IETE Journal of Research

ISSN: 0377-2063 (Print) 0974-780X (Online) Journal homepage: http://www.tandfonline.com/loi/tijr20

A Novel Segmentation Methodology for Cursive Handwritten Documents

Subhash Panwar & Neeta Nain

To cite this article: Subhash Panwar & Neeta Nain (2014) A Novel Segmentation Methodology for Cursive Handwritten Documents, IETE Journal of Research, 60:6, 432-439, DOI: 10.1080/03772063.2014.963174

To link to this article: http://dx.doi.org/10.1080/03772063.2014.963174

Published online: 15 Oct 2014.

A Novel Segmentation Methodology for Cursive Handwritten Documents

Subhash Panwar and Neeta Nain

Department of Computer Science and Engineering, Malaviya National Institute of Technology, Jaipur 302017, India

ABSTRACT

Segmentation in handwritten documents is a very challenging task because in handwritten documents curved text lines and non-uniform skews appear frequently. Basically segmentation means to analyse the document image into its sub component as text line, words, or ligatures and finally character. Various handwritten text recognition systems take segmented character as input to recognize them and some recognition systems which are based on holistic approaches use word level segmentation. In this paper, we have implemented a complete framework of segmentation for handwritten documents with various languages. A novel connectivity strength parameter is used for deciding the groups of the components which belong to the same line. Over-segmentation is also removed with the help of depth first search approach and iterative use of the connectivity strength function. We have implemented and tested this approach with English, Hindi, and Urdu text images taken from benchmark database and find that it is a language adaptive approach which provides encouraging results. The average accuracy of the proposed technique is 97.30%.

Keywords:
Connectivity strength function, Cursive handwritten text, Depth first search, Segmentation.

1. INTRODUCTION

Text segmentation of cursive handwritten documents is much more difficult than that of printed documents because lines appearing in handwritten documents are often non-uniformly skewed and curved. Moreover, the spaces between text lines are not often the same compared to the spaces between within-line characters and also some text lines may interfere with each other. Hill and dale writing styles are also a problem for segmentation. Therefore many text line detection techniques, such as projection analysis [1,2], Hough transform [3], and K-nearest neighbour connected components (CCs) grouping [4], are not able to segment handwritten text successfully. Many approaches are used only for specific languages, so still a uniform approach to handle all kinds of challenges and various language scripts is not available.

Figure 1 shows an example of unconstrained handwritten document with various segmentation challenges.

Text document image segmentation can be roughly categorized into three classes: top-down, bottom-up, and hybrid. Top-down methods partition the document image recursively into text regions, text lines, text words, and text characters. These methods always assume that all text are in straight lines. Bottom-up methods group small units of image (pixels, CCs, characters, words, etc.) into text lines and then text regions. Bottom-up grouping can be viewed as a clustering process, which aggregates text image components according to proximity and does not rely on the assumption of straight lines. Hybrid methods combine bottom-up grouping and top-down partitioning in different ways.

All the three approaches have their advantages and disadvantages. Top-down methods work well for typed text where the text lines are relatively horizontal, but it does not perform well on curved and overlapping text lines. The performance of bottom-up grouping relies on some heuristic rules or artificial parameters, such as the between-component distance metric for clustering. On the other hand, hybrid methods are complicated in computation, and the design of a robust combination scheme is non-trivial.

In graph representation of image each component is represented as vertex and the distance calculated between the CCs is represented as an edge with weight. Then, we may find out the minimum spanning tree of the given image, and thus the segmentation is made by comparing with predetermined distance which may be an inter-word distance or intra-word distance [5].

