Egyptian Informatics Journal (2016) 17, 189–197

Cairo University

Egyptian Informatics Journal

www.elsevier.com/locate/eij
www.sciencedirect.com

Text segmentation in degraded historical document images

A.S. Kavitha [a], P. Shivakumara [b,*], G.H. Kumar [a], Tong Lu [c]

[a] Department of Studies in Computer Science, University of Mysore, Karnataka, India
[b] Faculty of Computer Science and Information Technology, University Of Malaya, B-2-18, Malaysia
[c] National Key Lab for Novel Software Technology, Nanjing University, Nanjing, China

Received 2 March 2015; revised 1 October 2015; accepted 6 November 2015
Available online 2 January 2016

KEYWORDS

Text enhancement;
Sobel and Laplacian operations;
Indus document;
Clustering;
Text line segmentation

Abstract Text segmentation from degraded Historical Indus script images helps Optical Character Recognizer (OCR) to achieve good recognition rates for Hindus scripts; however, it is challenging due to complex background in such images. In this paper, we present a new method for segmenting text and non-text in Indus documents based on the fact that text components are less cursive compared to non-text ones. To achieve this, we propose a new combination of Sobel and Laplacian for enhancing degraded low contrast pixels. Then the proposed method generates skeletons for text components in enhanced images to reduce computational burdens, which in turn helps in studying component structures efficiently. We propose to study the cursiveness of components based on branch information to remove false text components. The proposed method introduces the nearest neighbor criterion for grouping components in the same line, which results in clusters. Furthermore, the proposed method classifies these clusters into text and non-text cluster based on characteristics of text components. We evaluate the proposed method on a large dataset containing varieties of images. The results are compared with the existing methods to show that the proposed method is effective in terms of recall and precision.

1. Introduction

India is a multilingual country, where all the states have provision to specify their own official language, which results in many official languages and various documents in different languages. Though work on segmentation of text is improved significantly, the recognition of old scripts like Indus is still difficult because of its complexity. Indus documents consist of symbols that look like ornamental in images [1]. Generally, these symbols are carved by hand on irregular surfaces such

* Corresponding author.
E-mail addresses: kavitha_sanjay_as@yahoo.co.in (A.S. Kavitha), hudempsk@yahoo.com (P. Shivakumara), ghk.2007@yahoo.com (G.H. Kumar), lutong@nju.edu.cn (T. Lu).

Peer review under responsibility of Faculty of Computers and Information, Cairo University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.eij.2015.11.003

Figure 1 Sample Indus document images.

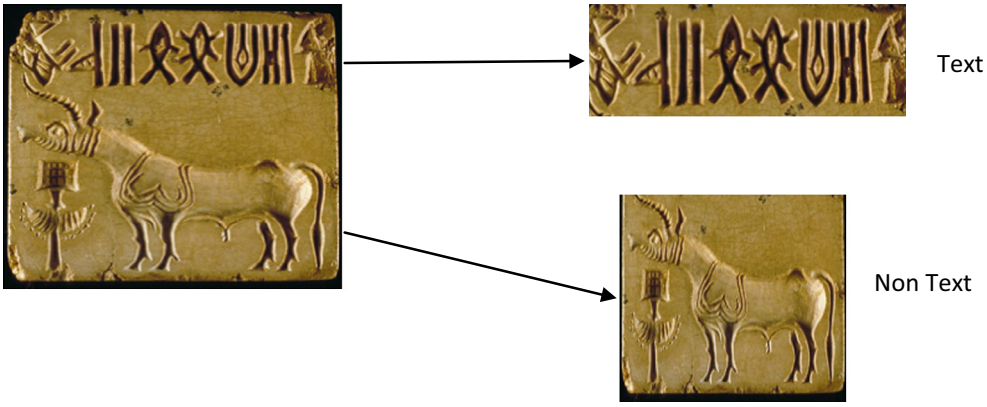


Figure 2 Illustrating text and non-text components in Indus documents.

as stones during the period of 3000 BC–1500 BC. As a result, Indus script is found in seal form that was used by people for the purpose of communication in the past. Fig. 1 shows a few images of these documents, where texts are associated with animal-like pictures in various forms such as a single horn and two horns. This complexity makes the segmentation problem more challenging and interesting. Due to the huge collection of such documents and the lack of scholars in the field of epigraphy, it is difficult to interpret all the scripts manually as it consumes a large amount of time. In order to reduce manual efforts, there is a need for the digitization of the scripts to preserve vital information for future study. Developing an automatic algorithm for converting raw script data to digital data involves four steps, namely text line segmentation, word segmentation, character segmentation, and character recognition. Text line segmentation is an important step as it facilitates other steps to achieve good recognition rates. In addition, text line segmentation is hard for the document like Indus due to the irregular structures of text components and unpredictable background variations [2,3]. Therefore, in this work, we focus on text line segmentation from Indus scripts. We can see some of the efforts toward text line segmentation [18–21] in the literature. Most of the methods are developed based on geometrical features such as aspect ratio and size for text line segmentation. Therefore, these methods may not be suitable for text line segmentation from Indus document images, where one cannot expect uniform size and structure due to complex background. Hence, we can conclude that there is an immense scope for developing a new method for segmenting text lines from Indus document images.

