IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-7, Special Issue-8, Apr 2019 E-ISSN: 2347-2693

A Detailed Survey of Text Line Segmentation Methods in Handwritten Historical Documents and Palm Leaf Manuscripts

R. Spurgen Ratheash[1*], M. Mohamed Sathik[2]

[1, 2] Sadakathullah Appa College, Tirunelveli, Tamilnadu, India
Manonmaniam Sundaranar Univesity, Abishekapatti, Tirunelveli, India

Corresponding Author: spurgen@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7si8.99103 | Available online at: www.ijcseonline.org

Abstract— The revolution of document analysis provides the handwritten palm leaf manuscripts and historical documents, epigraphic into digital. Digital document is available as auto recognition of those historical documents which is the second revolution in document research. Many algorithms are available from the period of the 19th century to the 20th century with the researchers across the world. In order to achieve auto recognition, the line segmentation is a foremost process. Though automatic segmentation of text lines is still a burning research, many technical issues remain unsolved yet. The present survey has been carried out as the survey of newly proposed and modified methods of text line segmentation in palm leaf manuscripts and handwritten historical documents published during the period 2008 – 2018. It could benefit the researchers who do research in handwriting recognition.

Keywords— pre-processing, path finding approach, performance measure

I. INTRODUCTION

The digital world provides the potential way to convert handwritten documents into electronically usable data. This study is known as Document Image Analysis (DIA). The olden day documents such as palm leaf manuscripts and handwritten historical documents make DIA a Herculean task by means of strains, yellowing, low intensity variations, random noises, fading and degradation. In DIA, the text line segmentation is considered the most significant process step because the text line segmentation efficiency affects the accuracy of the whole recognition system. It contains line segmentation, word segmentation, character segmentation and text recognition modules. The line segmentation task is extracting and separating the text regions into individual lines. Many of the language scripts such as Japanese, Chinese and Latin have attained a consummate position. However, the Asian scripts pose many issues[1] from fluctuation in the base line, variability in skew between different as well as same line, flexible writing style of different writers, presence of touching components among two adjacent lines. Ceaseless survey is needed to give out many methods reported every decade in text line segmentation. It gives one the motivation to carry out a quality survey after the publishing concerned in 2008.The paper precedes pre-processing in section II, survey of line segmentation methods 2008-2018 given in table format in section III, performance measure in section IV, and conclusion in section V.

II. PREPROCESSING

The line segmentation requires noise free documents. The digitized palm leaf and historical handwritten document are practically impossible to provide noiseless documents. The pre-processing techniques are used to remove the noises and extract the text from their dominating backgrounds. Most of the papers provide the following pre-processing methods:

A. Binarization

Converting RGB or Gray scale valued image into binary image using various threshold techniques[2]. The thresholding provides Global and Local thresholding. Otsu's thresholding method produces best result for text data in document analysis.

B. Skew Correction

The skewed lines of binary images can changed into proper horizontal lines by skew correction. The skew has categorized three types as follow: Global, Multiple, and Non-Uniform Skew[3]. In skew correction the vertical projection profile method is used to rotate the image across different angles. The maximum value of standard deviation is calculated for an optimal rotation in the projection profile.

C. Extracting Connected Components

The edge of the text image calculated using edge detection methods and Stroke Width Transform is applied to identify the strokes in each pixel. The small dots, patches, and unwanted noise components of the images are removed from the text image[4].

III. SURVEY OF LINE SEGMENTATION

A. Adaptive Partial Projection (App)

In 2012, APP is derived from the partial projection method. It divides the image into vertical columns and the projection profile is applied on that column in order to achieve 'smoothing'. It removes peaks and valleys in the histogram of the image. The line extraction process is (i) find the number of lines, divide the image into vertical columns, calculate the horizontal projection profile, smoothen the histogram, find the base lines, (ii) find the valleys of smooth histogram, test all the valleys, check for incorrect top and bottom

line, test the number of base lines, test for connected component. APP method is used on palm leaf manuscripts to segment the lines that applied on 264 text lines from 60 palm leaf manuscripts, 129 lines (48.86%) and 177 lines (67.05%) were correctly segmented. The method provides the error when the text is overlapping with the succeeding or preceding lines[5].

