Text Line Segmentation of Handwritten Documents in Hindi and English

[1]Sunanda Dixit
Assistant Professor, ISE Department, DSCE Bangalore, India
sunanda.bms@gmail.com

[2] Sneha
Student, ISE Department, DSCE, Bangalore, India
[2]smartiesneha@yahoo.co.in

[3]Nilotpal Utkalit
Student, ISE Department, DSCE, Bangalore, India
[3]utkalit.nilotpal@gmail.com

[4] Suresh H.N.
Department of IT
BIT, Bangalore, India
hn.suresh@rediffmail.com

***Abstract*—**Text line segmentation is a major task of handwritten document processing. In this paper we present a method to detect and segment unconstrained handwritten documents written in Hindi and English. Document image is first binarized and connected components are identified. Based on Hough lines the text lines are identified. Skew angle is determined by calculating the slope of the detected line and then the skewness is minimized. Segmentation is then performed and the result is refined by removing the noise which basically comprises components from adjacent lines.

Keywords- *Handwritten document, Hough lines, Text line segmentation, skew angle detection, connected component labeling, Hough peaks.*

*****

I. INTRODUCTION

Text line segmentation of a document image is considered as a critical stage towards unconstrained handwritten document recognition. Line segmentation is the first and the most critical pre-processing step for a document recognition, followed by word segmentation, word recognition and other indexing steps. Different types of handritten documents give arise to different types of problem.These problems might occur due to different writing styles of different people, different scripts of languages,overlapping of words,adjacent line touching,etc.

In this paper we concentrate only on the text line segmentation of handwritten documents written in Hindi and English script.The method used is common for both the scripts with slight changes in the parameters used(discussed later in detail in section3). The different methodologies used in the paper are Hough transformation, skew detection and correction and separation of adjacent touching lines.

The paper is organised as follows: Section 2 describes the problem statement. Section 3 is dedicated to literature survey. In the section 4, the proposed methodology is detailed. In section 5, we present the experimental results, and finally,Section 7 describes conclusion and future work.

II. PROBLEM

Handwritten text line segmentation and skew estimation is a critical task compare to printed text document. To improve the efficiency of OCR segmentation plays a vital role. The problem is to segment the Hindi and English handwritten text lines.

III. LITERATURE SURVEY

A Block based hough transform method is proposed for text line and word segmentation [1]. A block based hough transformation is being used which uses the gravity centers of parts of connected components to comprise a set of points of the initial image as the input and the lines that fit best to this set of points are calculated.

A statistical approach to line segmentation in handwritten documents is proposed[2]. A new technique for handwritten document segmentation is given this paper. The algorithm

ISSN: 2321-8169
733 – 739

used uses Gaussian densities and distance metrics. The algorithm works as follows, Thresholding and chain code document representation then obtain initial set of candidate lines and line drawing algorithm, in which all the lines are drawn parallely from left to right and modelled using Gaussian density. Followed by piece-wise projection profile if available.
A Natural Learning Algorithm based on Hough Transform for Text Lines Extraction [3] is proposed by Yao. A natural laerning alorithm refers to method which is similar to the learning procedure of human being is being used in the algorithm. Following are the steps in the natural learning algorithm used, Initially Hough Transform is applied to the minima points of the connected components in a small strip of the image to get initial lines.These minima points will be put into the existing cluster,or in a pool using moving window cluster.Then Hough transform will be applied to generate new clusters for the points in the pool. The above steps are till small strips of images are obtained. The clusters obtained initially are considered as initial clusters and the algorithm is repeated till the end of cluster.