The paper is structured as follows. In Section 2, we give a brief survey of related work. Section 3 discusses the proposed work in detail. Finally, Section 4 discusses experimental results for the proposed method and the comparisons with the existing methods.

2. Previous work

There are several methods proposed for text extraction from scanned, handwritten, degraded and historical document images in the literature [4–10]. Most of the methods require plain and homogeneous background with a high contrast images for achieving good segmentation results. However, when we look at Indus documents as shown in Fig. 1, we cannot assume that such documents have plain backgrounds and structured text lines because these documents are handwritten with different tools on different surfaces. We consider Indus documents as a type of degraded historical document images, and text line segmentation from these documents still remains an unsolved problem. In this section, we will review the

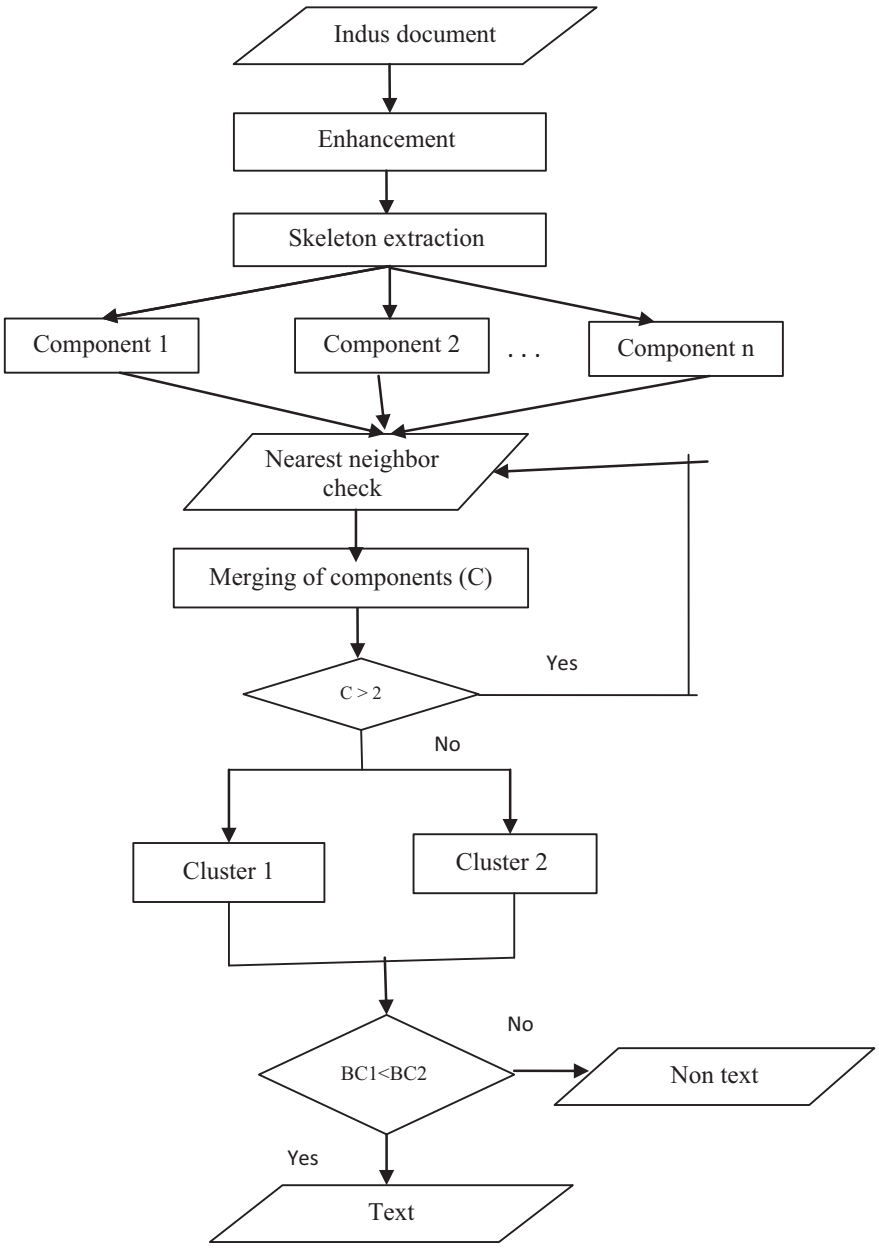


Figure 3 Flow diagram of the proposed text line segmentation: BC1 and BC2 denote the number of branches in cluster-1 and cluster-2, respectively.

literature on text line segmentation from historical handwritten degraded documents (i.e., Indus documents).

