A Survey of Text Line Segmentation Methods for South Indian Languages

Shreeharsha B N
Department of Computer Science and Engineering
The National Institute of Engineering,
Mysuru, India

Ravindra R[1]
Department of Computer Science and Engineering
The National Institute of Engineering,
Mysuru, India

Padmini M S
Assistant Professor,
Department of Computer Science and Engineering,
The National Institute of Engineering,
Mysuru, India

Srinidhi S H
Department of Computer Science and Engineering
The National Institute of Engineering,
Mysuru, India

***Abstract*—Converting a handwritten script into a digitalized image is one of the major tasks involved in Optical Character Recognition (OCR) process. Line Segmentation is considered as the basic pillar of character recognition system as it deals with identifying text lines and distinguishes them from unwanted objects. The accuracy of the character recognition system depends upon the complexity and efficiency of the segmentation algorithm used. Despite some successful methods developed for OCR, they fail to be a better segmentation method for scripts of Indian languages. Skewness in lines, overlapped lines and characters, variable inter and intra-word gaps, different writing styles pose a major problem for developing a segmentation method with high accuracy. This paper aims at investigating different segmentation techniques like Projection Profile, Normalized Chain Codeand Seam Carving method for handwritten documents. These methods are subjected to tests with different datasets of varying complexity. Experimental results of the above mentioned methods are briefly discussed in this paper.**

Keywords*—Handwritten Documents, Line Segmentation, Optical Character Recognition, South Indian Languages.

I. INTRODUCTION

Optical Character Recognition and Document Image Analysis are the fields which are under continuous research for several years. OCR plays an important role in transforming a document written on paper media into text on digital form which can be searched and analysed.

The process of OCR is shown in the Fig. 1. Major phases in Optical Character Recognition are,

1. Image pre processing
2. Segmentation
3. Extraction
4. Recognition

As shown in the Figure 1, Document is placed over the Scanner for Scanning where it is converted into any one of the digital image formats like JPG, JPEGand PNG. The pre-processing step involves binarization [1], which is the process of converting a grayscale image with 0-255 pixel values into a binary image with 0 & 1 pixel values by thresholding. Majority of the noise will be removed and also it helps to save disk space as binarized image requires less space to store. The size of the character is reduced by Thinning in which Skeletoning of the image is done by different techniques such as Otsu's [2] method. In the next phase document is subjected to line segmentation in which text lines are identified. The segmentation process further continues with word and character segmentation. Segmented characters are extracted and the recognition phase recognises the extracted characters thus converting the entire document into a digitalised image in which information can be analysed. [3]

Segmentation is a process of extracting objects of interest from an image. The first step in segmentation process is detecting the lines. The subsequent steps are detecting the words in every line and detecting the characters in each word.

Individual lines in a document are detected during Line Segmentation process. Hence Line Segmentation is a crucial task in OCR process as the final output depends on the initial detection efficiency.

This paper is structured as follows: Section II presents various challenges that become a hurdle for developing a better segmentation method. A comparative study of some improved and efficient segmentation algorithms is done in Section III. The dataset used for the experiment and analysis of the results is presented in Section IV. The conclusion and summarization of the survey is described in Section V.

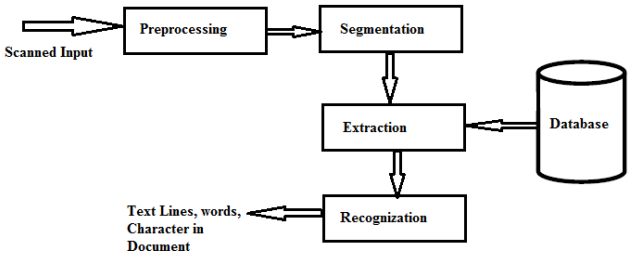


Fig. 1 Process of OCR.

