DOI: http://dx.doi.org/10.26483/ijarcs.v9i1.5428

Volume 9, No. 1, January-February 2018

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

A COMPARATIVE ANALYSIS OF LINE AND WORD SEGMENTATION FOR HANDWRITTEN DOCUMENT IMAGE

Neerugatti Varipally Vishwanath
Assistant Professor, Department of Electronics and Communication Engineering, St.Peter's Engineering College, Hyderabad,India

Murugan R
Associate Professor, Department of Electronics and Communication Engineering, St.Peter's Engineering College, Hyderabad,India

D.HariSaiRam' T.Leela Sai and Shashank Nanda Kumar
UG students, Department of ECE,
St. Peter's Engineering College, Hyderabad, India

Abstract: Segmentation of handwritten image is the challenging task in Optical Character Recognition. Due to the improper segmentation of this task, many of the methods produce poor recognition rate. Text characteristics can vary in size, font, alignment, color, orientation, and contrast and background information. These characteristics variations turn the process of word detection complex and difficult. Since handwritten text can vary greatly depending on the user skills, disposition and cultural background. The segmentation should be possible based on zooming, a line portion of content, and a word section from line and character fragment from word. This should be possible by the utilization of level, vertical technique. This paper surveys numerous essential and propelled division strategies of written by hand archive pictures.

Keywords: handwritten document image analysis,Text line segmentation, Word segmentation, Optical Character Recognition

1. INTRODUCTION

THE Document image segmentation to text lines and words is a criticalstage towards unconstrained handwritten document recognition. Variation of the skew angle between text lines or along thesame text line, existence of overlapping or touching lines, variablecharacter size and non-Manhattan layout are the challenges of textline extraction. Due to high variability of writing styles, scripts, etc., methods that do not use any prior knowledge and adapt to the propertiesof the document image, as the proposed, would be more robust[1].

Line extraction techniques may be categorized as projection based, grouping, smearing and Hough-based. Globalprojections-based approaches are very effective for machineprinted documents but cannot handle text lines with differentskew angles. However, they can be applied for skew correctionin documents with constant skew angle. Hough-based methodhandle documents with variation in the skew angle between textlines but are not very effective when the skew of a text line variesalong its width. Thus, we adopt piece-wise projections which candeal with both types of skew angle variation[2].

On the other hand, piece-wise projections are sensitive tocharacters' size variation within text lines and significant gapsbetween successive words. These occurrences influence theeffectiveness of smearing methods too[3]. In such cases, the resultsof two adjacent zones may be ambiguous, affecting the drawing oftext-line separators along the document width. To deal with theseproblems we introduce a smooth version of the projection profilesto oversegment each zone into candidate text and gap regions. Then,we reclassify these regions by applying an HMM formulation thatenhances statistics from the whole document page. Starting fromleft and moving to the right we combine separators of consecutiveones considering their proximity and the local foreground density.

Grouping approaches can handle complex layouts, but they failto distinguish touching text lines[4]. In our approach, we deal withsuch a case by splitting the respective connected component (CC)and assign the individual parts to the corresponding text lines.

In word segmentation, most of the proposed techniquesconsidera spatial measure of the gap between successive CCs and define athreshold to classify "within" and "between" word gaps[5].Themeasures are sensitive to CCs shape, e.g. a simple extension of thehorizontal part of character "t". We introduce a novel gap measurewhich is more tolerant to such cases. The proposed measure resultsfrom the optimal value of the objective function of a soft-marginlinear SVM that separates consecutive CCs.

Preliminary versions of the text-line and word segmentation algorithmswere submitted to the Handwriting Segmentation Contestin ICDAR07, under the name ILSP-LWSeg, and performed the bestresults[6].

The organization of the rest of the paper is as follows: The recent related works presents in Section II. The description of text-line extraction from handwritten document images is in section III.In section IV, the various proposed technique of segmentation in text lines into wordsis presented.The comparisons and conclusions arepresented in Sections 5 and 6, respectively.

2. RELATED WORK

In this section, the brief review of recent work on textline and word segmentation in handwritten document images were presented.

As far as we know, the following techniques either achievedthe best results in the corresponding test datasets, or areelements of integrated systems for specific tasks. One of themost accurate methods uses piece-wise projection profiles to obtain an initial set of candidate lines and bivariate Gaussiandensities to assign overlapping CCs into text lines [7].