The survey on the contributions toward inscription can be found in Soumya and Kumar [11]. The authors proposed preprocessing techniques and segmentation methods of Kannada present handwritten texts based on connected component analysis. However, none of the methods is tested on unstructured layout documents. Omar and Lu [4] proposed an algorithm to extract text lines from historical document images using Steerable Directional filters. An adaptive local connectivity map is generated based on paragraph map to extract paragraphs. The objective of this paper is to find the orientation for each paragraph. Patterns are validated using projection profiles. Text lines in each paragraph are extracted by finding the central point of each connected component. However, the method can deal with the documents only containing text lines.

Recently, Gatos et al. [5] proposed a work toward segmenting historical handwritten document images into text lines and text zones. Text zones are extracted by finding vertical lines by rules. White run pixels are used for segmentation. Bounding box coordinates for each connected component are used to calculate the height of a character. However, the method requires a uniform height of characters with clear background. Kleber et al. [6] proposed a method for detecting the skewness of scanned documents, which is needed for image analysis. The method determines the skew of a document page by the Nearest Neighbor Clustering. Interested points are found using DOG to evaluate the skewness. However, the method can only

(a) Sobel Gradient Image

(b) Laplacian gradient Image

(c) Intersection of two gradient Images

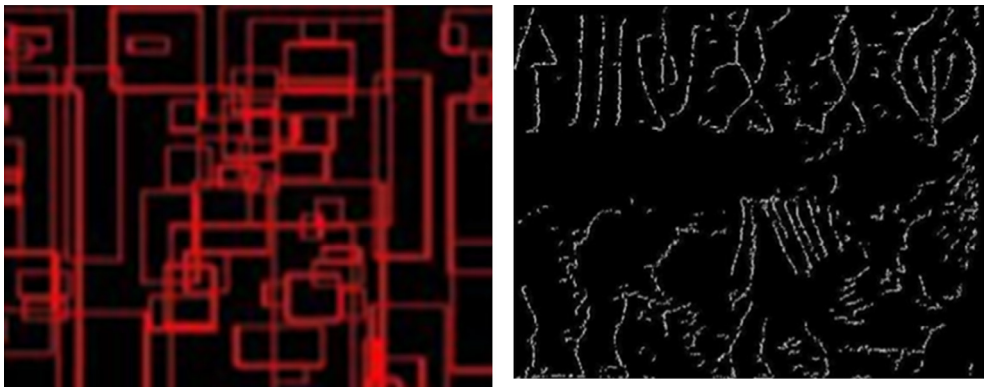
(d) Skeleton

Figure 4 Intermediate results of text enhancement method.

Figure 5 Example for pruning components.

work for the characters on plain surfaces but not for ornamental characters on irregular surfaces. Garz et al. [7] proposed binarization-free clustering to segment curved text lines in Historical documents. Text lines are detected by representing word segments using graphs, in which an edge is a link between two segments. The method works well for the texts written in ink; however, it cannot work on unconnected characters. Rabaev et al. [8] detected characters in damaged documents and then grouped them to text lines by analyzing the evolution maps of connected components. A sweep line moved from left to right is further used to check whether elements lie in the same line. However, the method can only detect lines of equal-size texts which are chosen in their dataset.

In Garzet al.'s method [9], parts of characters considered as interest points are extracted. Words are identified in high density regions and characters are separated by ascenders and descenders. However, the method concentrates for median words but not the manuscripts containing variable heights. Method is robust to background noise such as stains but not to abrupt background. Messaoud et al. [10] introduced three



(a) Overlapping bounding boxes (b) False segmentation

Figure 6 False components removal.

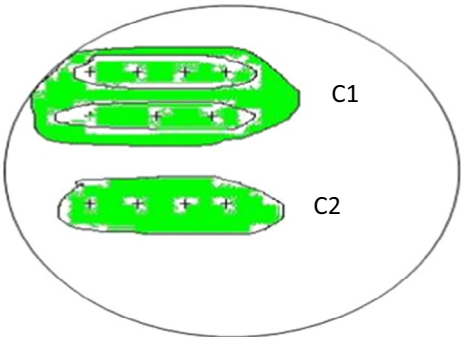


Figure 7 Nearest neighbor criterion for text and non-text clusters.

steps to achieve multilevel text line segmentation by proposing a histogram approach, grouping connected components for text line estimation, using the nearest neighbor to overcome the overlapping problem. However, the approach is prone to errors from unstructured background.

