JCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-6, Issue-4 E-ISSN: 2347-2693

A Comparative Study of Segmentation Techniques used in Handwritten Documents

Ms. S. A. Bhopi[1*], M. P. Singh[2]

[1]Department of Computer Science, MGM's College of CS & IT, SRTMU, Nanded, Maharashtra (India)
[2]Department Of Computer Science, IET, Dr. B. R. Ambedkar University, Agra, U.P. (India)

Corresponding Author: smitabhopi@gmail.com, Tel.: +91-9881525159

Available online at: www.ijcseonline.org

Received: 30/Mar/2018, Revised: 09/Apr/2018, Accepted: 25/Apr/2018, Published: 30/Apr/2018

***Abstract*—** Handwritten document image segmentation is key step for OCR (Optical Character Recognition) System. It is an important step because inaccurately segmented text lines will cause errors in the recognition stage. The selection of segmentation algorithm being used is the essential factor in deciding the accuracy of the OCR system. Devnagari is the most popular script in India. Devnagari is the script for Sanskrit, Hindi, Marathi, Kashmiri, Sindhi, Bihari, Bhili, Konkani, Bhojpuri and Nepali languages. It has vowels, consonants, vowel modifiers and compound characters, numerals. Optical Character Recognition for Devanagari is highly complex due to its rich set of conjuncts. The nature of handwriting makes the process of text line segmentation very challenging. Several techniques to segment handwriting text line have been proposed in the past. Our purpose is to provide a learning-based approach for segmentation of handwritten document images. This paper presents a quantitative comparison of three algorithms for page segmentation: Projection Profile, Run-length Smearing and Bounding Box along with some morphological operations like erosion, dilation etc. We have implemented these algorithms on our own dataset of handwritten documents. We have experimented and compare the accuracy and results of these methods.

Keywords — OCR, Line and Word Segmentation, Projection Profile, Bounding Box, Run length Smearing

I. INTRODUCTION

In this paper, we compare previous work done on text line segmentation in handwritten documents. There are two types of segmentation bottom-up and top-down. The bottom-up approach use the connected components based methods merge neighboring connected components using simple rules on the geometric relationship between neighboring blocks. Where as in the top-down algorithms projection based methods is used .which is one of the most successful method for machine printed documents since the gap between two neighboring text lines in machine printed documents is typically significant, thus the text lines are easily separable. The projection based methods cannot be directly used in handwritten documents, unless gaps between lines are significant or handwritten lines are straight.

In the segmentation is an image of handwritten document image is decomposed into sub images of lines, words and characters. It is one of the essential steps in an optical character recognition (OCR) system. It is an important step because inaccurately segmented text lines will cause errors in the recognition stage. It makes a major contribution to the error rate of the system.

The rest of the paper is organized as follows. Section II explains the challenges in text line segmentation of handwritten documents. Section III describes some segmentation methods. Section IV provides the related work done in the area of handwritten document image segmentation. Section V presents an extensive performance evaluation and quantitative comparison. Section VI will have the concluding remarks of the study we have done.

II. CHALANES IN SEGMENTATION OF HANWRITTEN DOCUMENT

The variation in the handwriting of each person in handwritten documents makes the segmentation procedure a challenging task. There are many problems encountered in the segmentation procedure of handwritten documents due to skew angles between lines, overlapping words and adjacent text lines touching. The appearance of slant in the text line, punctuation marks and the non-uniform spacing of words are the major difficulties in word segmentation process. Freestyle and unconstrained handwriting text line segmentation is considered a complex and challenging task due to the following characteristics [1]

Fluctuating lines or skew variability [2, 3]. Lines of text in general are not straight. The inter-line distance variability and inconsistent distance between the components may vary due to writer movement. It may be straight, straight by segments, or curved [1]. According to Okun [4], three types of skew exist in documents:

- A Global skew: all the page blocks have the same orientation
- multiple skew: unaligned paragraphs or slant is different in different blocks of the page such as the FLAUBERT's drafts [2] which contain several blocks of text arranged in a non linear way, and numerous editorial marks such as erasures and word insertion, and
- non uniform text line skew or varying text line slope: slant is different along the same line of text, for example curvilinear text lines.