II. CHALLENGES

In thissection, some of the challenges involved in developing segmentationalgorithms for Indian languages are discussed. Major problems among them are:

A. *Overlapping and touching components*

Overlapped lines are descenders and ascenders located in adjacent lines. Touching lines are ascenders and descenders belonging to consecutive lines which are connected. These components are very hard to differentiate before identifying the text lines as described in [4].

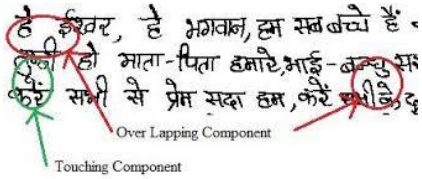


Fig. 2 Touching and Overlapping Lines.

B. *Influence of author's writing style*

It is always seen that handwritten document does not contain even style of writing. Line spacing, line orientation are not always unique. While writing, author unknowingly may provide more line spacing at some place in text while lesser at some place. Also there can be different line orientations especially when there are annotations or corrections. Inter and intra-word gaps will also vary from one author to another.

C. *Impact of document image with poor quality*

Smudges and ink dots present in the other side of the input image produce binarization error as shown in the Fig. 3 Smudges means spreading of the ink on the periphery of the surrounding pixels as shown in the Fig. 4. Also the poor scanning of document, poor paper quality and variable intensity of background causes errors in binarizing the image.

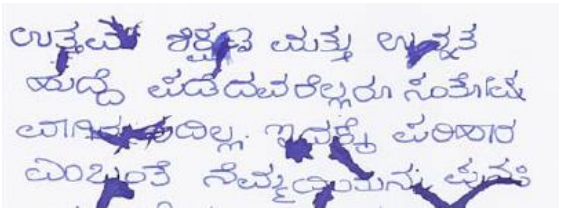


Fig. 3 Smudges.

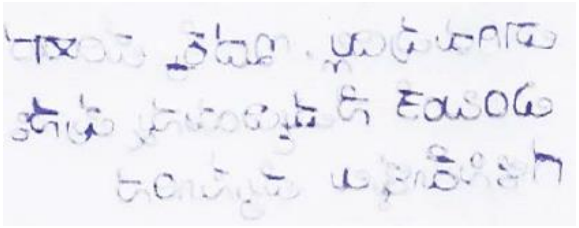


Fig. 4 Seeping Ink.

D. *Complexity involved with script of the languages*

South Indian languages like Kannada, Tamil, Telugu and Malayalam pose a major problem in segmentation because of complex nature of their script. When compared to English script, South Indian languages contain certain number of characters which include vowels and consonants. Scripts are composed of characters and in conjunction with consonant modifiers which form a meaningful word. [5]

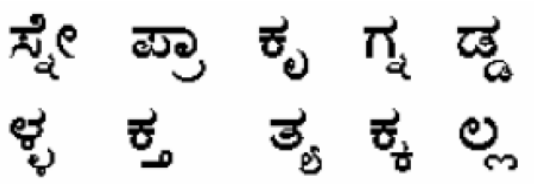


Fig. 4 Conjunct Consonants.

E. *Other factors*

1) *Skewed lines:*

Document skew has been recognized as a universal problem of document imaging. Hand placement or paper placement normally cause a skew of 2-3˚. In some cases skewness may extend up to 10˚. Lines with a skew more than 2-3˚will reduce the accuracy of OCR. If the skewness crosses a limit of 5˚, the result becomes unreliable according to [6].

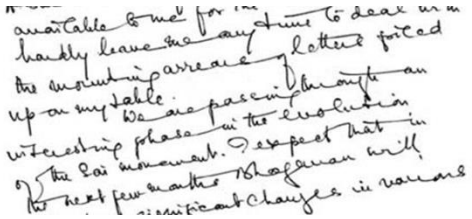


Fig. 6 Skewed Lines

2) *Fluctuating lines:*

Lines of text are partially or fully connected to other text lines and lines in which words are placed unevenly from the base line are called fluctuating lines (Fig. 7). Fluctuation from the base line will mislead the segmentation algorithm.