The method in [12] segments text lines based on the distances among the bounding boxes of components in an image, while the method in [13] segments a document that has text lines by drawing snakes (curves) over ridges (the central point of each line). The latter assumes a uniform height for text lines. However, it is not true for Indus documents. The advantage of the two methods is that they segment text lines irrespective of scripts and are said to be robust to non-structured layouts in Indus documents. Since these two existing methods are developed for segmenting text lines from plain background images, they may give lower accuracies for Indus documents.

In summary, it is noted from the above discussions on text lines segmentation from historical document images such as handwritten and scanned document images that none of the existing methods gives a satisfied solution for unstructured documents and the documents containing irregular structures of characters. In addition, the existing methods require a high resolution and plain background for achieving good results.

We hardly found literature on epigraphical documents like Indus documents. Murthy et al. [14] proposed a method for the segmentation of touching lines and characters in Epigraphical documents. The method in [15] discussed about the characterization of Indus scripts using the probabilities of symbols sequences and the syntactic rules generated by the analyzed sequences or the correlations among symbols. Rajesh et al. [16] proposed a probabilistic method to analyze the sequences of Indus scripts and predict missing letters.

Based on the above discussions, we can conclude that none of the existing methods can be tested on the documents like Indus for text line segmentation. Therefore, text detection from unstructured layout documents with low contrast and varied font sizes on irregular surfaces is still considered as an open challenge in the document analysis community.

3. Proposed method

It is noted from the above discussions and Fig. 1 that an image can have any contrast and a text can have any character shape. In order to enhance low contrast texts, we propose to explore the combination of Laplacian and Sobel edge images as we are inspired by the work presented in [17] for video text detection, in which Laplacian and Sobel combination has been used for increasing low contrast texts in video. In this work, we perform intersection operation for Laplacian and Sobel images of the input image to obtain an enhanced image. Laplacian operation enhances both low and high contrast pixels at near edges, and at the same time it produces noise pixels due to background variations. Sobel operation only enhances high contrast pixels without producing noise pixels. Therefore, we perform intersection operation to choose only the pixels which are significant for text line segmentation.

For the enhanced image, we further apply skeleton to reduce pixel widths of edge components. This operation preserves the structures of edge components in the enhanced image, and at the same time it saves the number of computations. We observe that generally most Indus documents contain texts along with pictures like animals. Since a picture looks like animals with different shapes, when the branches of the picture look more cursive, the picture contains a larger number of branches. With this notion, we propose a method to eliminate the components which has more cursive branches in the skeleton image. This results in pruned text image components.

The above steps help in increasing the gap between text and non-text components spatially. To segment text lines, we