B. Matched Filtering

Matched Filtering method consists of three major stages such as (i) Foreground Pixel Density (FPD) – binarize the image using otsu's threshold, extract the skeleton using thinning, form a smallest rectangle contains foreground pixels, randomly select circle regions on that rectangle and compute the mean in all foreground pixels. (ii) Centers of Text Line extraction (CTL) – filter binary image using convolution of G and L to generate filtered image, binarize the filtered image and remove the connected components, perform filling, thinning, removing spurs to get coarse skeleton image(CSI), refine CSI to get centers of text line by using top-down grouping techniques. Overlapping Connected Components (OCC) separation – get the overlapping connected component according to CTL and perform morphology to reduce noises, extract the skeleton, detect all junction points and get the overlapping CC. It is applied on HIT-MW database and produces 99.91% of Detection Rate(DR), 100% of Recognition Accuracy (RA), 99.95% of Performance Measure (FM)[6].

C. Mid Point Detection

The binarized image stored into a matrix, height of the line is calculated, and divide into vertical strips equal to 100 pixels. Horizontal Profile Projection finds white spaces between the two adjacent lines and identifies the midpoint then calculates the difference between adjacent midpoints. The difference is greater than the height of the line that it can be identified as the touching and overlapping lines. The number of lines is calculated by the midpoints and marks the segmentation points. Subsequently, save the matrix and display the output from the matrix. This method is applied on different Gurmukhi script scanned document written by different writers for well spaced document, overlapping and touched lines and it achieves 94% of accuracy in overall performance [3].

D. Thinning

The scanned Bangle image is blurred and scans the vertical column from the top to identify the continuous change of their intensity values. This intensity provides the information of changing the text lines and it suggests the separation point between the two lines. The touched lines and overlapping lines are separated by fixing the junction point in the middle of the touched lines. The junction point fixes above the segmentation point that identifies as the upper line and below the point consider as lower line. The overlapped area is treated as suspected area that are specified by the rectangle and then verified further to fix the segmentation point. The method applied on ICDAR 2013 Bangla documents and it provides 93.16% of Detection Rate (DR), 92.32% of Recognition Accuracy (RA), 92.74% of Performance Measure (FM)[7]

E. Competitive Learning Algorithm

The binary image processed to extract the Connected Components (CC) using two-scan algorithm, common height (H_{cc}), and width (W_{cc}) is calculated using median width and height and center of mass of each CC are calculated. Adaptive Partial Projection technique is used to detect the number of lines in the document. The vertically divided binary image is set to $4W_{cc}$, then the columns are built by Y – Projection profile histogram and smooth twice is used to move the average filter method with the window size of H_{cc} and H_{cc}. The peaks of histogram are identified as the median lines in the column. The competitive learning algorithm is used to identify the midpoint of the line. The method is applied on EFEO database and it provides 99.528% precision, 99.534%recall and 99.531% Performance Measure (FM)[4].

F. Fully Convolutional Network

The dense layers are removed from the Convolutional Neural Network forms FCN to process the image in variable size. Instead of decision taking FCN work as encoder and decoder. This method provides the solution for semantic segmentation problem. In this approach, the pooling layer is used to reduce image resolution in order to decrease the computation count and increase the filter size. In dilated convolution "A trous" algorithm is used with wavelet transform. It provides two advantages; size of the receptive field can be controlled without reducing the resolution not increasing the number of parameters. Secondly, reduce the number of parameters and depth of the network. The method is applied on ICDAR 2017 with 1600 images contains 10000 lines, 1500 images for training and 100 for validation. It also applied on cBAD dataset with 755 pages, 216 for training and 539 in testing. It provides the result of 0.66 Precision, 0.86 Recall, 0.75 Performance Measure (FM)[8].

G. Second Order Derivative Analysis

The text lines are blurred and it appears as a blob by Gaussian function. It gives an advantage of steerable that means the response of the filter which can be calculated as base filters. After processing all the scales, one gets line orientation, scale and strength of each pixel. Finally, one considers the pixels which have the strongest response within their line. Compare the neighbors which belong to the same line with the small difference between their orientation and scales. After further filtering, apply non-maxima suppression approach in selected ridge. When it falls within the area of other local maxima, which has higher filter response, the pixel with mild-low value is marked as red. This method applies on IAM database, and many dataset and it provides 99.8% Precision, 93.1% Recall and 91.6% Performance Measure(FM)[9].