We deal with the problem of touching components and hill and dale styles using graph representation of document image. After finding the CCs we group them according to the sequence of appearance in document

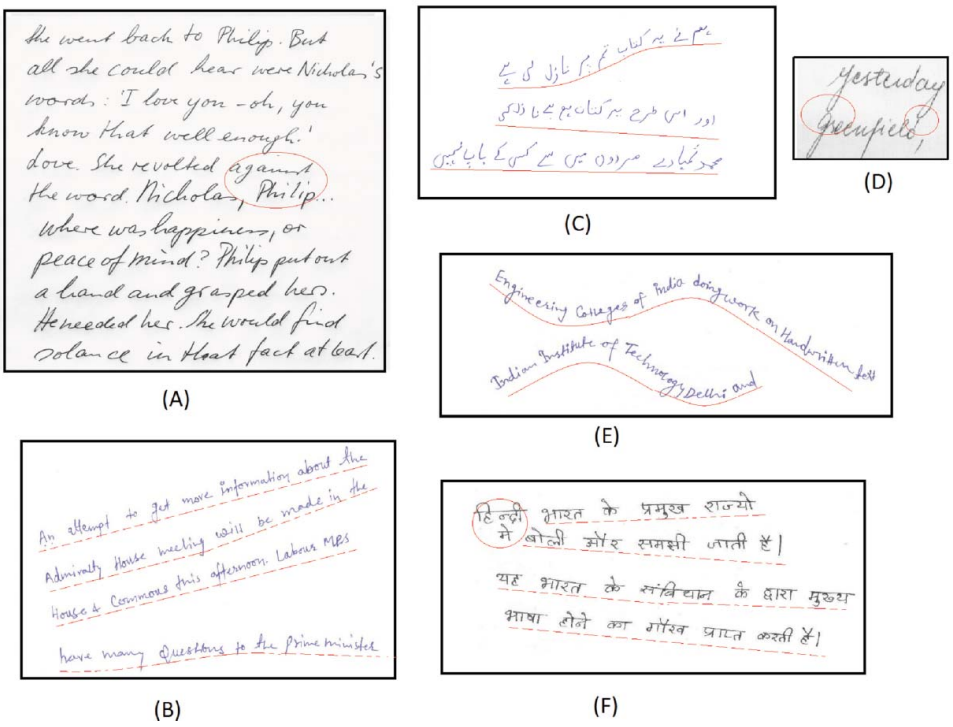

Figure 1: Various handwritten document samples of different styles which explain various segmentation challenges: (A) sample image from IAM data set having overlapped components in red circle, (B) multiple skewed document image, (C) curved text line in Urdu handwritten images, (D) touching lines shown in red circle, (E) hill and dale writing style, (F) multi-skewed and overlapped text line in Hindi text document.

using the proposed connectivity strength parameter and depth first search (DFS) approach.

We make the following contributions:

1. We calculate the connectivity strength of two CCs and the components are connected if they have maximum strength.
2. The segmented components are sequenced using DFS approach and connectivity strength according to appearance in document.
3. Touching components are separated using average vertical width estimation.

In this paper, we discuss extensive quantitative comparison experiments on a large handwriting data set using three scripts as English, Hindi, and Urdu.

This paper is organized as follows. In Section 2, we review some related materials. Connectivity strength parameter and component sequencing using the DFS approach are described in Sections 3 and 4, respectively. Section 5 presents an extensive performance evaluation and quantitative comparison of the proposed method and the previous methods on a large handwriting data set. This paper ends with a summary and a brief discussion of our future work.

2. RELATED WORK

Survey [2] on text segmentation can be categorized according to various methods such as projection profile-based methods, Hough transform methods, grouping methods, methods for processing overlapping and touching components, smearing methods, and other methods [2,6]. Many different methods can be employed for text line segmentation. As the related work discussed in [7] the projection-based method is one of the mostly used approaches for printed documents. It uses the gap between two neighbouring text lines and segments the image from the gap. Such approaches which use projection profile are not appropriate for cursive and unconstrained handwritten text documents because gaps between two cursive text lines are not significant. In [8], authors divide the image into multiple columns and find the projection profile for each column and combine the results of adjacent columns but results of two adjacent columns are ambiguous in the presence of curved lines. So we conclude that projection-based methods are only suitable for printed and straight text and top-down approaches also have the disadvantage that they cannot process hill and dale writing styles. Another approach called the Hough transformation is comfortable with little amount of curved text but unable to handle multiple skewed text.