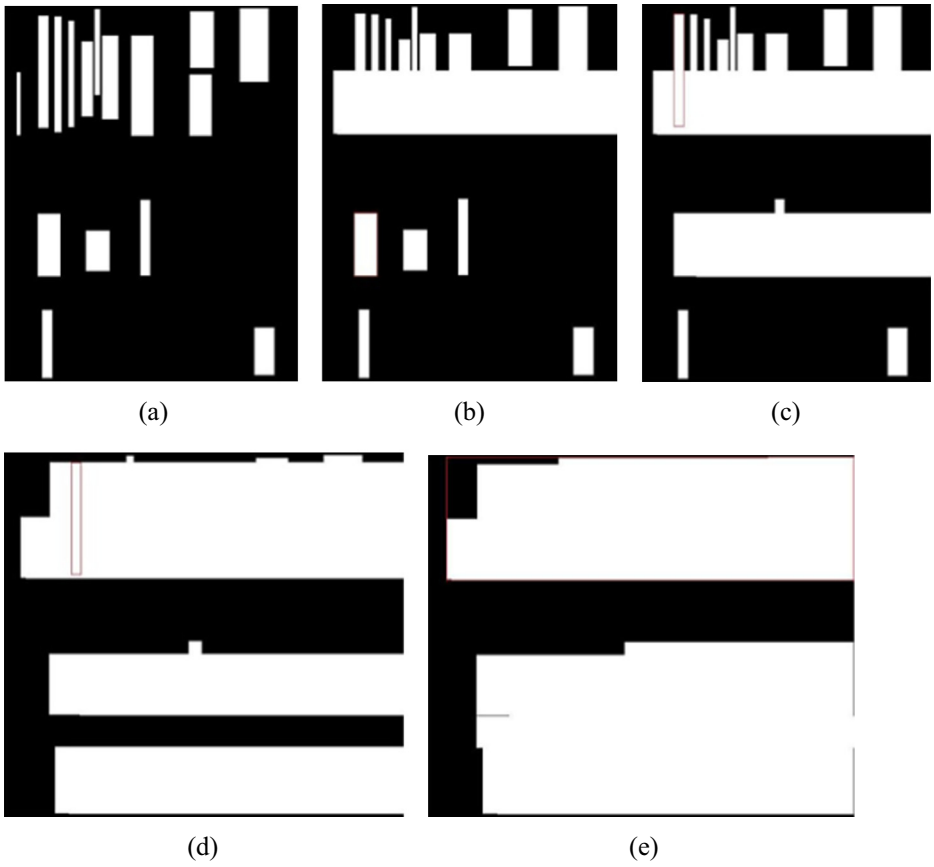


Figure 8 Illustrating grouping process.

propose clustering based on the nearest neighbor criterion to group the components that have closed proximity. The clustering process continues until the method gets two clusters for the whole pruned image. This is because it is expected text components to be formed one cluster, and non-text components to be formed another cluster. This is valid because the space between text and non-text components is larger than the space between the respective text and non-text components. The proposed method studies the number of the branches in each cluster to classify text and non-text clusters. The cluster that contains less number of branches is considered as a text cluster because text components in this cluster are not cursive as another cluster containing animal-like pictures. This results in text line segmentation. One sample example is shown in Fig. 2, where we can see text lines and non-text lines are segmented by the proposed method. The block diagram of the proposed method is shown in Fig. 3.

3.1. Text enhancement

As discussed in the above discussions, we need to enhance low contrast text components. For this purpose, we consider Laplacian and Sobel operations on the input image as these are well known gradient operations to enhance the information in the image. It is true that since Sobel operation is the first order derivative, it gives fine details for high contrast pixels. Therefore, it enhances only high contrast edges of text components but not the edges of low contrast components as in Indus document images. To overcome this problem, we propose to use Laplacian operation, which enhances both low contrast and high contrast pixels because this operation involves the second order derivative. Besides, the Laplacian operation introduces noises for complex background information. To retain enhanced edges and suppress background noises, we propose to perform intersection operation of the Sobel and Laplacian operation outputs. For example, the results of Sobel and Laplacian operations on the input images are respectively shown in Fig. 4(a) and (b), where one can notice Sobel enhances high contrast information, while Laplacian enhances both low and high contrast information along with noises. To take the advantage of both Sobel and Laplacian, we perform an intersection operation as shown in Fig. 4(c), where it is noted that only significant information is highlighted. The proposed method applies skeleton to reduce pixel width to a single pixel to save the number of computations as shown in Fig. 4 (d). Let P1(i, j) and P2(i, j) be the pixel values at position (i, j) in A and B, respectively. The intersection of these gradient images is computed as true if P1(i, j) and P2(i, j) both have positive gradient. The algorithmic steps of enhancement are represented as follows.

The masks to compute Sobel gradient Image are given by,

$$Gx = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} \quad \text{And} \quad Gy = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ +1 & +2 & +1 \end{bmatrix}$$

(a) Text cluster (b) Non-text cluster

Figure 9 Text and non-text cluster separation.

The Sobel gradient of the image for array $G[i, j]$ is obtained by,

$$Gx = ((2 * C(i+2,j+1) + C(i+2,j) + C(i+2,j+2)) - (2 * C(i,j+1) + C(i,j) + C(i,j+2))) \quad (1)$$

$$Gy = ((2 * C(i+1,j+2) + C(i,j+2) + C(i+2,j+2)) - (2 * C(i+1,j) + C(i,j) + C(i+2,j))) \quad (2)$$