H. Path Finding Approach

To extract the text components from palm leaf manuscripts, the edges of the image are calculated by Canny Edge detection and Stroke Width Transform. X –projection profile is used to identify the starting and ending point of the lines, Y – projection profile for each column to smooth the histogram. Further, the boundary of the line is identified by A* algorithm method and two cost functions are described such as Intensity difference cost function and Vertical cost function. Then, the image is compared with Ground Truth value. This method is applied on 100 pages of Khmer palm leaf manuscripts called Sleuk Rith Set collected form EFEO database and it produces the results of 92.15% Detection Rate(DR), 93.70% Recognition Accuracy, 92.92% Performance Measure(FM)[4]

IV. ANALYSIS ON PERFOMANCE MEASURE

In the survey of line segmentation from historical handwritten document and palm leaf manuscripts has two categories of performance measures. First, the result has the parameter of

Precision, Recall and Performance Measure. The Precision and Recall are specifically used for Information Extraction. Precision gives the result of number of document retrieved that relevant and Recall is the number of relevant document that retrieved. Precision and Recall are inversely proportional to each other and understanding of this difference is much more important to build an efficient classification system. For example, among the 15.6 million results the relevant links to my question is 2 million. 6 million of results were relevant but not produce by the particular search engine. Here, Precision calculated by 2M/15.6M = 0.13 that means all the retrieved links were relevant and Recall is calculated by 2M/8M = 0.25 that means retrieve all the relevant links[10]. The Precision and Recall is explained by the following mathematical formulae;

Precision P = TP/(TP+FP)

Here, True Positive that gives the number of correctly retrieved documents divided by the Total number of document retrieved.

Recall R = TP/(TP+FN)

Here, True Positive that gives the number of correctly retrieved documents divided by the Total number of relevant document retrieved.
The combination of Precision and Recall produces the Performance Measure (F – measure)

*F=2 * ((Precision * Recall) / (Precision + Recall)*

Second, the result provides the parameter of Detection Rate (DR), Recognition Accuracy (RA) and F- Measure. The text lines detected by the way of one-to-one matches between text lines detected detection in historical handwritten documents and palm leaf manuscripts. The efficiency measures by the F- measure values.

V. CONCLUSION

Text line segmentation is a Herculean task. This is because the method of lines is under the condition of skewed, multi skewed, well spaced, touching lines and overlapping lines. The Viterbi algorithm concentrates on skewed lines with high variability, skeletonization algorithm methods that are highly supported for multi skewed text lines. The fringe map method supports for the Telugu document text lines and the adaptive partial projection method is used to segment the touching characters in the document. The projection profile methods are highly supportive of the skewed approach and the text line in the ground truth. The matching score is computed as

Match Score (i,j) =

where T(s) is function that counts the number of points in set S; $G_j$is the set of all points inside the union of all selected regions of isolated characters in ground truth belonging to text line j; $R_i$the set of all points inside the region of result text line i; and I_{IC} the set of all points inside the union of all selected regions of ground truth isolated characters in the whole document page[4][11][12].

The region pair is considering by the match score is above to the acceptance threshold T . Assume that N is the number of text lines found in the ground truth, M to be the number if text lines by the approach, and o2o to be the number of one – to – one match pairs. The detection rate is defined by,

DR = o2o/N

and recognition accuracy is

RA = o2o/M

the evaluation metric F-measure score is calculated by combining the above two relation

FM = 2.DR.RA/DR+RA

In analysis, among the two methods the second method of result parameters; Detection Rate (DR), Recognition Accuracy (RA) is the correct way to measure the text line

line in the historical handwritten documents. Fully Convolution Network method provides the segmentation method for connected characters and overlapping characters. Table-1 provides a detailed survey of line segmentation and its limitations. The present survey analyzes the various line segmentation methods presented during the last decade. In all the algorithms, the results for the touching, overlapping text lines, skewed lines in the handwritten text documents and palm leaf manuscripts are analyzed. The survey concludes that there is no successive algorithm to segment the touched and overlapped lines in the Tamil palm leaf manuscripts. There is a vast space to be filled up with an efficient method to segment the Tamil palm leaf manuscript's touched and overlapping lines in the future.