Trial comes about on an accumulation of 720 archives (English, Arabic and youngsters' penmanship) demonstrate that 97.31% of content lines were portioned accurately.The writersmention that “a more intelligent approach to cut anoverlapping component is the goal of future work”. A recentapproach [8] uses block-based Hough transform to detect lines and merging methods to correct false alarms. Although thealgorithm achieves a 93.1% detection rate and a 96%recognition rate, it is not flexible to follow variation of skewangle along the same text line and not very precise in theassignment of accents to text lines.

Li et al. [9] examine the content line identification undertaking as a picture division issue. They utilize a Gaussian window to change over a double picture into a smooth dark scale. Then they adopt the level set method toevolve text-line boundaries and finally, geometrical constrainsare imposed to group CCs or segments as text lines. Theyreport pixel-level hit rates varying from 92% to 98% ondifferent scripts and mention that “the major failures happenbecause two neighboring text lines touch each othersignificantly”. A similar method [10] evaluates eight differentspatial measures between pairs of CCs to locate words inhandwritten postal ad- dresses. The best metric proved to bethe one which combines the result of the minimum run-lengthmethod and the vertical overlapping of two successive CCs.

Additionally, this metric is adjusted by utilizing the results ofa punctuation detection algorithm (periods and commas).Then, a suitable threshold is computed by an iterativeprocedure. The algorithm tested on 1000 address images andperformed an error rate of about 10%. Manmatha andRothfeder [11] propose an effective for noisy historicaldocuments scale space approach. The line image is filteredwith an anisotropic Laplacian at several scales in order toproduce blobs which correspond to portions of characters atsmall scales and to words at larger scales. The optimum scaleis estimated by three different techniques (line height, pageaveraging and free search) from which the line height showed best results. Much more challenging task is line segmentationin historical documents due to a great deal of noise. Feldbach and Tonnies [12] have proposed a bottom up method forhistorical church documents that requires parameters to be setaccording to the type of handwriting. They report a 90%correct segmentation rate for constant parameter values whichrises to 97% for adjusted ones.

Another integrated system for such documents [13] creates aforeground/background transition count map to find probablelocations of text lines and applies min-cut/max-flow algorithmto separate initially connected text lines .The method performshigh accuracy (over 98%) in 20 images of GeorgeWashington's manuscript.

3. OPTICAL CHARACTER RECOGNITION SYSTEM

Machine-printed text recognition system originated in he late 1950s and has been widely used since mid 1990s in desktop computers. Many of the world's information is heldin hard copydocuments. OCR system releases this information via text on paper through an electronic shape. Once in this form, theinformation retrieval system can be used to locate matter of interest, and word processing software can beused foraltering the text. OCR technology has beendeveloped somuch that today's system is indeed useful in dealing with an expansive variety of machine-printed documents. processing a neatly printed image can deliver results withan accuracy of 99%or more [14].

When a scanner scans a pageof text into a system(i.e. aPC or a laptop), the text is saved in theform of an electronic file composed of minute dots, alsoknown as pixels that is appeared in figure 1. A omputer does not consider these set of pixels as text, in fact, itis considered as animage of the text [15].

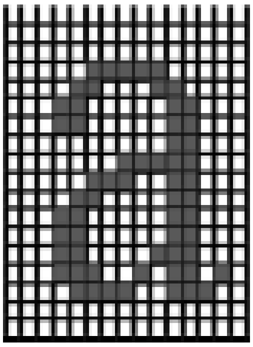

Figure 1: Image of text

These images cannot be processed by the word processors. So, to be able to edit the group of pixels theymustfirst beconvertedinto words. For this, the picture must undergo a complex phenomenon called the Optical Character Recognition[16]. The following image shows the general OCR system:

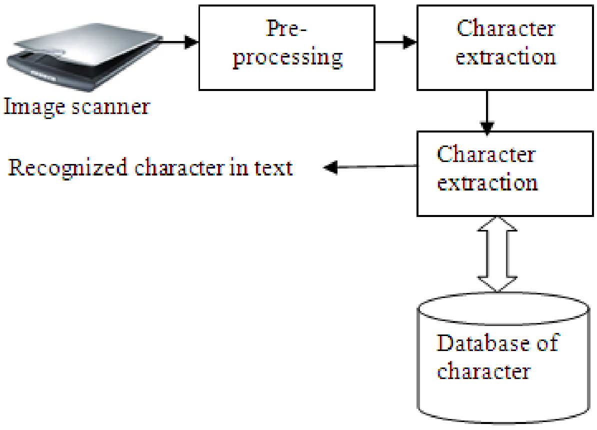

Figure 2: General OCR System

In figure 2 the first image is corresponds to the working of the scanned image,i.e, binarization, noiseremoval, refinement, skew correction and detection. The pre-processing which corresponds the character extraction; it is pretreated to perform the line, word and character separation.[17]. The last phase is mindful for the feature extractionand selection resulting in image recognition.

A. *Steps of Optical Character Recognition*

The steps of OCR system is shown in the following figure 3

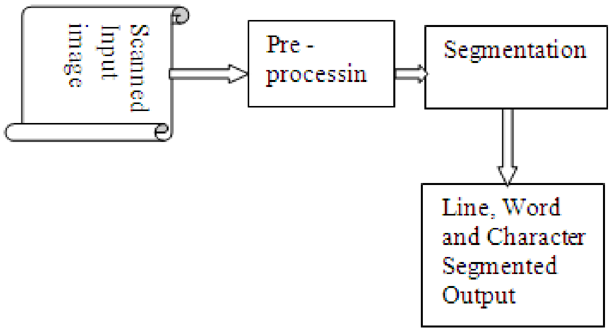

Figure 3. Block diagram of hand written document image line ,word and character segmentation.

1. Image acquisition

This includes checking a report and putting away it as a picture. Their answer (number of dabs per inch, dpi) decides the rate of process [17].

2. Pre-processing

Procedure of speaking to the examined picture of further handling. Preprocessing plans to deliver information that are simple for the OCR framework to work precisely. It decreases commotion and mutilation, evacuates skewness and performs skeleton sing of the picture, consequently streamlining the preparing whatever is left of the stages [18].

3. Segmentation

After the pre-processing a 'spotless' report is acquired. The following stage is division. In this stage, portioning the report into its sub-segments. It isolates the diverse legitimate parts, similar to content from designs, line of a section, and characters of a word. Division is a vital period of OCR, in light of the fact that it can reach in partition of words, lines or

characters specifically influence the acknowledgment rate of the content. Actually right acknowledgment in light of right division. The precision of the manually written character acknowledgment framework accurately relies upon division process [19].

The segmentation task is very difficult in document images because hand written characters exhibit my properties like shape, structure, touching, not proper alignment, lot of variations in character size and skew angle.So, many research people have been published research on line, word and character segmentation which is quite good still there is requirement.For any hand written character recognition system the segmentation process is must .Figure shows the different segmentation process. in figure 4 shows the different line segmentation process.

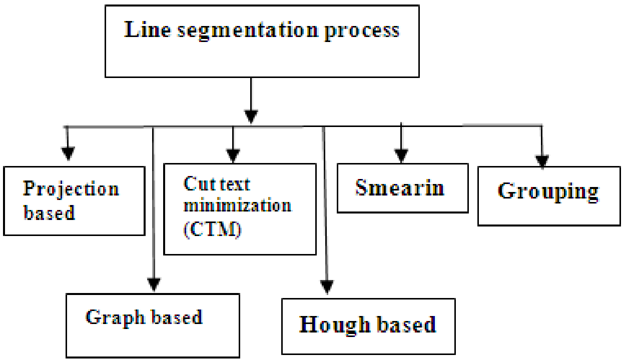

Figure 4. Different segmentation process.

4. Feature Extraction

An arrangement of standards put away on OCR motor looking at against character's shape and its highlights that recognizes each character distinguish a character. The fundamental piece of the acknowledgment framework configuration is the determination of a steady illustrative arrangement of highlights. It is the most significant issue in the planning issues associated with building an OCR framework [20].

5. Classification

The fundamental basic leadership phase of an OCR framework is order. Order utilizes the highlights removed in the element extraction stage to distinguish the content section [20].

4. SEGMENTATION

The preprocessing stage yields a spotless archive. The adequate measure of shape data, high pressure and low commotion on standardized picture is gotten. The following stage subsequent to preprocessing is division. Division is the way toward sectioning the entire record into sub segments. Division is of two sorts, outside and inside division. While outside division is the detachment of the sentences, sections, and other such written work units. Interior division is the seclusion of the characters and letters [21].