Smearing methods also have their benefits. They are efficient and computationally inexpensive. In smearing method the continuous foreground pixels along the horizontal direction are filled by black pixels and fill the area of black pixels is formed [5,9]. In this approach also it is assumed that the distance threshold between words is predefined. Hence, the smearing approach groups the text lines but needs some predefined threshold.

The performance of CC-based methods on a handwritten document is improved by estimating the local orientation of the text line and using it to guide the merging of CCs [4].

In [10] authors use the image meshing for line detection locally in the presence of multi-orientation of lines. Wigner-Ville distribution and projection histogram are used to determine the local orientation. This local orientation is then enlarged to limit the orientation in the neighbourhood.

In [11] the text line is segmented using affinity propagation. They first estimate the local orientation at each primary component to build a sparse similarity graph and then use a shortest path algorithm to compute similarities between non-neighbouring components. Affinity propagation and breadth first search are used to obtain coarse text lines.

In [12], the line segmentation algorithm is based on locating the optimal succession of text and gap areas within vertical zones by applying Viterbi algorithm and a text line separator drawing technique is applied and finally the CCs are assigned to text lines.

Table 1 presents the existing approaches and a performance evaluation of the proposed algorithm for segmentation.

We are proposing an effective bottom-up grouping method for text line segmentation for unconstrained handwritten text documents. Our approach is based on grouping of CCs, connectivity strength function (CSF), and DFS for finding the exact sequence of components according to the appearance in document.

3. CONNECTIVITY STRENGTH FUNCTION AND ITS SIGNIFICANCE

In this section, we describe the motivation for estimating a connectivity strength parameter for the CC of document image and describe the significance of the parameter.

The CSF is derived as as follows: let there be two connected components C_1 and C_2 having centroids as (x_1, y_1) and (x_2, y_2), respectively. The minimum distance (d) between the two components is

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

and the vertical distance Y_d is

$$y_d = (y_2 - y_1)$$

then the *CSF* is defined as

$$\text{CSF} = \frac{|d - y_d|}{y_d}$$

For each pair of CC, we compute the value of CSF.The decision for grouping the components depends on CSF as

$$\text{CSF} = \begin{cases} 0 & \text{belongs to different lines} \\ \infty & \text{belongs to same line} \end{cases}$$

where $\text{CSF} = 0$, only when $d = y_d$, which means the two components have the minimum connectivity strength as they are orthogonal and hence belong to different lines, which are almost parallel. And the $\text{CSF} = \infty$, only when $y_d = 0$. This means the

Table 1: Evaluation of line segmentation algorithms

Author	Merits	Demerits	Accuracy (in %)
Vikas et al. [13]	High accuracy for printed text	Unable to handle skewed lines	100%
Satadal Saha et al. [14]	Less computational complexity	Fails to segment closely spaced lines	87.5%
Malakar S. et al. [15]	Handles skewed text to some extent	Incorrect segmentation of document with uneven skew	89.35%
Fei Yin et al. [5]	Robust to handle various documents with multi-skewed and curved text lines	Document with overlapping of two neighbouring text lines	98.02%
Proposed approach	Robust to handle curved, uneven, and multi-skewed documents and touching components	Fails to handle completely overlapped text lines	98.92%