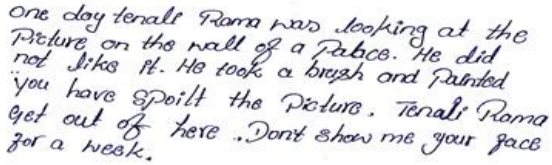


Fig. 7 Fluctuating Lines

III. METHODOLOGY

Several methods have been developed for segmentation process. Many of them require certain constraints and limitations on input documents. Some methods will work in real time situations with less constraints on input documents. Three methods among those are discussed in this section.

A. Horizontal Projection Profile

A projection profile [7] is a histogram giving the number of ON pixels accumulated along parallel lines. In a two dimensional representing plane, two types of projection profile namely horizontal projection profile and vertical projection profile are defined. A horizontal projection profile is a one-dimensional array where each element denotes the number of ON pixels along a row in the image. Similarly a vertical projection profile gives the pixel sum of columns.

To segment the document image into several text lines, we use the valleys of the horizontal projection computed by a row-wise sum of black pixels. The position between two consecutive horizontal projections where the histogram height is least denotes one boundary line. Using these boundary lines, document image is segmented into several text lines.

Algorithm:

Input: Scanned image of handwritten document.
Output: Identified text line segments.
Step 1: Scanning the input document
Step 2: Binarization – The process of separating foreground and background information based on obtained threshold value.
Step 3: Noise reduction – Removal of image noise generated by degradation due to aging, photocopying or during image capture.
Step 4: Thinning – Skeletoning the image
Step 5: Generation of Histogram to get threshold value.
Step 6: Line Segmentation using the data of pixel intensity of the image obtained by histogram.

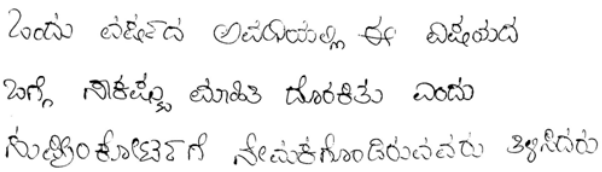


Fig. 8 Original Image

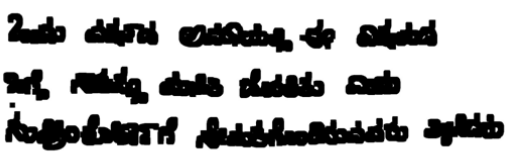


Fig. 9 Binarized Image

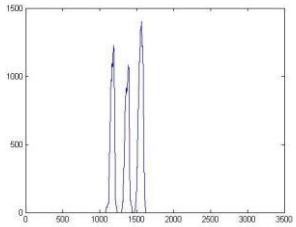


Fig. 10 Horizontal Projection Profile

B. Normalised Chain Code

Algorithm:
Step 1: Input a document image D
Step 2: Preprocess the document image D to obtain enhanced image EI
Step 3: Thin the image EI
Step 4: Convert the image into binary image BI.
Step 5: partition the image BI into n number of vertical segments VS_i of BI of size n, where n is the number of columns in each segmented partition.
Step 6: Compute the horizontal projection profile $H_k(VS_i)$ of the each vertical segmented block VS_i, where i=1, 2, 3…n and k=1, 2, 3…
Step 7: For each $H_k(VS_i)$ mark the row indices in the regions of minimum valleys and store it in array A_s.
Step 8: Obtain the combined array A, where A = A1 U A2 U A3 U…..As and size of (A)= M x N Where M is max number of rows in array $A_i$and N is number of columns equal to total Number of arrays A_i.
Step 9: For each distinct value in combined array A calculate the count of repetition in the A and Obtain a distinct marked row array DA of size M x 2 where column 1conatin each distinct row in all Ai and column 2 contains count of each distinct row in column 1.
Step 10: Obtained Normalized Distinct Array (NDA) from Distinct Array (DA).
Step 11: For each value v in NDA, compute the path of traversal for identifying the segmentation curve between the text lines in the document based on the back ground intensity of document.
Step 12: If any black pixel is encountered in the path, then calculate the 8 directional chain code from that black pixel till a specified threshold.
Step 13: if chain code sequence belongs to 08888 or 07777 or 03333 or 06666 or 04444 or 02222 then consider the encountered black pixel in the path as segmentation path else move to the next row in the NDA.
Step 14: Continue this process till all the rows in the NDA
Step 15: Stop.

