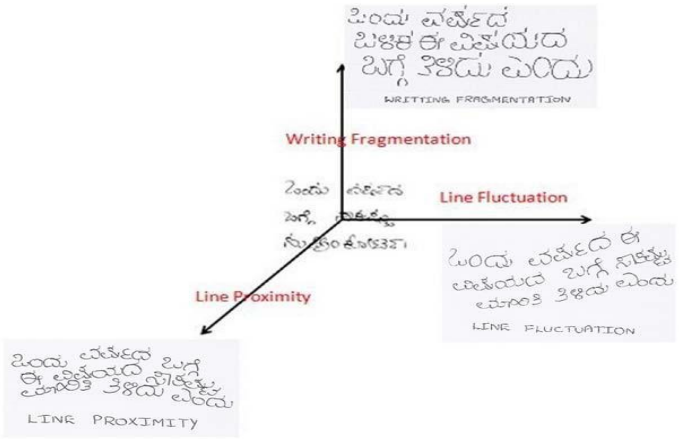


**Fig. 9 The three main axes of document difficulties for text line segmentation**

***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. 10 (pronounced as ramanu).Case (ii) A word with subscript as shown in Fig. 11(pronounced as prashastavaagi).

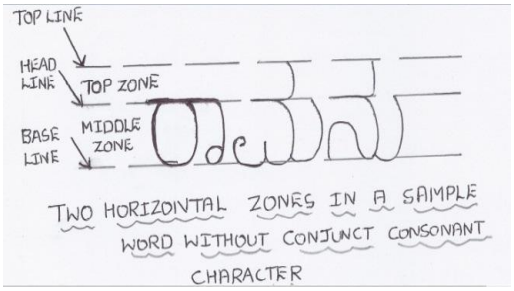


**Fig. 10 A word without subscript**

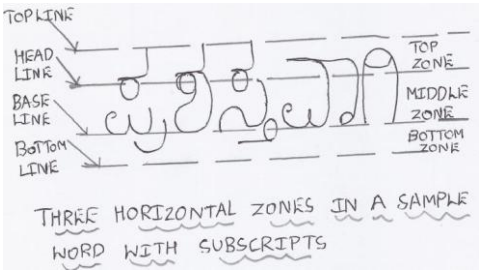


**Fig. 11 A word with subscript**

## IV. PREPROCESSING

- The main aim of Preprocessing is to transform the input image into an image.
- It contains feature extraction. Skew detection and correction, binarization, noise removal and morphological operations.

**Skew detection and correction**: Fourier transform method is used to detect skews and correct skews in the image [9].

**Binarization**: Global thresholding method is used to converting a gray scale into binary image.

**Noise removal**: Median filtering method is used for noise removal.

Advantages 1) to reduce noise
2) and preserve edges.

**Morphological operations**: Morphology is a technique for extracting the image components.

**Advantage** 1) Representation and description of region shapes.

**Different morphological operations are**

**Dilation**: This operation grows the objects in an image.

- Dilation is represented by $\oplus$. it is used for structuring element for probing and expanding the shapes contained in the input image

**Erosion**: This operation is the reverse of dilation operation. It removes pixels on object boundaries or The objects in a binary image shrink or thin using erosion operation.

**Close:** Dilation followed by erosion is known as close operation.

**Hit or Miss Transformation**: This is used for shape detection and is used to identify particular patterns of foreground and background pixels.

After applying the preprocessing techniques on the input image to get preprocessed image.Preprocessed image is given as an input to the segmentation phase of the OCR system.

## V. TEXT LINE SEGMENTATION METHODS

dividing an image into multiple parts is called segmentation.

The main aim is an altering of an image in to multiple segments .

It is more meaningful and easier to deriving. In this stage various segmentation techniques are discussed.

### *A. Projection Profile*

A projection profile is a technique giving the number of ON pixels gather together along parallel lines. Mainly Horizontal projection profile is used for text line segmentation.

**It contains two steps**

- •Pre-processing with morphological operations
- •Text line extraction.

**Algorithm for Projection profile**

Step 1:Binarise the original image.
Step 2:Morphological ( Dilation and erosion) operations are applied.
Step 3: when segment falls, the row number can be obtained.
Step 4: continuous white space and once a black pixel is found. In this case row number is marked as starting index of a line,
Step 5:the row number is marked as Ending index of a line, When there is a continuous white space is found.
Step 6:Height of the segment can be obtained using starting and ending index of a line.
Step 7:Crop the image using height and save it in a separate file.

**Advantages**

- Projection Profile gives the best segmentation rate compare to other proposed methods
- this method works well for clearly separated lines
- this method cannot divide the touching or overlapping lines and instead it will merge those lines.

**Disadvantages**

- Projection Profile cannot divide the touching or overlapping lines.

**Advantages**

- To Increase the strength of the histogram.

### *B.* Weighted Bucket Method

**Algorithm**

Step 1:Scanned input document.
Step 2:Document preprocessing.
Step 3:Find connected component.
Step4: Remove destructor from image.
Step 5:Normalizing of components using standard error.
Step 6: Grouping of component syllable.
Step 7:Create an array of empty bucket.
Step8:Select the component syllable with highest weight and link to the bucket.
Step 9:Go over the original list of component syllabus, putting each syllable in the bucket that falls within the range.
Step10: Repeat step7 and 8 till all the component syllables are segregated into buckets.
Step 11: Delete any empty buckets left.
Step 12: Sort each bucket to arrange the syllable from left to right.

## VI. PROPOSED METHOD

**Modified horizontal projection profile and connected component method**

It 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.

**In first part**:

step 1:Line segmentation function is applied to the preprocessed image 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 betweenevery 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.

Step 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.

Step 3. All the values stored in the different positions of array B are added and the result is stored in a variable say 'sum'.

Step 4. Average is calculated by taking the ratio of 'sum' by 'n'. Here 'n' is the number of positions in the array 'B'.

Step 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.

Step 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).

Step 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.

Step 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.

**In the second part** of line segmentation function, the image containing overlapping undergoes morphological operation such as Hit-Miss Transformation and close operation to convert the image into image containing lines. Lined image is passed through the morphological operation 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.

Finally, original text is extracted from the lined image along with the partitioned line. Each text line containing original text is cropped and saved in separate sub images.

## VII. RESULTS

the results of proposed method for line segmentation**.**

| Author | Segmentation Method | Size of Dataset | Segmentation rate |
|---|---|---|---|
| Proposed method | Modified horizontal projection profile and connected component method | 100 | 97.5% |

## VIII. CONCLUSION

In this paper, we have proposed a Modified horizontal projection profile and connected component method for text line segmentation of Kannada handwritten documents. The method was tested on totally unconstrained handwritten Kannada documents. An average segmentation rate of 97.5%.

## REFERENCES

1. 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

2. 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.

3. 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

## AUTHORS PROFILE

**SHAKUNTHALA B S** Currently working as Asst.Professor in the Department of ISE at Kalpataru Institute of Technology, pursuing Ph.D degree in Computer Science and Engineering from VTU, Karnataka , India..

**Dr. C S Pillai** Currently working as Professor in department of CSE at ACS college of engineering, Bangalore, Karnataka , India.