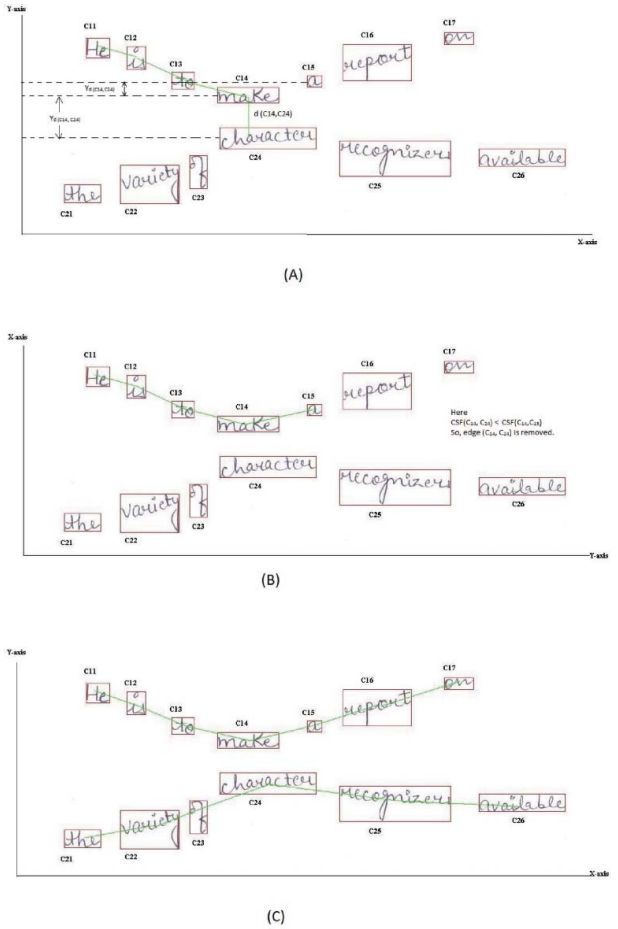

Figure 2: (A) Mis-qualified linkage between two lines. (B) Removal of mis-qualified linkage using CSF. (C) Complete forest of components using CSF.

connectivity between the two components is the strongest as they both belong to the same line. The angle between the two components is zero aligning them on the same line as shown in Figure 2.

For example consider Figure 2, which illustrates how the mis-qualified words linkage is removed using CSF. For removing such connections (edges) from the traversing sequence of DFS, we further use a CSF as explained below which is very useful in deciding the groups of the components which belong to the same line. The strokes labelled with C_{11}, C_{12}, C_{13}, C_{14}, C_{15}, C_{16}, and C_{17} belong to a single text line and strokes labelled with C_{21}, C_{22}, C_{23}, C_{24}, C_{25}, and C_{26} belong to different text lines. Using DFS of sorted list of CCs the component C_{24} appears after the component C_{14} as shown in Figure 2(A) in mis-aligned link.After applying CSF between its neighbour components we found that strength of C_{14} and C_{24} is less than the strength between C_{14} and C_{15}. So we remove the link between C_{14} and C_{24} and make connection between C_{14} and C_{15}. The final result of stroke segmentation is shown in Figure 2(C).

4. PROPOSED SEGMENTATION APPROACH

The proposed algorithm proceeds by computing the bounding box and centroid for the 8-connected components of the document image. Due to writer variations or word formations we detect some overlapped or disjoint components. Thus, we correct overlapped bounding boxes by using the extreme coordinates and recalculate its centroid and extreme coordinates. Next, we find the cost metric C where $c(i,j)$ is the distance between two connected components CC_i and CC_j. We calculate the cost using Eq. (1).

$$c(i,j) = \begin{cases} |x_i - x_j| & \text{if } CC_i \,\&\, CC_j \text{ are horizontally parallel} \\ |y_i - y_j| & \text{if } CC_i \,\&\, CC_j \text{ are vertically parallel} \\ \sqrt{|x_i - x_j|^2 + |y_i - y_j|^2} & \text{otherwise} \end{cases} \quad (1)$$

After finding the cost matrix, we calculate the minimum possible CSF which is used as the threshold value Th_{CSF}. Where ever two components have CSF below the threshold Th_{CSF}, remove the connection between the components.

Finally we apply DFS on the sorted vector V finding the sequence vector V_{seq} containing exact stroke sequence as appearing in the document image.

Figure 3 shows the complete block diagram of the proposed approach. After applying the proposed approach on Figure 1 we remove the mis-aligned components from the text lines and generate the forest of given document image as shown in Figure 4, where our forest is defined as a group of trees and where every tree is a text line.