C. Seam Carving Method

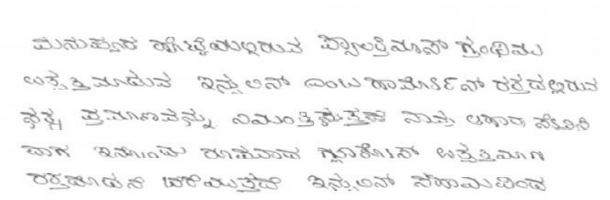
Seam Carving method [8] uses a two-stage procedure to extract text lines from a grayscale image. First, seam carving is used to generate the medial seams of the handwritten page. Geodesic distance transform is given as the input to the optimization procedure, in which each pixel's

value is its shortest path length to the nearest background pixel.

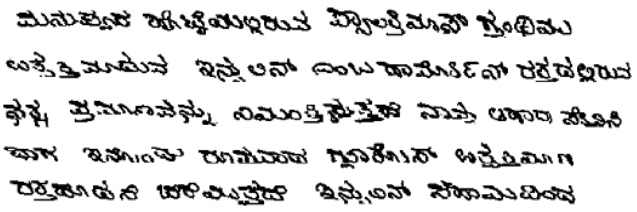
In the second step, seam seeds are generated and a greedy algorithm is applied, these seeds are propagated to generate two separating seams: one above the medial seam and one below the medial seam. These separating seams define the upper and lower boundaries of the text line.the use of seam carving for medial seam computation can result in seams that jump over neighboring lines, particularly in cases where the gaps between words are much large compared to the distance between two consecutive text lines. The histogram matching approach, however, is more robust, because it avoids jumping over neighboring lines with considering the multiple orientation of the text.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

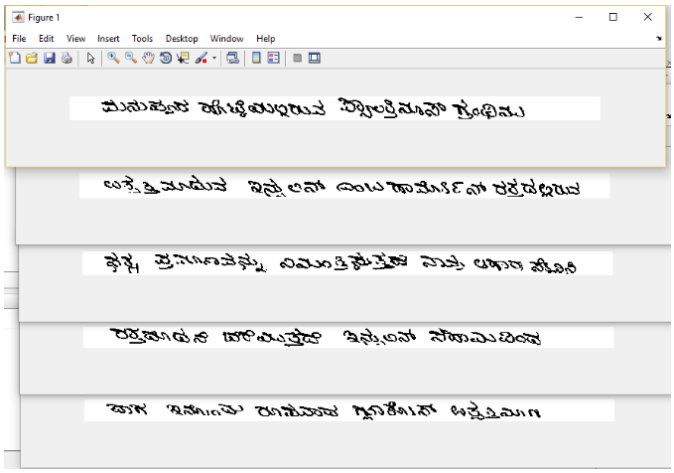
This chapter presents the results obtained using the segmentation methods described in the methodology section of this paper. The dataset is collected by author from different individuals belonging to different categories like age, educational background in a separate unruled A4 sheet without any restrictions. The participants are given to write text pages by different types of pens. The documents which are collected from different individuals are then scanned in gray-scales. We have considered 100 documents from the dataset taking 25 documents from each category for experimentation. A comparison of the accuracy of the results is tabulated in table 1.