Table -1: Survey of Text line segmentation methods during 2008 - 2018

Sl.No	Year	Authors	Proposed Method	Database	Applied for	Drawbacks	Text Languages
1.	2008	Themos et al[13]	Viterbi Algorithm	ICDAR 2007	Skew with high variability, Non strict left and right margins	Minimum efficiency to detect high variability of writing styles, sizes	English, French, German and Greek
2.	2008	Sanchez et al[14]	Skeletonization algorithm	PROHIST project data base	Multi skewed text lines	Not efficient in overlapped text lines	English
3.	2009	Rodolfo et al[15]	Morphology and Histogram Projection	IAM handwritten Data base	Handwritten document	Minimum Efficiency in upper character in the text line	English
4.	2010	Rajiv kumar et al[16]	Variable size windowing	Handwritten Gurumukhi document	Handwritten Gurumukhi Script	Not much efficient when the characters are combined in nature	Gurumukhi
5.	2010	Naresh kumar Garg	Block covering	Handwritten document	Handwritten	Not efficient in overlapping,	Bangla,

		et al[17]	method	from 15 writers	Devanagari, Hindi	broken parts and thick parts in upper modifiers	Devanakari, Telugu
6.	2010	Jija Das Gupta et al [18]	Handwritten text line segmentation approach	IAM database	Offline English letter	Not much efficient in overlap and touching component	English
7.	2011	Vijaya Kumar et al[19]	Fringe Map based method	Handwritten Telugu Character	Printed Telugu Script document	Minimum efficiency in non constant space exists between lines	Telugu
8.	2011	Rajib et al[20]	Skew Detection Algorithm	Handwritten Bangla words	Skewed Bangle words	Not efficient in some part of the word is skewed and the rest is not skewed	Bangla
9	2012	Ines Ben et al[21]	Multi level framework	IAM – HistDB Database, ICFHR 2010	Touching elements as well as skewed text lines	Minimum Efficiency in overlapping characters	English
10.	2012	Rapeeporn et al[5]	Adaptive Partial Projection	Thai palm leaf manuscripts from Mahasarakham University, Northeaster Thailand	Touching character in consecutive lines	Minimum Efficiency in vowels which are too close to the upper or lower line, long prolonged components, and connected components of consecutive lines.	Thai
11.	2012	HandeAdiguzel et al[22]	Connected component and projection profile	Ottoman printed book	Well Spaced lines	Minimum efficiency in small sized components placed in wrong lines	Ottoman documents
12.	2014	Xi Zhang et al[23]	Seam carving	ICDRA2013 Handwritten segmentation contest dataset	Minimum touched and well space lines	Inefficiency in large components which touch multiple text lines	English and Greek
13.	2014	Rapeeporn et al[24]	Combined method for segmentation	Thai palm leaf manuscripts	Text line segmentation	Error in touching characters, noise surrounded characters, incorrect line separation	Thai
14.	2014	Youbao Tang et al[6]	Matched Filtering and Top down Grouping Method	ICFHR 2010, HITMW, Contest database	Text line extraction	Inefficiency in overlapping lines	English, French, Chinese, German, Greek
15.	2015	Payal Jindal et al[3]	Midpoint Detection Technique	Gurumukhi Handwritten document	Skewed lines, overlapped lines, Connected components	Not efficient for complex overlap text lines	English, Gurumukhi
16.	2015	Mullick et al[4]	Thinning	ICDAR2013	Text line segmentation	Not efficient in skew lines, overlap and low space	English, Greek, Bangla
17.	2016	Dona Valy et al[25]	Competitive Learning Algorithm	EFEO data base	Skewed and fluctuated lines	Not efficient in cursive characters and need improvement in touching components	Khmer
18.	2016	Banumathi et al[26]	Projection Profile Technique	Kannada handwritten document	Skewed lines and spaced lines	Not efficient in Low space lines, Overlapping lines	Kannada
19.	2016	Quang Nhat Vo et al[27]	Fully Convolutional Network (FCN) and Line Adjacency Graph (LAG)	ICDAR 2013 Handwritten Segmentation Contest dataset	Touching Characters	Not efficient in complex touching characters, Error in Indian documents where the characters vertically connect to the text lines.	English, Greek, Bangla
20.	2017	Himanshu Jain et al[2]	Bottom up procedure	ICDAR2009 Text segmentation Contest Dataset	Touching and overlapping character	Not efficient while using minimum artificial parameters on overlapping characters	Greek, French, German, English
21.	2017	Guillaume et al[8]	Fully Convolutional Network	ICDAR 2017 and cBAD , ICDAR 2015 and ANDAR, RIMES Dataset	Well spaced lines	Not efficient in touching and overlapping characters	English
22.	2017	Dona Valy et al[7]	Path Finding Techniques	EFEO database National Library Budhist Institute	Skew, fluctuated, discontinued text lines	Not efficient in touched and overlapped characters	Khmer
23.	2017	Kathirvalavakumar et al[28]	Projection Profile and Connected component	Tamil printed document	Skew slant lines	Not efficient in handwritten document, Touching and overlapping characters in above	Tamil