Figure 5(a) shows an example of hindi handwritten document. Figure 5(b) shows the result using proposed approach. The complete process can be enumerated as shown in Algorithm 3

5. EXPERIMENTAL RESULTS

The experiments are done on various varieties of handwritten document images including different languages as Hindi, English, and Urdu. To cover all the cases such as skewed lines, curved lines, and touching lines, some images are randomly selected from the large IAM [16] database of handwritten documents and some are generated from different writers with different languages. The various cases are enumerated below.

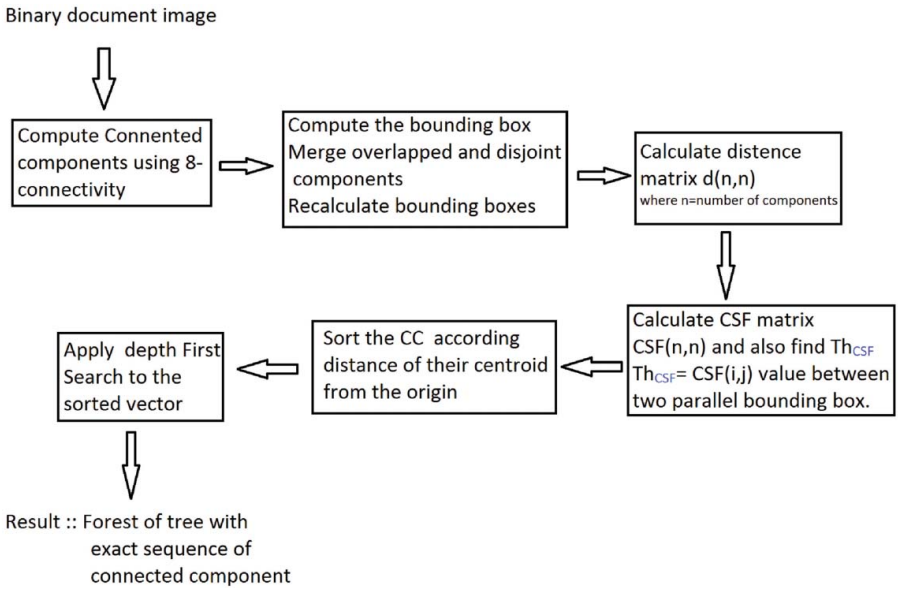

Figure 3: Block diagram of the proposed segmentation approach.

1. Curved lines: Figure 6(A) shows an example of curved handwritten lines. The projection profile techniques [1] for such curved lines fail completely in such line segmentation. Also Figure 6(B) shows the result of the proposed approach.

   Figure 7 shows an example of Urdu handwritten document with curved lines and also shows the result using the proposed approach.
2. Skewed lines: an example image of handwritten document with skewed lines is shown in Figure 8(A); we apply the CSF and find the exact forest of the text line which is shown in Figure 8(B). Here also the traditional methods would fail. Again it is ascertained that the CSF improves the accuracy of line segmentation in presence of skewed line in the documents.
3. Touching lines: to overcome the touching line problems, first we find the average height of the components and then we find the components having height more than twice the average height. And just break the component into two components as dividing by half.

The experimental results of the proposed line segmentation approach show that proposed CSF improves line segmentation accuracy significantly in all the cases. The proposed method was also compared with other state-of-the-art methods in experiments on a large database of IAM [16], handwritten documents data set and its superiority were demonstrated. The accuracy rate

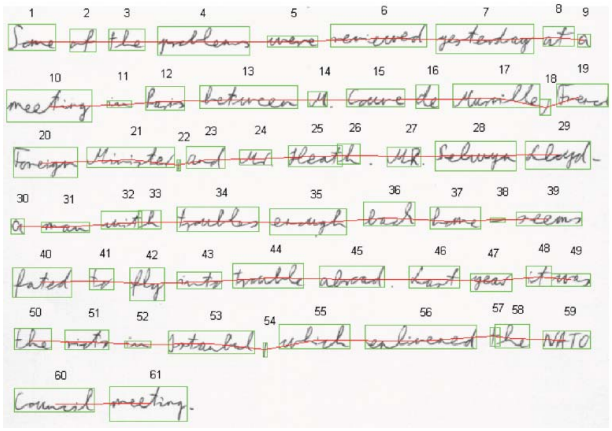

Figure 4: Forest with sequence of components for Figure 1 using the proposed approach.