The magnitude of pixel gradient is given by

$$G[i,j] = \left|\sqrt{Gx^2 + Gy^2}\right| \quad (3)$$

Laplacian Gradient for array $C[i, j]$ is computed as

$$Kx = ((C(i+2,j+1) + C(i,j+1) - 2 * C(i,j))) \quad (4)$$

$$Ky = ((C(i+1,j+2) + C(i+1,j) - 2 * C(i,j))) \quad (5)$$

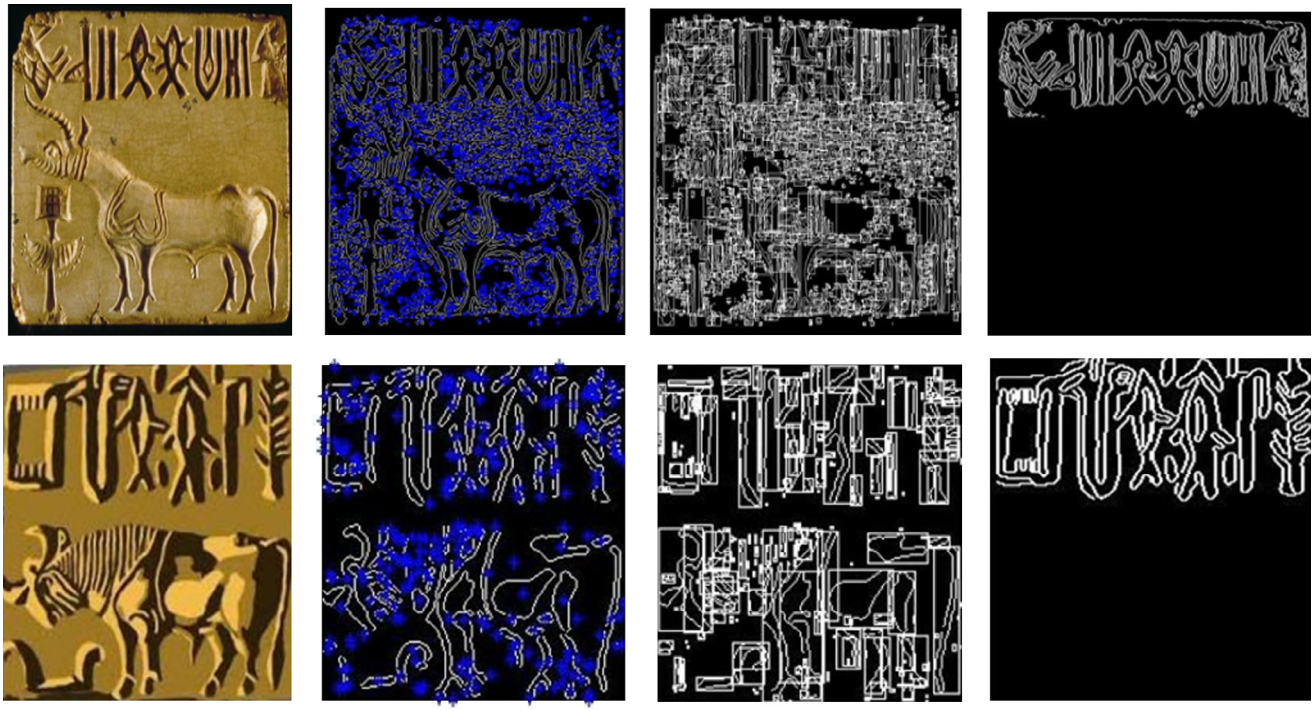
$$K[I,J] = Kx + Ky \quad (6)$$

The intersection of gradient images is obtained by

$$G[I,J] \cap K[I,J] \quad (7)$$

3.2. Pruning text components

Due to complex document images, the above enhancement step may enhance non-text components as shown in Fig. 4 (d). It is also true that a skeleton algorithm sometimes creates disconnections between the components. To avoid disconnections, we perform smoothing using morphological operation. This results in a smoothed image, where we can see connected components without disconnections. Then the proposed method fixes the bounding box for each component in the smoothed image. If the bounding box of components overlaps with the bounding boxes of other components, it will be



(a) Input (b) Bukhari et al.[13] (c) Diemet al.[12] (d) Proposed method

Figure 10 Text line segmentation by the existing and the proposed methods.

Table 1 Matching matrix of the proposed method on the classification of text lines.

Proposed method	Text (%)	Non-text (%)
Text	**91**	9
Non-text	13	**87**

Those two values are classification rates for text (91%) and non-text (87%).

Table 2 Performance of the proposed and existing methods on text line segmentation.

Methods	G_t	T_p	F_n	F_p	Recall	Precision
Diem et al. [12]	515	20	495	500	0.038	0.03
Bukhari et al. [13]	515	65	450	500	0.126	0.11
Proposed	515	461	54	13	**0.89**	**0.97**

merged to make a single component. This connects sub-components, which have more than two pixel gaps. When we observe from the skeletons of an Indus document image, we are noted that the components that represent an animal-like picture have more cursive branches compared to the components that represent text as discussed. The proposed method then counts the number of branches for the results obtained by overlapping bounding boxes. If the number of a branch is larger than a certain threshold value, we discard it as a non-text component as shown in Fig. 5, where the marked components will be removed from the image. Fig. 6(a) shows the result after merging the nearest sub-components into a component using overlapping bounding boxes. Fig. 6(b) shows the result after discarding the components that have a large number of branches.