A. Segmentation processes

The Segmentation processeshave the following steps:

1. Line segmentation

Line division is the procedure in which from the picture, we extricate just lines or separate the lines. Flat projection of an archive picture is most ordinarily used to extricate the lines from the report. The even projection will have isolated pinnacles and valleys for the lines that are all around isolated and are not tiled, which fill in as the separators of the content lines. These valleys are effectively identified and used to decide the area of limits between the lines. Word division is the procedure in which from the line division, we separate just words. As we realize that there is a separation between single word another word, this idea is utilized for word division [22].

2.Word segmentation

Word division is a procedure of separating a string into its part words. Word part is the way toward parsing linked content to induce where word breaks exist. By utilizing

vertical projection profile, one can get segment wholes. By searching for minima in level projection profile of the page, we can isolate the lines and after that different words by taking a gander at minima in the vertical projection profile of a solitary
line. By utilizing the valleys in the vertical projection of a line picture, one can remove words from a line and furthermore extricating singular characters from the word [22].

3. Character segmentation

In character division, we extricate just characters from word. Character division is a troublesome advance of OCR frameworks as it separates important areas for examination. This progression breaks down the pictures into classifiable units called character. As indicated by Casey and Lecolnet [22].

5. COMPARISONS

Many segmentation methods are proposed for achieving high recognition accuracy. The comparison of line word and character segmentation if found in table 1.

TABLE 1:
Line, word and character segmentation analysis

S.no	Authors	Year	Approach	Feature/used	Efficiency
1	Payal Jindalet.al[23]	2015	Mid-point detection	Spaces that separates two lines	95%
2	O.Surinta et.al[24]	2014	A path planning algorithm	Smart combination of simple soft cost-functions	99.90%
3	Karmakar et.al[25]	2014	Space between two lines	Space recognition technique	100%
4	Y.Tang et.al[26]	2014	Matched filtering	Top-down grouping	99.95%
5	Snehdeep et.al[27]	2014	Midpoint analysis	Midpoint	93.05%
6	H.S.Vishwas et.al[28]	2014	Connected component labeling method	Projection profile and search for foreground pixel	Good
7	Sonam jain et.al[29]	2014	Correlation	Windowing method	100%
8	Micheal.B et.al[30]	2013	Dynamic multiplayer perception	Textblock,core-text line decoration, background periphery	96.30%
9	M.Javed et.al[31]	2013	Horizontal project profile curve	Local minima points	96.96%
10	M.M.Mehdi et.al[32]	2013	Smart data structure	Image scaling and re-scaling	100%
11	Alireza et.al[33]	2011	Piece-wise potential separating line	Rowstrip	92.35%
12	Vijaya kumar.k et.al[34]	2011	Peak fringe map number	Filtering	97.24%
13	W.Boussellaa et.al[35]	2010	Block covering analysis unsupervised technique	Fuzzy c-means algorithm	91.72%
14	N.Priyanka et.al[36]	2010	Histogram based	Run length based smearing	99.50%
15	Fei Yin et.al[37]	2009	Vertical bays(VB) framework	Gaussian component	94.30%
16	G.Louloudis et.al[38]	2008	Hough transform Enhancing text line	Connected component analysis.	95.80%
17	Fei Yin et.al[39]	2008	Minimal spanning tree(MST)	Connected component analysis	98.02%

18	Parthapratim Roy et.al[40]	2008	Morphological operation and run length smearing algorithm	Foreground portion, erosion boundary information background portion	92.68%
19	S.Basu et.al[41]	2006	Face obstruction method	Hypothetical water flow	91.44%
20	Yili et.al[42]	2006	Structure using a gausian window	Level set method to evolve text line boundaries	92%
21	William A.Bet.al[43]	2006	Min-cut/max-flow graph algorithm	Adaptive local connectivity map	Good
22	Zhiuin Shi et.al[44]	2004	Adaptive local connectivity map	Local project profiles	95%

6. CONCLUSION

OCR of any handwritten/printed documents involves various stages ranging from scanning of the document to get its digital image, text line extraction, word extraction, character segmentation to the character recognition. In this paper, we made a comparative analysis of text line extraction, word extraction, character segmentation for the last twenty years. This analysis will help to know the complete frame work for text line extraction, word extraction,character segmentation.In future work we will propose one efficient algorithm for text line extraction, word extraction, character segmentation on unconstrained handwritten documents.