Text extraction from grey scale historical document images using adaptive local connectivity map [ALCM] is proposed in [4]. The method emphasises on reducing the scale of the image for text line detection. This reduced scale is termed as grey scale. On a grey scale the line patterns appear distinct and the touching between lines loses prominence.
In this method each pixel value at a pixel location represents a connectivity property of its neighboring pixels in the original document image. Then by using fuzzy runlength technique these images are binarised. The algorithm is designed for complex historical documents. It is also general enough to be used for any type of images such as binary images,machine printed or even mixed script.
Text line extraction in handwritten document with Kalman filter applied on low resolution image is proposed by A.Lemaitre [5]. The proposed method is on low resolution image based on theory of Kalman filtering. Kalman filtering is based on a notion of perceptive vision, which says that at a certain distance, text lines of documents can be seen as line segments. It works on global vision, which makes it possible to decrease noise around lines. This method is proposed to work on grey level image where a segment is defined as a succession of connected run-lengths (a set of connected black pixels within a column), which have approximately the same thickness and from which the middle points of the run-lengths are on a line segment. This method makes it possible to deal with difficulties met in ancient damaged documents.

Detecting text lines in handwritten documents Using gaussian window is proposed by [6]. In this model a script independent text line detection which detects an image segmentation problem by enhancing text line structure using a Gaussian window, and adopting a level set method to evolve text line boundaries.

The algorithm works as follows: Firstly the text line structure is enhanced by blurring with a Gaussian window,followed by conversion of binary image to grey scale.Then an initial estimate of text line boundaries is estimated using the level set method.

Text line extraction from multi-skewed handwritten documents of English and Bengali is proposed by S.Basu [7]. This method involves an assumption that hypothetical water flows,from both left and right sides of the image frame,face obstruction from characters of text lines.The stripes of areas left unwetted on image frame are finally labelled for extraction of text lines.
In this hypothetically assumed situation water flowing across the image frame does not wet the areas which face obstruction from the characters of the text lines.These are referred as unwetted

ISSN: 2321-8169
733 – 739

stripes.Once labellings of document images into stripes is done,the entire image is divided into two different types of stripes,one containing text lines and the other containing line spacings. This is how the technique works for multi-skewed handwritten document images.

[8] proposed an approach based on minimum spanning tree (MST) clustering. First,the connected components of the document images are grouped into a tree by MST clustering. The edgesof the tree are then dynamically cut to form text lines by using a new objective function for finding the number of clusters. The reduced hypervolume is used as a criteria for finding the edge to cut and use a novel objective function based on the characteristics of documents to determine the final number of clusters. The minimum spanning tree clustering (MST) algorithm is known to be computationally efficient and capable of detecting clusters with irregular boundaries.This approach is free of artificial parameters,and can apply to various documents of multi-skewed and curved text lines.

Handwritten text line segmentation by clustering with distance metric learning is proposed by Fei Yin [9]. The proposed method uses effective bottom up algorithm based on Minimal Spanning Tree (MST) clustering with distance metric learning. The connected components of document image are grouped into a tree structure based on any given distance metric. The text lines are extracted by dynamically cutting the edges of a tree using a new objective function.This algorithm does not require any external parameters. The minimal spanning tree (MST) algorithm is suitable for clustering the connected components into text lines, because it is computationally efficient and capable of detecting clusters with irregular boundaries.

Morphology Based Handwritten Line Segmentation Using Foreground and Background Informationis proposed by PP.Roy [10]. Morphological operation and run-length smearing algorithm (RLSA) is used to segment individual text lines from unconstrained handwritten document images.This RLSA is firstly applied to individual word as a component. Then the foreground portion of the smoothed image is eroded to get some seed components from the individual words of the document. Finally,using the positional information of the seed components and the boundary information, the lines are segmented.

IV. PROPOSED METHOD

Our proposed methodology consists of four steps. (1) Image Binarization and classification. (2) Hough Transform mapping and detecting hough lines. (3) Skew detection and correction. (4) Separation of components belonging to adjacent line.