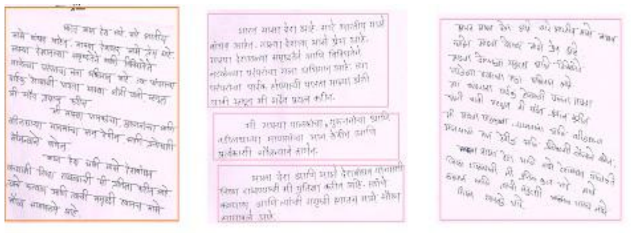
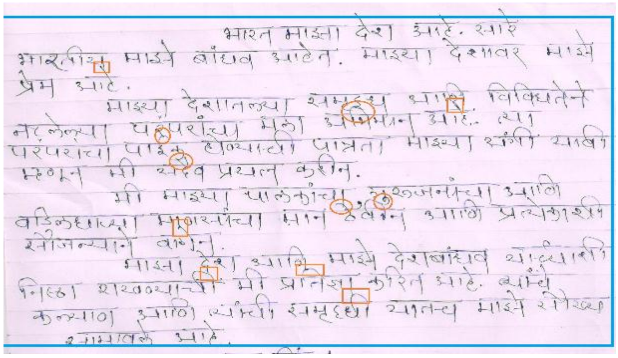


Figure 1. a. Global Skew b. Multiple Skew
c. Non uniform text line Skew

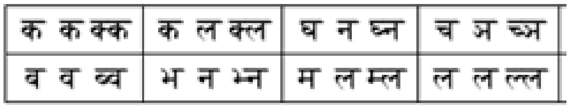
Line proximity i.e. Small gaps between neighboring text lines will cause touching or overlapping of ascenders or descenders. Text lines may be touching or overlapped, when upper-strokes and down-strokes of two consecutive lines are near or touching, or ascenders and descenders of adjacent lines interfere.[6]



भारत माझा देश आहे. सारे
भारतीय माझे बांधव आहेत. माझ्या देशावर माझे
प्रेम आहे.
माझ्या देशातल्या समृद्ध आणि विविधतेने
नटलेल्या परंपरांचा मला अभिमान आहे. त्या
परंपरांचा पाईक होण्याची पात्रता माझ्या अंगी यावी
म्हणून मी सदैव प्रयत्न करीन.
मी माझ्या पालकांचा, गुरुजनांचा आणि
वडिलधाऱ्या माणसांचा मान ठेवीन आणि प्रत्येकाशी
सौजन्याने वागेन.
माझा देश आणि माझे देशबांधव यांच्याशी
निष्ठा राखण्याची मी प्रतिज्ञा करीत आहे. त्यांचे
कल्याण आणि त्यांची समृद्धी यातच माझे सौख्य
सामावले आहे.

Figure. 2:a. Overlapping components separated (circle) b. touching component separated into two parts (rectangle) in Devnagari Script.

Writing fragmentations ,Characters are made up of more than one connected component. This applies to Indian scripts such as Marathi , hindi, Telugu, Tamil, Bangla, and Malayalam and Arabic writing with massive presence of diacritical points.



क क क्क	क ल क्ल	घ न घ्न	च ञ च्ञ
व व व्व	भ न भ्न	म ल म्ल	ल ल ल्ल

Figure 3 Writing Fragmentation

III. Segmentation Methods

Image segmentation is the division of an image into **regions** or **categories**, which correspond to different objects or parts of objects. *Every* pixel in an image is allocated to one of a number of these categories. A good segmentation is typically one in which:

- pixels in the same category have similar grayscale of multivariate values and form a connected region,
- Neighboring pixels which are in different categories have dissimilar values.