			Techniques			and below text lines	
24.	2018	David Et al[9]	Second order derivatives	IAM, GRPOLY-DB, ICDAR 2009, cBAD data set	Skewed Lines	Not efficient in Overlapping characters	English

REFERENCES

[1] Made Windu Antara Kesiman, Dona Voly., "Benchmarking of Document Image Analysis Tasks for Palm Leaf Manuscripts from Southeast Asia",*JImaging,4,43;doi:10.3390/jimaging 4020043*, 2018.

[2] Himanshu Jain, Archana Praveen Kumar, "A Bottom Up Procedure for Text Line Segmentation of Latin Script", IEEE, *pp 1182 – 1187*, 2017.

[3] Payal Jindal, Dr. Balkrishnan Jindal, "Line and Word Segmentation of Handwritten Text Documents written in Gurmukhi Script using Mid Point Detection Technique", IEEE, *Proceedings of 2015 RACES UIET Panjab University Chandigarh 21-22nd December2015,* 2015.

[4] Dona Valy, Michel Verleysen, Kimheng Sek, "Line Segmentation for Grayscale Text Images of Khmer Palm Leaf Manuscripts", IEEE, 2017.

[5] Rapeeporrn Chamchong, Chum Che Fung, "Text Line Extraction Using Adaptive Partial Projection for Palm Leaf Manuscripts from Thailand", IEEE, *2012 International Conference on Frontiers in Handwriting Recognition, IEEE, PP 586-5591,* 2012.

[6] Youbao Tang, Xiangqain Wu, Wei Bu, "Text Line Segmentation Based on Matched Filtering an Top-down Grouping for Handwritten Documents", IEEE, *2014 11th IAPR International Workshop on Document Analysis System, PP 365 – 369*, 2014.

[7] Mullick, Banerjee, Bhatacharya, "An Efficient Segmentation Approach for Handwritten Bangla Document Image, IEEE, 2015.

[8] Guillaume Renton, Clement Chatelian, "Handwritten text line segmentation using Fully Convolutional Network", IEEE, *2017 14th International Conference on Document Analysis and Recognition, PP 5 – 9*, 2017..

[9] David Aldavert, Marcal Rusinol, "Manuscript Text Line Detection and Segmentation using Second Order Derivatives", IEEE, *2018 13th IAPR International Workshop ON Document Analysis Systems, PP 293 – 298*, 2018.

[10] https://towardsdatascience.com/model evaluation-i-precision-and-recall-166ddb257c7b

[11] Vishal Chavan, Kapil Mehrotra, "Text Line Segmentation of Multilingual Handwritten Documents Using Fourier Approximation", IEEE, *2017 Fourth International Conference on Image Information Processing(ICIIP), PP 250 – 255*, 2017.