3.3. Text line segmentation

It is noted from Fig. 6(b) that we can still see spaces between text and non-text regions due to the elimination of unwanted components. Next, we need to group text and non-text components separately to extract text lines from the image. Since the space between two regions generally looks larger than the space between two components, we propose the nearest neighbor clustering method for grouping the nearest components. For each component in the image, the method finds the nearest neighbor component using Euclidean distance as defined in Eq. (8). The component that gives the minimum distance is considered for grouping. This process continues until the method gets two clusters for the whole image. Since it is a two class problem (text and non-text) and it is known that the average space between two regions is larger than that between two components, the grouping process continues until it gets two clusters as shown in Fig. 7, where we can find the grouping results in two regions. If the distance between two regions does not satisfy a certain threshold, there are chances of getting more than two clusters. This is a rare case for Indus script because according to observation generally each image contains only one text lines with animal picture background. It can be illustrated mathematically as follows. Let $C = \{C1, C2, C3, Cn\}$ be the finite set of components. Let C_i be candidate components considered for merging. The distance between a candidate component and another component is calculated by finding the boundary values of the components. C_j is the set of all the other components excluding C_i. The distance between the two components is considered for merging. Let (X_1, Y_1) and (X_2, Y_2) be the extreme coordinates of two components facing each other, and the distance between the two components is obtained by

$$\text{Euclid\_dist}_{X1,X2} = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2} \tag{8}$$

The proximity of closeness defined by the minimum distance criteria is calculated by

$$Cn = \text{Min}\left\{d(a,b) : a \epsilon C_i, b \epsilon C_{j=1,n} - C_i\right\} \tag{9}$$

We merge two nearest components by

$$Cnew = Rci \cup Rcj \cup Cn \tag{10}$$

The process of grouping is illustrated in Fig. 8, where (a)–(e) show the step by merging the nearest neighbor components and the final results. Since the grouping process involves unsupervised nearest neighbor clustering criterion, it results in clusters without identifying them. Therefore, we propose to extract features which represent text components, namely, the number of branches in text components in the cluster for classifying the cluster as a text one. The cluster that gives a less number of branches is considered as a text cluster because text components usually have fewer branches compared to non-texts like an animal picture.

Let NB_{c1} and NB_{c2} be the numbers of the branches in cluster1 and cluster2, respectively. If $NB_{c1} < NB_{c2}$ then the contours of clusters c1 and c2 are located as text and non-text, respectively. If $NB_{c2} < NB_{c1}$ then the contours of clusters c2 and c1 are located as text and non-text, respectively. The results can be seen in Fig. 9, where we can see text and non-text regions separately.

4. Experimental results

Since text line segmentation from Indus documents is a new problem as introduced, we create our own dataset consisting of 500 images from archeology survey of India, Mysore and magazines. This dataset includes varieties of text lines of different handwritings with different tools on different surfaces. As a result, this dataset is said to be complex compared to scanned document images. To measure the performance of the proposed method, we use classification rate with confusion matrix. To generate confusion matrix, we count the cluster that represents text and the cluster that represents non-text for calculating classification rate. The matching matrix of the proposed method is reported in Table 1, where it is noted that the proposed method gives a good classification rate for text and non-text classifications.

For evaluating text line segmentation, we use well-known measures such as recall and precision as in Eqs. (11) and (12). The definitions for recall and precision are as follows. Let G_t be the total number of the text lines counted in 500 images, T_p be the number of the text lines segmented from the proposed method, F_n be the number of the lines that are

not segmented, and F_p be the number of non-texts that are classified as texts.

$$\text{Recall} = \frac{Tp}{(Tp + Fn)} \tag{11}$$

And

$$\text{Precision} = \frac{Tp}{(Tp + Fp)} \tag{12}$$

To show the effectiveness of the proposed method, we implement two recent methods on text line segmentation for comparative studies. The method in [12] segments text lines based on the distance between the bounding boxes of the components in an image. The method in [13] segments the documents having text lines by drawing snakes (curves) over ridges (central points of each line). The latter method assumes a uniform height for all text lines. However, this is not true for Indus documents. The reason to choose these two methods is that they segment text lines irrespective of scripts and datasets and are said to be robust to non-structure layouts as in Indus documents. Since these two existing methods are developed for segmenting text lines from plain background images, the existing methods report poor accuracies for our Indus documents. The qualitative results of the proposed and the existing methods are shown in Fig. 10, where one can notice that the proposed method is able to segment text lines correctly for both the images, while both the existing methods fail to segment the first image in Fig. 10 due to the inherent limitations of the existing methods such as the requirements of both high resolution and plain backgrounds. For the second image shown in Fig. 10, the existing methods segment text and non-text lines correctly since the image contains enough spaces between text and non-text lines. Since the existing methods aim at segmenting text lines, they focus on the segmentation of text and non-text lines without separating text and non-text lines as shown in Fig. 10 on the second image. The quantitative results of the proposed and existing methods are reported in Table 2, where we can see that the proposed method is the best at recall and precision compared to the existing methods.