Segmentation algorithms are based on one of two basic proper-ties of intensity values discontinuity and similarity. There are three general approaches to segmentation, termed thresholding, edge-based methods and region-based methods.

A. In thresholding, pixels are allocated to categories according to the range of values in which a pixel lies.
B. In edge-based segmentation, an edge filter is applied to the image, pixels are classified as *edge* or *non-edge* depending on the filter output, and pixels which are not separated by an edge are allocated to the same category
C. In Region-based segmentation algorithms operate iteratively by grouping together pixels which are neighbors and have similar values and splitting groups of pixels which are dissimilar in value.

IV. Related work

Since 1960's character segmentation and recognition is an active field of research. It is still an open problem in the field of pattern recognition and image processing. Text line segmentation can be roughly categorized as bottom-up or top-down.

A top-down page segmentation technique known as the recursive X-Y cut decomposes a document image recursively into a set of rectangular blocks [8]. The connected component based methods merge neighboring connected components using a few simple rules on the geometric relationship between neighboring blocks. Connected component grouping [10] methods are sensitive to topological changes of the connected components, and it is not easy to derive script independent merging rules based on connected components.

Rodolfo P. dos Santos and Gabriela S. Clemente propose an efficient algorithm to segment handwritten text lines[18]. The text line algorithm uses a morphological operator to obtain the features of the images. A sequence of histogram projection and recovery is proposed to obtain the line segmented region of the text. A horizontal histogram projection is performed which results in the text lines positions. A threshold is applied to divide the lines in different regions. To eliminate false lines another threshold is used

Projection based methods may be one of the most successful top-down algorithms for machine printed documents[7].The gap between two neighboring text line is typically significant, the projection of text lines is easily separable in the orthogonal direction. The gaps between two neighboring handwritten lines may not be equal or handwritten lines are not straight [14] these methods cannot be used directly in handwritten documents. Another disadvantage of the top-down approaches is that they cannot easily process complex non-Manhattan layouts.

Arivazhagan, M. (2007) presented a piece-wise projection profile technique to segment a handwritten document into distinct lines of text by obtaining an initial set of candidate lines [7]. The lines traverse around the connected component by associating it to the line above or below.

Text line extraction from unconstrained handwritten documents is a challenge because the text lines are often skewed and curved and the space between lines is not obvious. To solve this problem, Yin & Liu [14] has propose an approach based on minimum spanning tree (MST) clustering with new distance measures.

Louloudis, Gatos, Pratikakis, & Halatsis proposed technique is based on a strategy that consists of distinct steps. The first step includes preprocessing for image enhancement, connected component extraction and average character height estimation [11]. In the second step, a block-based Hough transform is used for the detection of potential text lines.

A top-down page segmentation technique known as the recursive X-Y cut decomposes a document image recursively into a set of rectangular blocks[8] . It uses black pixels instead of using image pixels to achieve improvement in computation.

Two novel approaches to extract text lines and words from handwritten document are presented by Papavassiliou, V., Stafylakis[12]. The Viterbi algorithm is used for line segmentation which is based on locating the optimal succession of text and gap areas within vertical zones along with text-line separator drawing technique. The connected components are assigned to text lines in the end. A gap metric that exploits the objective function of a soft-margin linear SVM that separates successive connected components is used for word segmentation.

V. PERFORMANCE EVALUTION AND COMPARISION

For experimental purpose we have created our own dataset of handwritten characters. We have collected 65 handwritten marathi documents written by different individuals belonging to different categories in a separate sheet without any restrictions. It is done so to collect various samples of handwriting with different writing style, size, width etc. We have digitized the handwritten documents using scanner at 300 DPI in color mode and stored the scanned images in the jpeg format.

Before segmentation we have to perform some preprocessing on every document image as follows

- Convert to black and white
- Remove all object containing fewer than 30 pixels
- Perform morphological operations like erosion and dilation for normalization of image

After preprocessing is over we can apply the image to various segmentation algorithms to verify the results as follows
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 [15].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.