4.1. IMAGE BINARIZATION AND CLASSIFICATION

In this phase image is first converted to two dimensional matrix. i.e. to gray scale. We can then detect the edge and finally the connected components are identified. Mean height (MH) of connected component and average width (AW) are computed. Based on a certain threshold the noise in the document are identified and removed. We then classify the document into two different subsets, one consisting of all the small characters such as punctuation marks and accents. The other consists of the majority of the characters of the document. The average height and width of the connected component belonging to the second subset is calculated. These calculation forms the basis of identification of our desired regions.

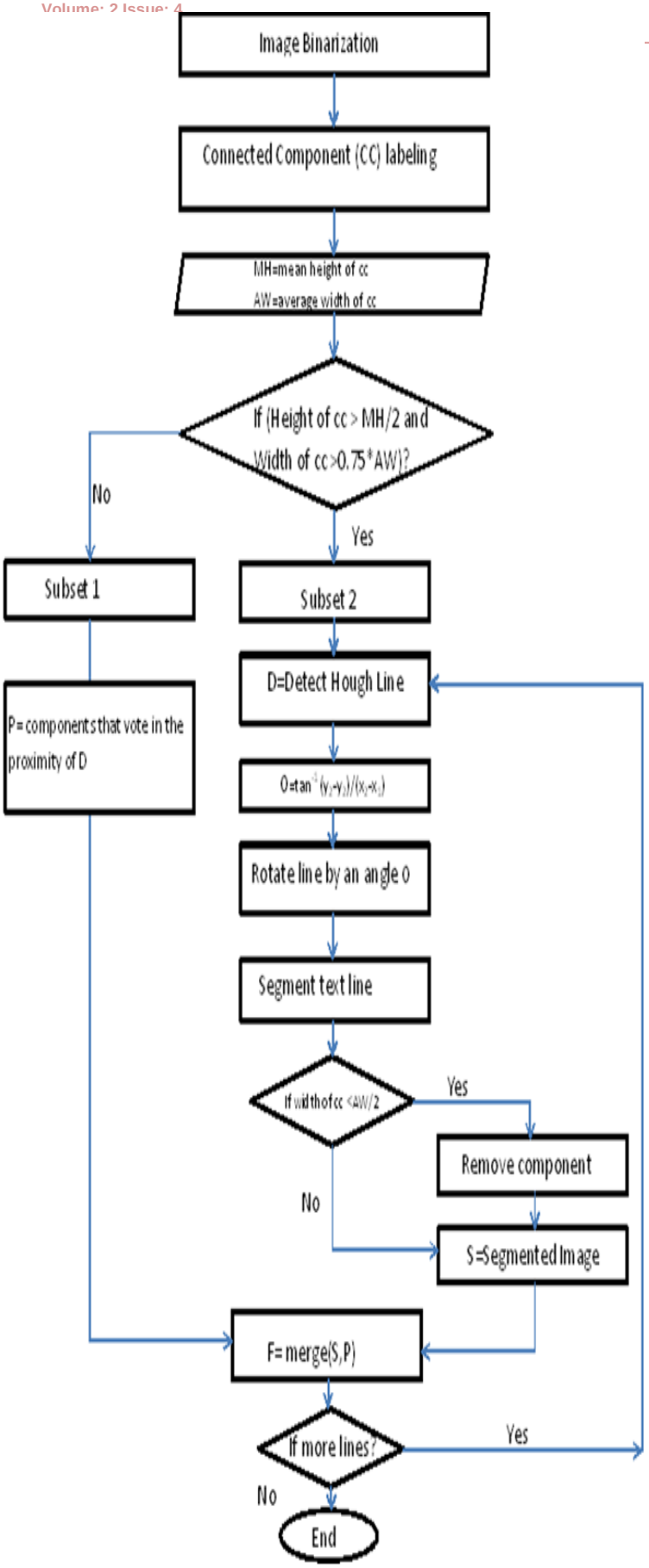

Fig 1. Flow diagram of proposed method.