(a)

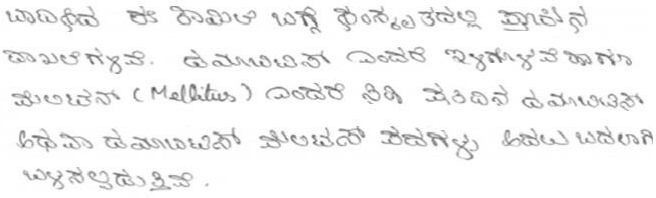


(b)

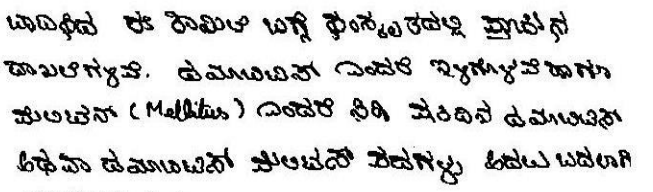


(c)

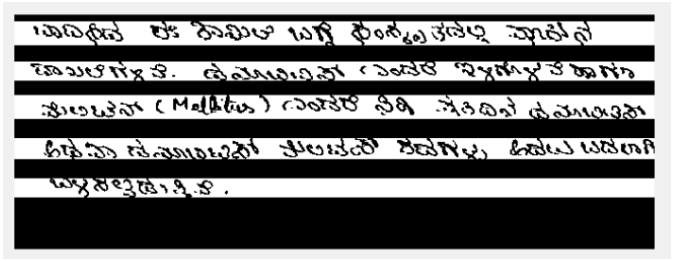
Fig. 11 Horizontal Projection Profile Method - Original Input image (a), Binarized Image (b), Images containing segmented text lines (c).



(a)

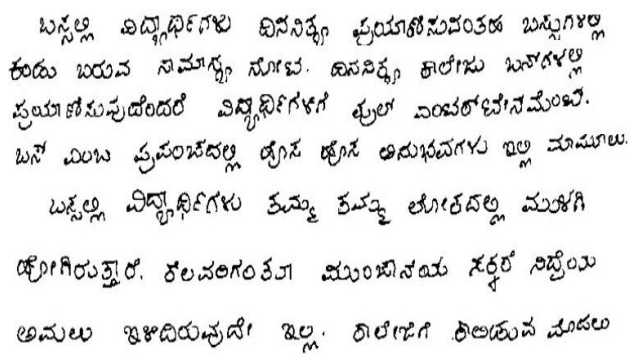


(b)

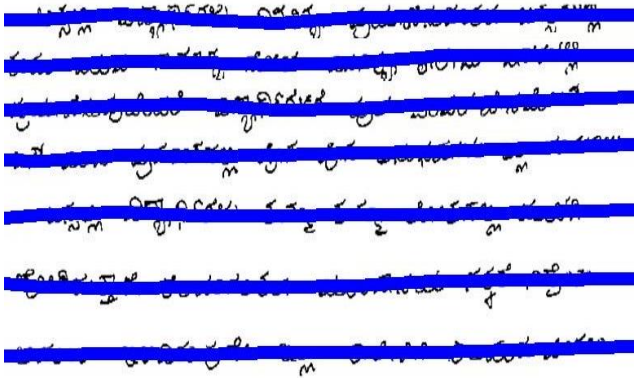


(c)

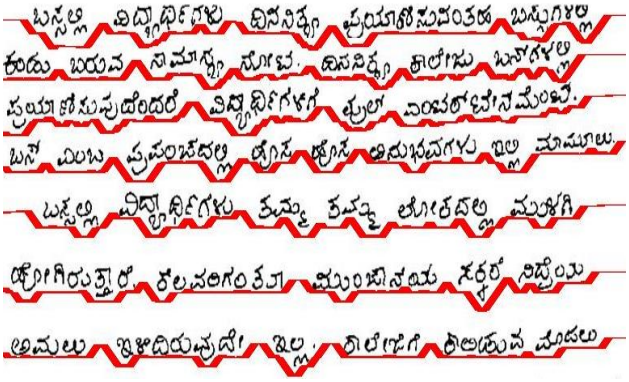
Fig. 12 Normalized Chain Code Method – Original Input Image (a), Binarized Image (b), Image with Identified Line Segments (c)