The text line algorithm uses a morphological operator to obtain the features of the images. Following, a sequence of histogram projection and recovery is proposed to obtain the line segmented region of the text. First, a Y histogram projection is performed which results in the text lines positions. To divide the lines in different regions a threshold is applied.

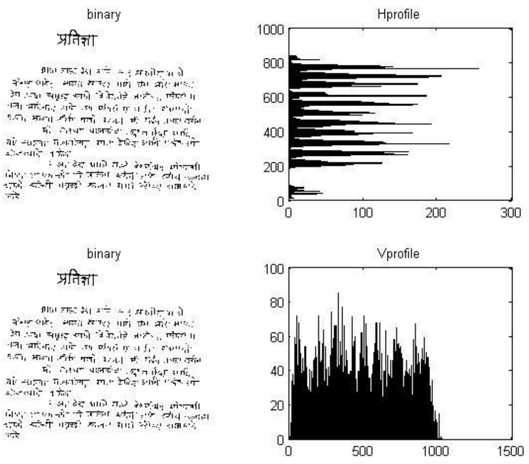


Figure 4 Horizontal and Vertical Projection Profile

In Run **Length Smearing method method [16]**, 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

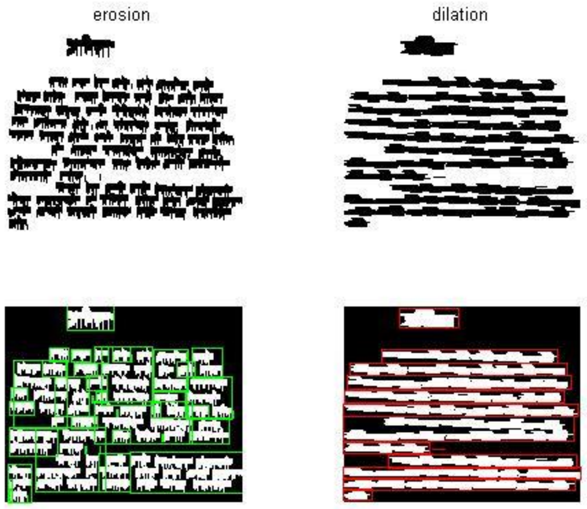


Figure 5. Run length Smearing

A technique based on Bounding Box[16] 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 [16].

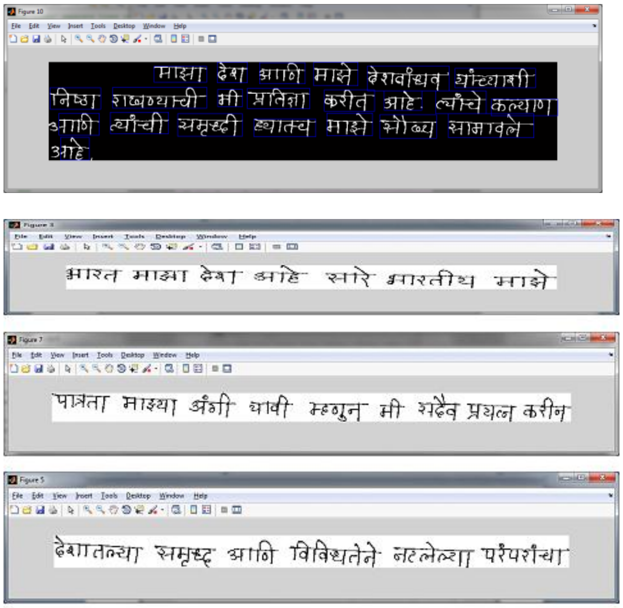


Figure 10
File Edit View Insert Tools Desktop Window Help
Figure 8
File Edit View Insert Tools Desktop Window Help
Figure 7
File Edit View Insert Tools Desktop Window Help
Figure 5
File Edit View Insert Tools Desktop Window Help

माझा देश आणि माझे देशबांधव यांच्याशी निष्ठा राखण्याची मी प्रतिज्ञा करीत आहे. त्यांचे कल्याण आणि त्यांची समृद्धी ह्यातच माझे सौख्य सामावले आहे.