4.2. DETECTING HOUGH LINES

A hough transform is a process of finding the value of rho, theta and an accumulator array. These rho and theta value defines a line in the Cartesian space which is identified by:

$$\rho = x \cos \Theta + y \sin \Theta$$

We compute hough transform matrix for subset two of the document. Based on this matrix, hough peaks are computed, hough peaks corresponds to the peaks in the computed hough matrix. These peak values allows us to detect line segments corresponding to the connected component. We then compute the hough lines based on certain threshold and fill the gaps between line segments. In this way we are able to detect the text lines in the subset. The variation of the slope of this line with respect to a reference line gives us the measure of skewness of the line.

4.3. SKEW DETECTION AND CORRECTION

As said earlier a reference line is chosen, for our case it is a straight line originating from the detected hough line and having a slope of '0'. Since the slope of our reference line is '0', the slope of the detected line itself gives the measure of the skewness. After detecting the skewness it is corrected by aligning the deviation from the reference line. Although different methods of skew detection could be used, it produces similar results.

4.4. SEPERATION OF COMPONENTS BELONGING TO ADJACENT LINE

The points that map to 1.2*MH above the line and 1.1*MH below the line forms the desired segment for English documents whereas for Hindi documents the point mapping to 1.1*MH above the line and 1.3*MH below the line are considered. All the connected components that maps to this region are not considered for the next iteration. The small characters that includes punctuation marks and accents are assigned to the text line if they vote to this region. All

 ISSN: 2321-8169
733 – 739

component that has length greater than AW/2 and negligible height with respect to MH are considered components from next line and are removed. This process is repeated unless no more line is detected.

V. RESULTS AND DISCUSSION

For the purpose of evaluation, lines are considered to be detected when it is successfully able to detect the punctuation like characters and also components from the adjacent lines are successfully separated, or is minimal in case the lines are very close to each other.

Proposed method when applied to documents which are mostly written in block letters and also where noise is well above practical levels produces an accuracy of about 94%.

For well written handwritten document containing significant level of noise the accuracy is more than 97%. For Hindi documents the accuracy is between 93% - 97% Further the algorithm is able to detect lines which are multi skewed i.e. where the text lines are skewed both in the clockwise and anti-clockwise direction.

VI. CONCLUSION

In this paper, we presented a text line segmentation approach based on the Hough line. The algorithm is valid for a wide variety of skewed as well as non-skewed document. We use connected component analysis to identify the small characters and normal characters in the document and also to calculate the average character length and width. The experimental results indicate the validation of the approach. Further research includes using the method for text line segmentation of decorated and calligraphic documents.

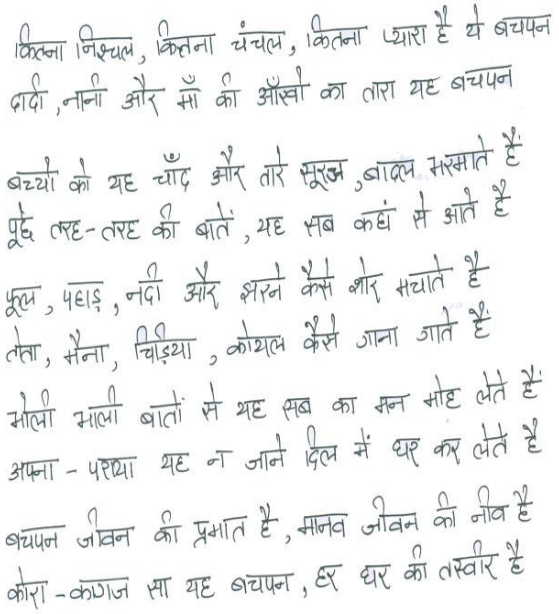

Fig2. (a)