[12] Ayush Padhan, Sidharth Behra, Paushpalata Pujari, "Comparative Study on Recent Text Line Segmentation Methods of Unconstrained Handwritten Scripts", IEEE, *International Conference on Energy, Communication, Data Analysis and Soft Computing(ICECDS-2017), PP 3853 – 3858*, 2017.

[13] Themos Stafylakis, Vasilis Papavassiliou, "Robust Test-Line and word segmentation for Handwritten Documents Images", IEEE, *PP 3393 – 3396*, 2008.

[14] Sanchez, Suarez, Mello, Oliveria, Alves,"Text Line Segmentation Images of Handwritten Historical Documents", IEEE, *Image Processing Theory, Tools and Applications*, 2008.

[15] Rodolfo P dos Santos, Gabriela S Clemente, Tsang Ing Ren, "Text Line Segmentation Based on Morphology and Histogram Projection", IEEE, *2009 10th International Conference on Document Analysis and Recognition PP 651 – 655*, 2009..

[16] Rajiv Kumar, Amardeep Singh, "Detection and Segmentation of Lines and Words in Gurmukhi Handwritten Text", IEEE, *2010 IEEE 2nd International Advance Computing Conference, PP 353 – 356*, 2010.

[17] Naresh Kumar Garg, Lakswinder Kaur, Jindal, "A new method for Line Segmentation of Handwritten Hindi Text", IEEE, *2010 Seventh International Conference on Information Technology, PP 392 – 397*, 2010.

[18] Jija Das Gupt, Bhabatosh Chanda, "A model based Text Line Segmentation method for Off-Line Handwritten Documents", *IEEE, 2010 12th International Conference on Frontiers in Handwritting Recognition, pp 125-129*, 2010.

[19] Vijaya Kumar Koppula, Atul Negi, " Fringe Map based Text Line Segmentation of Printed Telugu Document Images", IEEE, *2011 International Conference on Document Analysis and Recognition, PP 1294 – 1298*, 2011.

[20] Rajib Ghosh, Debnath Bhattacharyya, Tai hoon kim, Gang soo Lee, " New Algorithm for Skewing Detection of Handwritten Bangla Words", *Springer – Verlag Berlin Heidelberg, PP 153 – 159*, 2011.

[21] Ines ben Messaoud, Hamid Amiri, "A Multilevel Text Line Segmentation Framework for Handwritten Historical Documents", *IEEE, 2012 International Conference on Frontiers in Handwriting Recognition, PP 515-520*, 2012.

[22] Hande Adiguzel, Emre Sahin, Pinar Duygulu, "A Hybrid Approach for Line Segmentation in Handwritten Documents", *IEEE, 2012 International Conference on Frontiers in Handwriting Recognition, PP 503-508*, 2012.

[23] Xi Zhang, Chew Lim Tan, " Text Line Segmentation for Handwritten Documents Using Constrained Seam Carving", *2014 14th International Conference on Frontiers in Handwriting Recognition, IEEE, PP 98-103*, 2014.

[24] Rapeeporrn Chamchong, Chum Che Fung, "A Combined Method of Segmentation for Connected Handwritten on Palm Leaf Manuscripts",IEEE, *2014 IEEE International Conference on Systems, Man and Cybernetics, PP 4158 – 4161*, 2014.

[25] Dona Valy, Michel Verleysen, Kimheng Sek, "Line Segmentation Approach for Ancient Palm Leaf Manuscripts using Competitive Algorithm", IEEE, *2016 15th International Conference on Frontiers in Handwriting Recognition, PP 108-113,* 2016.

[26] Banumathi, Jagadeesh Chandra, " Line and Word Segmentation of Kannada Handwritten Text documents using Projection Profile Technique", IEEE, *2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT) PP 196 – 201*, 2016.

[27] Quang Nhat Vo, GueeSang Lee, " Dense Prediction for Text Line segmentation in Handwritten Document Images", IEEE, *ICIP 2016, PP 3264 – 3268*, 2016.

[28] Kathirvalavakumar Thangairulappan, Karthigai selvi Mohan, " efficient Segmentation of Printed Tamil Script into Characters Using Projection and Structure", IEEE, *2017 Fourth International Conference on Image Information Processing(ICIIP), PP 484 – 489*, 2017.