भारत माझा देश आहे सारे भारतीय माझे

पात्रता माझ्या अंगी यावी म्हणून मी सदैव प्रयत्न करीन

देशातल्या समृद्ध आणि विविधतेने नटलेल्या परंपरांचा

Figure 6. Bounding Box Technique

Using Morphological Operators [17] Erosion and Dilation are the two operators whose combination or series of combination can be applied with different structuring element depending upon the size of the image. In erosion, the input image is eroded with the structuring element to obtain the processed image. The processed image obtained is sharper. The reverse happens in dilation. These operators are usually applied for detecting and removing header line from the word.

Table 1. Segmentation of different handwritten samples

Handwritten Document Sample	**Segmentation Methods Accuracy in percentage %**					
	Projection Profile	**Projection Profile with Morphology**	**Run Length Smearing method**	**Run Length Smearing method with morphology**	**Bounding Box method**	**Bounding box with morphology**
Document 1	**93**	**97**	**90**	**90**	**93**	**97**
Document 2	**92**	**95**	**91**	**92**	**90**	**95**
Document 3	**88**	**92**	**83**	**84**	**88**	**92**
Document 4	**86**	**92**	**85**	**88**	**84**	**85**
Document 5	**95**	**97**	**92**	**91**	**95**	**97**
Document 6	**85**	**88**	**90**	**91**	**85**	**85**

Document 7	**95**	**97**	**91**	**92**	**95**	**97**
Document 8	**87**	**89**	**88**	**89**	**87**	**89**
Document 9	**93**	**96**	**91**	**92**	**93**	**96**
Document 10	**91**	**93**	**88**	**88**	**91**	**93**
Average	**90**	**93**	**89**	**89**	**90**	**91**

Table 2. Results of Segmentation Methods

Segmentation Method	**Segmentation Accuracy**
Projection Profile	90 %
Projection Profile with Morphology	93 %
Run Length Smearing method	89 %
Run Length Smearing method with morphology	89 %
Bounding Box method	90 %
Bounding box with morphology	91 %

VI. CONCLUDING REMARK

This paper has provided a comparative study of the methods for off-line handwriting text line segmentation previously proposed by researchers. Three different methods like Projection profiles, Run Length smearing method and bounding box methods are used for text line extraction of Handwritten Devnagari Documents. These proposed methods are experimented on our own dataset collected from different writers. We have applied all the methods on 65 handwriiten document samples . Among all other proposed methods Morphological operations with projection profile gives the best segmentation rate of 93% because this method works well for clearly separated lines . But this method cannot divide the touching or overlapping lines and instead it will merge those lines. Skewed or slanted and intersecting lines are main challenges in line segmentation problem . We can combine these methods to get better results of line segmentation in handwritten documents.

REFERENCES

[1] L. L. Sulem, A. Zahour, B. Taconet, "Text line segmentation of historical documents: a survey", IJDAR, Vol. 9, No. 2-4, pp. 123-138 , 2007.

[2] S. Nicolas, T. Paquet, L. Heutte, "Text Line Segmentation in Handwritten Document Using a Production System", Proceedings of the 9th IWFHR,Tokyo, Japan, pp. 245-250, 2004.

[3] A. Zahour, B. Taconet, P. Mercy, and S. Ramdane, "Arabic Hand-written Text-line Extraction", in Proceedings of the Sixth International. Conference on Document Analysis and Recognition, ICDAR 2001, Seattle, USA, pp. 281–285, September 10-13 2001.

[4] O.Okun, M. Pietikainen, and J. Sauvola, "Document skew estimation without angle range restriction," IJDAR 2, pp. 132 - 144, 1999.