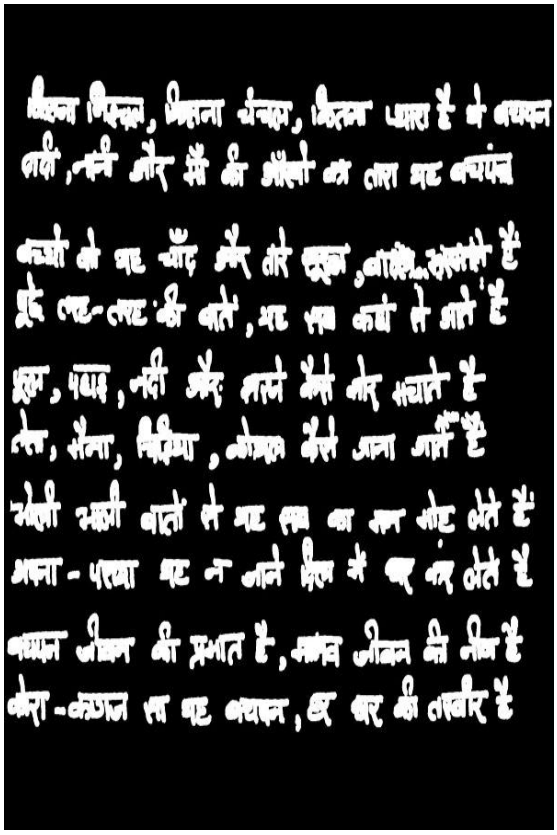

Fig2. (b)

ISSN: 2321-8169
733 – 739

Fig2 (c)

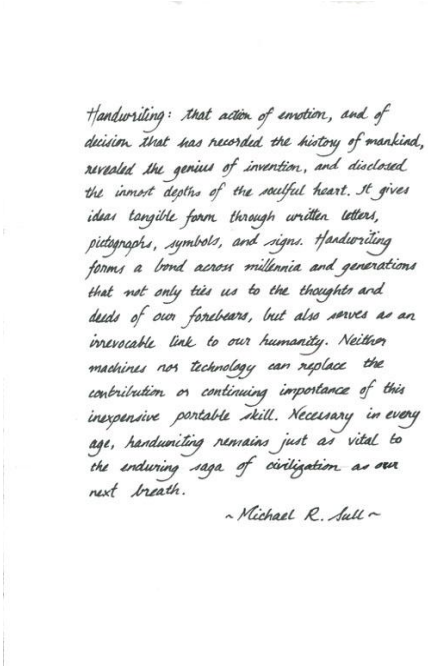

Fig2 (d)

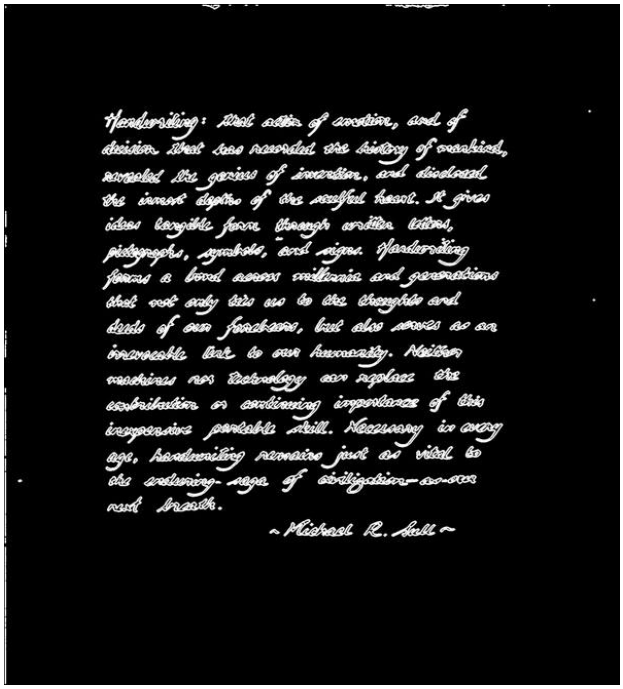

Fig2 (e)