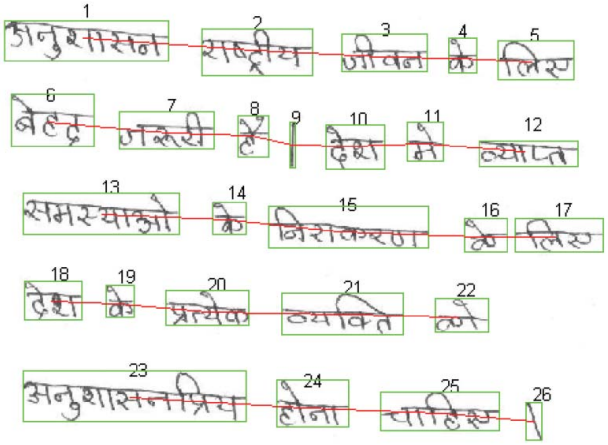

Figure 5: Example image of a Hindi handwritten document with forest and its exact CC sequence after applying proposed CSF.

Algorithm 1 Text segmentation.

Ensure: V_{seq} – sequence of words appearing in document.
Require: I – text document binarized image with background as 0.
Compute connected components (CC_i)s using 8-connectivity.
Compute centroids for all CC_is.
Compute cost matrix $c(n\times n)$ of CCs

$$c(i,j) = \begin{cases} |x_i - x_j| & \text{if } CC_i \ \& \ CC_j \text{ are horizontally parallel} \\ |y_i - y_j| & \text{if } CC_i \ \& \ CC_j \text{ are vertically parallel} \\ \sqrt{|x_i - x_j|^2 + |y_i - y_j|^2} & \text{otherwise} \end{cases} \quad (2)$$

Compute bounding box around each CC and find extreme coordinates $X_{min}, X_{max}, Y_{min}, Y_{max}$ of the corresponding CC_i.
Perform merging for overlapping and disjoint components and recalculate the CC_s as the correction causes changes in centroid and coordinates of CC.
Recalculate the bounding box and its respective cost matrix $c(n\times n)$
Compute the $CSF(n\times n)$
{where $CSF_{i,j} = \frac{|c_{i,j} - y_c|}{y_c}$ and $y_c = (y_j - y_i)$.}
Calculate Th_{CSF}
{Th_{CSF} = minimum possible CSF between two vertically aligned components}.
Sort vector V contains labels of CC, sorted according to the distance of their centroid from the origin.
$\forall$ CC_i in V
{ Compute V_{seq} using depth first search DFS and $CSF(n\times n)$ }
Let i be the initial CC_i in V, such that, $i \notin V_{seq}$
j be the CC with maximum value of $CSF(i,j)$, such that, $j \notin V_{seq}$
$CSF(i,j) < Th_{CSF}$
add j to V_{seq} containing i.
$[V_{seq}]$

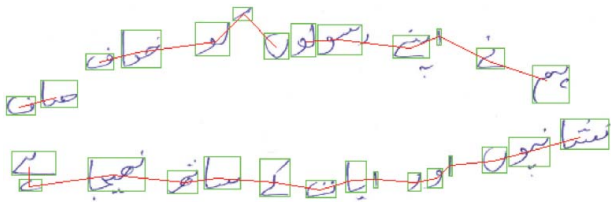

Figure 6: Example image of curved lines in handwritten text document and respective forest remains after applying CSF.