[5] N. Tripathy and U. Pal. ,"Handwriting Segmentation of Unconstrained Oriya Text," in International Workshop on Frontiers in Handwriting Recognition, pp. 306–311 , 2004.

[6] Pal U., Datta S. ," Segmentation of Bangla unconstrained handwritten text",Proceedings of Seventh International Conference on Document Analysis and Recognition, pp 1128 – 1132,2003.

[7] Arivazhagan, M. ." A statistical approach to line segmentation in handwritten documents. *Document Recognition and Retrieval" XIV, Proceedings of SPIE, San Jose, CA, USA, 6500,2007*.

[8] Ha, J., Haralick, R. M., & Phillips, I. T.," Recursive X-Y Cut using Bounding Boxes of Connected Components ", 952–955,1995.

[9] He, S., Samara, P., Burgers, J., & Schomaker, L. "Image-based historical manuscript dating using contour and stroke fragments. *Pattern Recognition* " , *58*.,2016

[10] Le, V. P., Nayef, N., Visani, M., Ogier, J. M., & Tran, C. De.,"Text and non-text segmentation based on connected component features ". In *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR* (Vol. 2015–November).

[11] Louloudis, G., Gatos, B., Pratikakis, I., & Halatsis, K. (n.d.). "A Block-Based Hough Transform Mapping for Text Line Detection in Handwritten Documents", Department of Informatics , Elsevire Pattern Recognition Tenth International Workshop on Frontiers in Handwriting Recognition, Oct 2006.

[12] Papavassiliou, V., Stafylakis, T., Katsouros, V., & Carayannis, G.."Handwritten document image segmentation into text lines and words ". *Pattern Recognition*, *43*(1), 369–377, 2010.

[13] A.N. Rajath. "An Adaptive Approach : Text Line Extraction from Multi-Skewed Hand Written Documents", *5*(6), 158–161,2015.

[14] Yin, F. E. I, & Liu, C.." Handwritten text line extraction based on minimum spanning tree clustering ". *International Conference on Wavelet Analysis and Pattern Recognition*, 1123–1128,2007.

[15] H. R. Mamatha and k. Srikantamurthy, "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

[16] Chethana, H. T., & Mamatha, H. R.. "Comparative Study of Text Line Segmentation on Handwritten Kannada Documents", *7*(1), 26–33,2016.

[17] Kinhekar, S..”Comparative Study of Segmentation and Recognition Methods for Handwritten Devnagari Script “, *105*(9), 34–39,2014.

[18] Santos, R. P., Clemente, G. S., Ren, T. I., & Calvalcanti, G. D. C..” Text Line Segmentation Based on Morphology and Histogram Projection “,2009 .

Authors Profile

Mrs. S.A.Bhopi pursed Bachelor of Science from DR Babasaheb Ambedkar Marathwada University of Aurangabad, Maharashtra in 1997 and Master of Science from Swami Ramanand Teerth Marathwada University Nanded in year 1999. She is currently pursuing Ph.D. and currently working as Assistant Professor in MGM’s College of Computer Science and IT, Nanded. Department of Computer Science and IT, affiliated to S.R.T.M.U. University of Nanded. She has 13 years of teaching experience.

Dr. M. P. Singh received his Ph.D. in Computer science from Kumaun University Nanital, Uthrakhand, India, in 2001. He has completed his Master of Science in Computer Science from Allahabad University, Allahabad in 1995. Further he obtained the M. Tech. in Information technology from Mysore. He is currently as Associate Professor in Department of Computer Science, Institute of Engineering and Technology, Dr. B.R. Ambedkar University, Agra, UP, India since 2008. He is engaged in teaching and research since last 16 years. He has more than 80 research papers in journals of international and national repute. He also has received the Young Scientist Award in computer science by international Academy of Physical sciences, Allahabad in year 2005. He has guided 18 students for their doctorate in computer science. He is also referee of various international and national journals.