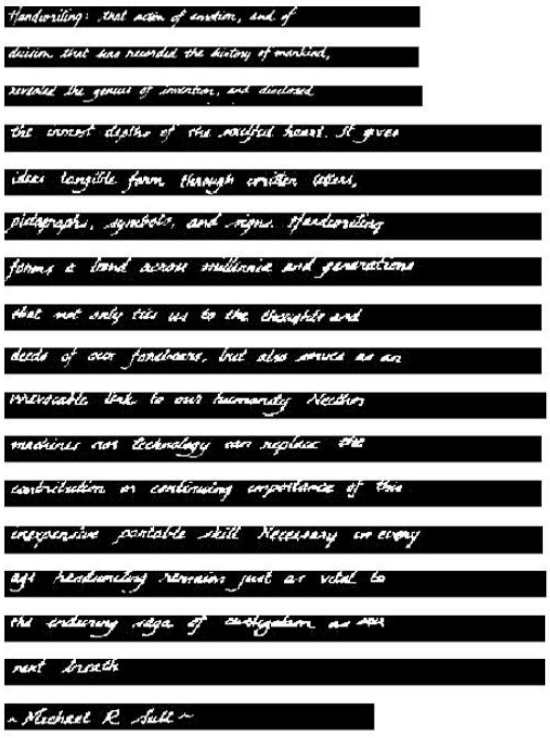

Fig2 (f)

Fig 2. *(a) Hindi handwritten document; (b) Hindi document after binarization; (c) Final segmented result of Hindi document; (d) English handwritten document; (e) English document after binarization; (f) Final segmented result of English document.*

REFERENCES

[1] G.Louloudis, B.Gatos, I.Pratikakis, C.Halatsis, Pattern Recognition, 2009.

[2] Manivannam Arivazhagan, Harsh Srinivasan, Sargur Srihari ,A statistical approach to line segmentation in handwritten documents.

[3] Yao Pu,Zhixin Shi, A Natural Learning Algorithm based on Hough Transform for Text Lines Extraction in Handwritten Documents,in: Proceedings of the 6th International Workshop on Frontiers in Handwriting Recognition, Korea,1998,pp637-646.

[4] Zhixin Shi,Srirangaraj Setlur,Venn Govindaraju, Text Extraction from Grey Scale Historical Document Images Using Adaptive Local Connectivity Map,in: The Second International Conference on Document Analysis and Recognition, Seoul,Korea,2005,pp.794-798.

[5] A.Lemaitre,J.Camillerapp,Text Line Extraction in handwritten document with Kalman filter applied on low resolution image, in: The Second International Conference on Document Image Analysis for libraries(DIAL'06), pp.12-23

[6] Yi Li,Yefeng Zheng,David Doermann, Detecting text lines in handwritten documents, in:The 18th International Conference on Pattern Recognition(ICPR'06)pp. 1030-1033.

[7] S.Basu,C.Chaudhuri,M.Kundu, M.Nasipuri,D.K.Basu,Text line extraction from multi-skewed handwritten documents,Pattern Recognition 40(6)(2007)1825-1839.

[8] Fei Yin,Cheng-Lin Liu, Handwritten Text Line Extraction based on mimimum spanning tree clustering, in:International Conference on Wavelet Analysis and Pattern Recognition, Beijing,China,November 2007, pp.1123-1128.

[9] Fei Yin,Cheng-Lin Liu,Handwritten text line segmentation by clustering with distance metric learning, in: International Conference on Frontiers in HandwritinG Recognition (ICFHR'08), Montreal,Canada, August 2008,pp. 229-234.

[10] Partha Pratim Roy, Umapada Pal, Josep Llado's, Morphology Based Handwritten Line Segmentation Using Foreground and Background Information, in:International Conference on Frontiers in Handwriting Recognition (ICFHR'08), Montreal, Canada, August 2008, pp. 241-246.