REFERENCES

[1] A. Chahi, I. El khadiri, Y. El merabet, Y. Ruichek, and R. Touahni, “Block wise local binary count for off-Line text-independent writer identification,” Expert Syst. Appl., vol. 93, pp. 1–14, 2018.

[2] Q. N. Vo, S. H. Kim, H. J. Yang, and G. Lee, “Binarization of degraded document images based on hierarchical deep supervised network,” Pattern Recognit., vol. 74, pp. 568–586, 2018.

[3] T. Mondal, N. Ragot, J.-Y. Ramel, and U. Pal, Comparative study of conventional time series matching techniques for word spotting, vol. 73. 2018.

[4] P. P. Roy, A. K. Bhunia, and U. Pal, “Date-field retrieval in scene image and video frames using text enhancement and shape coding,” Neurocomputing, vol. 274, pp. 37–49, 2018.

[5] F. Jia, C. Shi, K. He, C. Wang, and B. Xiao, “Degraded document image binarization using structural symmetry of strokes,” Pattern Recognit., vol. 74, pp. 225–240, 2018.

[6] B. Gatos, A. Antonacopoulos, N. Stamatopoulos, ICDAR2007 handwriting segmentation contest, in: Proceedings ofInternational Conference on Document Analysis and Recognition, 2007, pp. 1284–1288.

[7] Z. Razak, K. Zulkiflee, et al., Off-line handwriting text line segmentation: a review, International Journal of Computer Science and Network Security 8 (7) (2008) 12–20 .

[8] G. Louloudis, B. Gatos, C. Halatsis, Text line detection in unconstrained handwritten documents using a blockbased Hough transform approach, in: Proceedings of International Conference on Document Analysis and Recognition, 2007, pp. 599–603.

[9] Y. Li, Y. Zheng, D. Doermann, S. Jaeger, Scriptindependent text line segmentation in freestylehandwritten documents, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (8) (2008) 1313–329

[10] G. Seni, E. Cohen, External word segmentation of offline handwritten text lines,Pattern Recognition 27(1994)41–52.

[11] R. Manmatha, J.L. Rothfeder, A scale spaceapproach forautomatically segmenting words from historicalhandwritten documents, IEEE Transactions on PatternAnalysis and Machine Intelligence 27 (8) (2005) 1212–1225.

[12] M. Feldbach, K.D. Tonnies, Line detection andsegmentation in historical church registers, in:Proceedings of International Conference onDocumentAnalysis and Recognition, 2001, pp. 743–747.

[13] D.J. Kennard, W.A. Barrett, Separating lines of text infree-form handwritten historical documents, inProceedings of International Workshop on DocumentImage Analysis for Libraries, 2006, pp. 12–23.

[14] G. Louloudisa, B.Gatosb, I.Pratikakisb,C.HalatsisaText line and word segmentation of handwritten documents. Pattern Recognition (2008) pp. 3169 – 3183.

[15] Fei Yin, Cheng-LinLiu. Handwritten Chinese text line segmentation by clustering with distance metric learning Pattern Recognition (2009) pp. 3146 -- 3157.

[16] Vassilis Papavassiliou, Themos Stafylakis, et al..,Handwritten document image segmentation into textlines and words. Pattern Recognition (2010), pp. 369 – 377.

[17] Alireza Alaei UmapadaPal, et al..,. A new scheme for unconstrained handwritten text-line segmentation. Pattern Recognition (2011), pp. 917–928.

[18] Liwicki, M., Scherz, M., Bunke, H.: Word Extraction from On-Line Handwritten Text Lines. In: 18th International Conference on Pattern Recognition, Vol. 2, pp.929–933, (2006).

[19] Blumenstein, M., Verma, B.: A New Segmentation Algorithm for Handwritten Word Recognition. In: International Joint Conference on Neural Networks, Vol. 4, pp.2893–2898, Washington, DC , (1999).

[20] Chiang, J.-H.: Hybrid Neural Network Model in Handwritten Word Recognition. In:Neural Networks, Vol. 11(2) (1998), pp. 337–346.