5. Conclusion and future work

We have proposed a new method for segmenting text lines from degraded historical document images like Indus. The proposed method introduces a new combination of Laplacian and Sobel operations for enhancing low contrast pixels in the images. The characteristics of the components in the image are studied to eliminate unwanted components, which results in text components pruning in the image. We have proposed a grouping process, which involves the nearest neighbor criterion for merging text components. The iterative clustering process is then proposed to separate text and non-text regions. Our future plan would be extending the same method for other Indian scripts to show its ability and generic properties. We also focus on character segmentation from segmented text lines and character recognition.

References

[1] Mahadevan Iravatham. Dravidian proof of the Indus script via the rig Veda: a case study. Bull IRC 2014;4(1).

[2] Kavitha AS, Shiva kumara P, Kumar GH. Skewness and nearest neighbour based approach for historical document classification. Proc CSNT 2013:602–6.

[3] Kavitha AS, Shiva Kumara P, Kumar GH. An integrated method for classification of Indus and English document images. Proc ICERECT 2012:343–55.

[4] Omar A, Lu CC. Text line extraction for historical document image using steerable directional filters. Proc ICALIP 2014:312–7.

[5] Gatos B, Louloudis G, Stamatopoulos N. Segmentation of historical handwritten documents into text zones and text lines. Proc ICFHR 2014:464–9.

[6] Kleber F, Diem M, Sablatnig R. Robust skew estimation of handwritten and printed documents based on gray value images. Proc ICPR 2014:3020–5.

[7] Garz A, Fischer A, Bunke H, Ingold R. A binarization-free clustering approach to segment curved text lines in historical manuscripts. Proc ICDAR 2013:1290–4.

[8] Rabaev I, Biller O, El-Sana J, Kedem K, Dinstein I. Text line detection in corrupted and damaged historical manuscripts. Proc ICDAR 2013:812–6.

[9] Garz A, Fischer A, Sablatnig R, Bunke H. Binarization-free text line segmentation for historical documents based on interest point clustering. Proc DAS 2012:95–9.

[10] Messaoud IB, Amiri H, Abed HE, Margner V. A multilevel text line segmentation framework for handwritten historical documents. Proc ICFHR 2012:515–20.

[11] Soumya A, Kumar GH. Preprocessing of camera captured inscriptions and segmentation of handwritten Kannada text. IJARCCE 2014;3(5):6794–803.

[12] Diem M, Kleber F, Sablatnig R. Text line detection for heterogeneous documents. Proc ICDAR 2013:743–7.

[13] Bukhari SS, Shafait F, Breuel TM. Script-independent handwritten textlines segmentation using active contours. Proc ICDAR 2009:446–50.

[14] Murthy KS, Kumar GH, Shivakumara P, Ranganath PR. Nearest neighbour clustering approach for line and character segmentation in epigraphical scripts. Proc ICCS 2004.

[15] Rajesh P, Rao PN, Yadav N, Vahia MN, Hrishikesh Joglekar, Adhikari R, Mahadevan I. Entropy, the Indus script, and language. Comput Linguist 2010:795–805.

[16] Rajesh P, Rao PN. Probabilistic analysis of an ancient undeciphered script. Proc Comput Soc 2010:76–80.

[17] Shivakumara P, Sreedhar RP, Phan TQ, Shijian L, Tan CL. Multi-oriented video scene text detection through bayesian classification and boundary growing. IEEE Trans CSVT 2012:1227–35.

[18] Zhu A, Wang G, Dong Y. Robust text segmentation in low quality images via adaptive stroke width estimation and stroke based superpixel grouping. Lect. Notes Comput. Sci. 2015:119–33.

[19] Pintus R, Yang Y, Rushmeier H. Automatic text height extraction for the analysis of text lines in old handwritten manuscripts. ACM J Comput Cult Herit 2013:25.

[20] Gaurav SM, Nandish C. A survey and analysis of segmentation, feature extraction and classification in OCR system. IJAR 2015;5(1):24–6.

[21] Thakur P, Azam A. Edge detection through integrated morphological gradient and fuzzy logic approach. IJSETR 2015;4(5):1613–6.