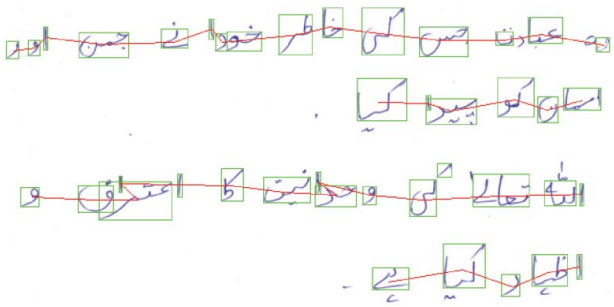

Figure 7: Example image of an Urdu handwritten document with forest remains after applying CSF.

of the proposed text line segmentation method is summarized in Table 2.

Our proposed approach achieves an average accuracy of 97.30%. We are able to handle skewed and curved lines successfully, compared to existing approaches as proposed by Vikas [13] using projection profile, which is unable to handle skewed lines; Satdal [14] used Hough transformation, which fails to segment closely spaced lines; Malakar [15] used run length smearing

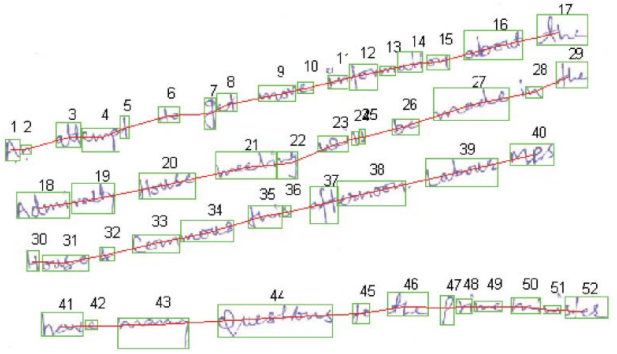

Figure 8: (A) Example image of skewed lines in handwritten text document. (B) Forest remains after applying CSF.

Table 2: Accuracy rate of proposed text line segmentation using CSF

Line types	Total no. of lines	Accurate detected lines	Accuracy rate
Printed lines	320	320	100
Skewed lines	2600	2520	96.92
Curved lines	1750	1670	95.42

approach, which could not handle documents with multi-skew. Fei yin [5] used clustering approach, which could not handle overlapping and closely spaced lines, while we can handle them by segmenting closely or overlapped lines at the midpoint of the height of the CC when the height of the CC is more than double the height of the average component.

6. CONCLUSIONS

In this paper, a language adaptive approach for handwritten text line segmentation with CSF has been presented and applied on Hindi, Urdu, and English language. We have used the IAM dataset for English handwritten documents and other documents collected from different writers with different languages as Hindi, English, and Urdu. The proposed text line segmentation approach with the novel use of CSF has the advantage of language adaptivity with highly curved and skewed text lines. From the experiments, 97.30% average accuracy was observed in the system. The results obtained by this segmentation is a forest of lines with exact sequence of words. It shows that the proposed system is capable of locating accurately the text lines in images and documents. We have also proposed a heuristic to handle touching lines.

REFERENCES

[1] F. Zamora-Martinez, M. J. Castro-Bleda, S. Espaa-Boquera, and J. Gorbe-Moya, "Unconstrained offline handwriting recognition using connectionist character N-grams," in *The 2010 International Joint Conference on Neural Networks (IJCNN)*, Barcelona, Spain, Jul. 18−23, 2010, pp. 1−7.

[2] L. Likforman Sulem, A. Zahour, and B. Taconet, "Text line segmentation of historical documents: A survey," *Int. J. Doc. Anal. Recognit.*, Vol. 9, pp. 123−38, Apr. 2007.

[3] L. Likforman-Sulem, A. Hanimyan, and C. Faure, "A Hough based algorithm for extracting text lines in handwritten documents," in *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, Montreal, Canada, 1995, pp. 774−7.