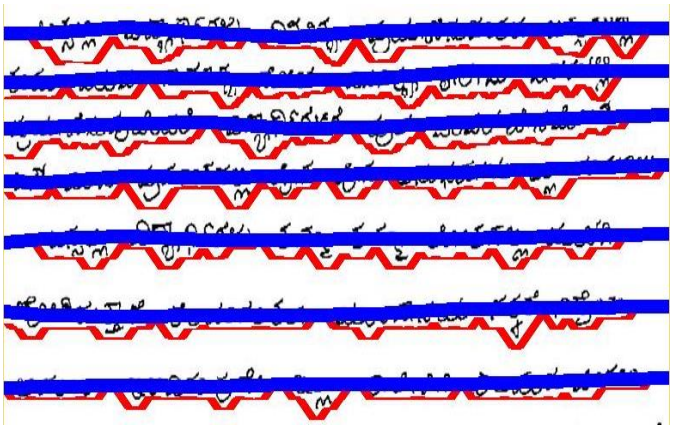
(a)



(b)



(c)



(d)

Fig. 13 Seam Carving Method – Original Input Image (a), Text Line Approximations (b), Minimum energy seams in line segments (c), Identified text line boundaries (d).

TABLE I
PERFORMANCE EVALUATION OF SEGMENTATION METHODS DISCUSSED IN THIS PAPER

Segmentation method	Size of dataset	Segmentation rate
Horizontal Projection Profile based method	100	83.47%
Normalized Chain Code	100	87.82%
Seam Carving Method	100	95.32%

V. CONCLUSION

The complexity involved with the nature of scripts of South Indian languages and other challenges discussed in this paper are the main hurdles for developing an Optical Character Recognition System for South Indian languages. An attempt is made in this direction and extraction of lines is done considering an input dataset of different handwriting styles. Accuracy obtained from the methods discussed in this paper are reduced because we have considered documents with different handwriting styles. These segmentation methods provide good segmentation rate for documents with good handwriting style with less line skewness.

Among several segmentation methods developed, three improved and efficient methods are discussed under the methodology section. Seam Carving method can be applied without binarizing the document. Compared to other two methods discussed, Seam Carving method yields a high accuracy with a segmentation rate of 95.32%.

ACKNOWLEDGMENT

The authors of this paper would like to thank everyone who contributed for providing to prepare the dataset to our survey.

REFERENCES

[1] Chien-Hsing Chou, Wen-Hsiung Lin, Fu Chang, *A binarization method with learning built rules for document images produced by cameras.*

[2] Miss Hetal J. Vala, Prof. Astha Baxi , *A Review on Otsu Image Segmentation Algorithm, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2, Issue 2, February 2013*

[3] G. Louloudisa, B.Gatosb, I.Pratikakisb, C.Halatsisa, *Text lineand word segmentation of handwritten documents2009 Elsevier Ltd*
[4] 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*
[5] Mamatha H R, Srikantamurthy K*Morphological Operations and Projection Profiles based Segmentation of Handwritten Kannada DocumentInternational Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868*
[6] Prasad Babu, R. Pradeep, B.S.Puneeth Kumar, M.Ravi Kumar*A Simple Text-line segmentation Method for Handwritten Documents International Journal of Computer Applications (0975 – 8878)*
[7] M. Thungamani and P. Ramakhanth Kumar,*A Survey of Methods and Strategies in Handwritten Kannada Line Segmentation, International Journal of Science Research Volume 01, Issue 01, 2011*
[8] Nikolaos Arvanitopoulos ,Sabine Susstrunk*Seam Carving for Text Line Extraction on Color and Grayscale Historical Manuscripts, 2014 14th International Conference on Frontiers in Handwriting Recognition*