[21] Gader, P., Whalen, M. ,Ganzberger, M., Hepp, D.:Handprinted Word Recognitionon a NIST Data Set. In: Machine VisionApplications, Vol. 8(1) (1995), pp. 31–40.

[22] Kurniawan, F., Khan, A. R., Mohamad, D.: Contour vsNon-Contour based Wordegmentation from Handwritten Text Lines: an experimental analysis. In: InternationalJournal of Digital Content Technology and its Applications Vol. 3(2) (2009),pp. 127–131.

[23] PayalJindal,Drbala Krishan Jindal " line and word segmentation of hand written text documents written in GURUMUKI script using mid point detection technique".Proceedings of 2015 RAECS UIET Punjab university chandigarh 21-22nddec 2015.

[24] O.Surinta,M.Holtkamp,Faik.K,J.PautV.O,L.Schomarker and

Marco.W."A*path planning for line segmentation of hand written documents"DOI 10.1109/ICFHR.2014.37.

[25] P.Karunakar,B.nayak and Nilaman.B."line and word segmentation of printed text document" IJCSIT,Vol 5(1),2014,157-160 ISSN : 0975-9646.

[26] Y.Tang,weibu,"text line segmentation based on matched filtering and top -down grouping for hand written documents" 11th IAPR,IWDAS 2014

[27] snehdeep, manojkumar,"segmentation of connected components and overlapping lines in Gurumukhi hand written documents",IJCA,(0975-8887) Vol 102-no.13,september 2014.

[28] H.S.Vishwas , Bindu,A.Thomas and C.Naveena"text line segmentation of unconstrained handwritten kannada historical script documents.

[29] sonamjain,Harawindarsinghsohal,"A Noval approach of word segmentation in correlation based OCR system",IJCA(0975-8887)Vol 99-no.18 august 2014.

[30] M.Cheal.B,Marcus.L,Rolf.I,"textline extraction using DMLP classifiers for historical manuscripts" 2013 12th ICDAR.15.20-5363/13.

[31] M.Javed,P.Nagabhushanam and .B.ChaudariExtraction of line word charecter segments directly from run length compressed printed text documents" 1403-7783.

[32] M.N.Mehdi, Aqsa Riaz" optimized word segmentation for the word based cursive hand writing recognition",IEEE 2013 european modelling symposium.978-1-4799-2578-0/3.

[33] Alireza.A,P.Nagabhusham,U.Pal,"piecewise painting technique for line segmentation of unconstrained handwritten text:a specific study with persian text documents" pattern Anal Applic (2011) 14:381-394.

[34] vijaykumar.k, AthulNegi,"fringe map based text line segmentation of printed telugu document images ",2011 ,ICDAR,1520-5363/11.

[35] W.Boussella,A.Zahour,H.Elabed,A.Benabdelhafid,and Adel Alimi"Unsupervised block covering analysis for text line segmentation of arabic ancient handwritten document images" 2010 ICRR 1051-4651/10.

[36] N.Priyanka,Srikantapal,Ranju Mandal," line and word segmentation approach for printed documents IJCA,RTIPPR 2010

[37] Fie Yin,Cheng-hi liu"Avariational Bayes method for handwritten textline segmentation" 2009,10th ICDAR 978-0-7695-2/09.

[38] G.Louloudis,B.Gatos,I.Pratikakis,C.Halatsis"Text line detection in handwritten documents"pattern recognition 41(2008)3758-3772.

[39] Fie Yin,Cheng-hi liu"Handwrittenchinese text line segmentation by clustering with distance metric learning",pattern recognition 42,(2009)3146-3157.

[40] ParthapratimRoy,U.Pal,JosephLlados,morphology based handwritten line segmentation using foreground and background information " 2008.

[41] S.Basu,C.Chaudhari,M.Kundu,M.Nasipuri,and D.K.Basu,"Text line extraction from multi-skewed handwritten documents",pattern recognition,40,2007,1825-1839

[42] Yili,Y.Zheng,DavidD,Stefan.J,"A new algorithm for detecting textline handwritten documents" 10th IWFHR oct2006.

[43] William.A.B,D.J Kennard "separating lines of text free from handwritten historical documents"2006-04-11 IEEE proceedings 2nd DIAL 8812-23 Lyon,France,April.

[44] ZhixinShi,Srirangaraj.S,VenuGovindraju "text extraction fom grayscale historical document images using adaptive local connectivity map".2004