[4] M. Kumar, M. K. Jindal, and R. K. Sharma, "K-nearest neighbour based offline handwritten Gurumukhi character recognition," in *International IEEE Conference on Image Information Processing (ICHP 2011)*, Himachal Pradesh, India, 2011, pp. 7−11.

[5] F. Yin, and C-L. Liu, "Handwritten Chinese text line segmentation by clustering with distance metric learning," *Pattern Recognit.*, Vol. 42, no. 12, pp 3146−57, Dec. 2009.

[6] A. Amin, and S. Wu, "Robust skew detection in mixed text/graphics documents," in *Proceedings of the 8th ICDAR*, Seoul, Korea, Aug. 29 to Sep. 1, 2005, pp. 247−51.

[7] Y. Li, Y. Zheng, D. Doermann, and S. Jaeger, "Script-independent text line segmentation in freestyle handwritten documents," *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 30, no. 8, pp. 1313−29, Aug. 2008.

[8] N. Tripathy, and U. Pal, "Handwriting segmentation of unconstrained Oriya text," in *Proceedings of the Ninth International Workshop Frontiers in Handwriting Recognition*, Tokyo, Japan, 2004, pp. 306−11.

[9] Z. Razak, K. Zulkiflee, et al., "Off-line handwriting text line segmentation: A review," *Int. J. Comput. Sci. Netw. Security (IJCSNS)*, Vol. 8, no. 7, 12−20, 2008.

[10] N. Ouwayed, and A. Belaid, "A general approach for multi-oriented text line extraction of handwritten documents," *Int. J. Doc. Anal. Recognit.*, Vol. 15, no. 4, pp. 297−314, Sept. 2012.

[11] J. Kumar, et al., "Handwritten Arabic text line segmentation using affinity propagation," *Proceedings of DAS10*, Boston, MA, 2010, pp. 135−42.

[12] V. Papavassiliou, et al., "Handwritten document image segmentation into text lines and words," *Pattern Recognit.*, Vol. 43, pp. 369−77, June 2010.

[13] V. J. Dongre, and V. H. Mankar, "Segmentation of printed Devanagari documents," in *Springer Proceedings, 1st International Conference on Communications in Computer and Information Science*, Vol. 198, Part 1, *ACITY*, Chennai, India, Jul. 2011, pp. 211−218.

[14] S. Saha, S. Basu, M. Nasipuri, and D. K. Basu, "A Hough transform based technique for text segmentation," *J. Comput.*, Vol. 2, no. 2, 13414, Feb. 2010.

[15] S. Malakar, S. Halder, R. Sarkar, N. Das, S. Basu, and M. Nasipuri, "Text line extraction from handwritten document pages using spiral run length smearing algorithm," in *2012 International Conference on Communications, Devices and Intelligent Systems (CODIS)*, Kolkata, India, Dec. 28−29, 2012, pp. 616−9.

[16] U. Marti, and H. Bunke, "The IAM-database: An English sentence database for off-line handwriting recognition." *Int. J. Doc. Anal. Recognit.*, Vol. 5, pp. 39−46, Nov. 2002.

Authors

Subhash Panwar received his BTech degree in computer engineering from University of Rajasthan, Jaipur, India in 2004, MTech degree in computer engineering from Motilal Nehru National Institute of Technology, Allahabad, India, in 2010, and pursuing PhD degree from the Malaviya National Institute of Technology, Jaipur, India. He is working in the Government Engineering College Bikaner, Rajasthan, India as an assistant professor from September 2004. His current research interests include image fusion, computer vision, computational intelligence and pattern recognition.

E-mail: panwar.subhash@gmail.com

Neeta Nain is an assistant professor in the Department of Computer Science and Engineering, Malaviya National Institute of Technology, Jaipur. She has more than two decades of teaching experience. Her research area is image processing, pattern recognition and computer graphics. Presently she is guiding research in and written text recognition, corpus development, and crowd analysis. She has published a number of articles in international journals and conferences.

E-mail: neetanain@yahoo.com

DOI: 10.1080/03772063.2014.963174;