Text Extraction from Degraded Historical Document Images

B. Dhivyabharathi
M.E.Communication Systems
Department of ECE
Saranathan College of Engineering, Trichy

Dr. M. Santhi M.E., Ph.D.,
Professor and Head of the Department
Department of ECE
Saranathan College of Engineering, Trichy

***Abstract*— Text extraction from degraded historical Indus script images is challenging due to complex background. In this paper, we present a new method for extracting text from the Indus documents. Canny edge detection operator is used to enhance the degraded low contrast pixels. Then, Skeleton of the enhanced image is generated to study the component structures. Further, Clustering method is used to separate the text components from the background. Nearest neighbor criterion is used for this purpose.**

Keywords—Canny edge detection operator, Skeleton algorithm, Clustering, nearest neighbor criterion.

I. I. INTRODUCTION

India is a multilingual country, where each state has unique official language and their historical documents written in different languages. Text extraction from these historical documents is still difficult. Because these documents are written in complex background and also the characters present in these documents are of different sizes, styles, orientations and alignments. Generally Indus documents consist of both text and non-text symbols. Such symbols are carved on stones and in irregular surfaces. It is difficult to recognize the text characters in an irregular background because there is a chance for degradation of these historical documents. Since the low contrast texts and the complex background makes the extraction process more challenging.

In the field of epigraphy, there are millions of historical documents available. Manual interpretation of such huge amount of documents consumes large amount of time and also it is impossible to store all such documents over a long period of time. In order to overcome these difficulties, digitization of these scripts is necessary.

In digitization, the raw script data is converted into digital data which involves four steps and are as follows: 1.Text line segmentation 2.Word segmentation 3.Character segmentation 4.Character recognition. Among these four steps, Text line segmentation plays an important role in achieving good recognition rates even though it is hard for Indus documents due to unstructured characters and unpredictable background variations. In this paper, we concentrate on text line extraction from historical document images. Most of the existing methods are suitable only for extracting text characters from uniform background. These methods depend on geometrical features like size and aspects ratio.

This paper is structured as follows. In section II, we provide information about related work. Section III, gives the proposed work in detail. Lastly, in section IV, we discussed experimental result for proposed method.

II.RELATED WORK

Text extraction from scanned, handwritten, degraded and historical document images can be done by several methods. In order to achieve best recognition results, most of the methods require plain and homogeneous background with high contrast images. In this section, we will review the literature on text extraction from degraded historical documents.

Steerable directional filter is proposed to extract the text from document [1]. By means of the paragraph map, an adaptive local connectivity map is created to extract the paragraphs. Then the orientation of each paragraph is calculated. By using projection profile analysis, the patterns are validated. Finally the text lines present in each paragraph are extracted by calculating the central point of each connected component.

The text lines and text zones are extracted from historical handwritten document images [2]. Text zones are identified by using vertical lines. Then the height of a character is calculated by finding the bounding box coordinates for each connected component.

The skewness of the scanned historical documents is identified [3]. Nearest neighbor clustering is used to determine the skew of the document. With the help of DOG, interested points are calculated for checking the skewness of the document page. However, this method is suitable only for the documents having uniform plain background.

A binarization free clustering technique is proposed in[4] to extract the carved text lines from historical documents. First the word segments are indicated using graphs, in which an edge is a link between two segments. Then the text lines are formed from the word segments. However, this method is not works well for unconnected characters.

Characters present in the damaged documents are identified in[5]. From the analysis of evolution maps of connected components, text lines are formed by grouping the identified characters. The movement of sweep line is used to check whether the elements lie in the same line.

A bottom up approach is used in [6] that fuse words to reduce their fusing distance. And also the text lines are represented by oriented rectangles for further layout analysis.

The central line of parts of text lines is computed using ridges over the smoothened image [7]. Then the state of the art active contours is applied over ridges. Finally we obtain the text line extraction results. The methods used in [6] and [7] are more advantageous that the extraction of text lines are irrespective of scripts. This method is reliable only for uniform background images and not for Indus document.

In summary, the existing methods provide better results

only for structured components with high resolution and plain background. So the text extraction from unstructured layout documents having low contrast and various font size characters on irregular surfaces is still becoming a problem in analysis of documents.

III.PROPOSED SYSTEM

In Fig.1. The sample Indus document image comprises of both text and non-text components. Due to the degradation of these documents the images become low contrast and the text components present in these documents can have different character shapes. To enhance the low contrast text, edge operators like Sobel, Laplacian and Canny operators are used. Laplacian operator enhances both low contrast and high contrast pixels and also it creates noise pixels due to background variations. Sobel operator enhances only high contrast pixels and it does not produce noise pixels.

After enhancement of the input document image, Skeletonization is performed to minimize the pixel widths of edge component. By doing Skeletonization, the structures of the edge components are preserved. Usually, most Indus documents have text along with animal like pictures. In order to extract the text and non-text components separately, clustering method based on the nearest neighbor criterion is used. Such separated text and non-text components are grouped as text cluster and non-text cluster respectively.

Text enhancement

As discussed in the above discussions, the enhancement of low contrast text components is required. In order to enhance the low contrast text components, edge operators like Sobel, Laplacian and Canny are used. Sobel is the first order derivative operator; it produces fine details for high contrast pixels. Therefore, the high contrast edges of the text components are enhanced by using this operator.

Fig.1. Sample Indus document Images

Text

Non-Text

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

The Sobel Gradient of the Images for array G [i, j] is given 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 pixels gradient is given by,

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

Indus document images consist of low contrast text components. To enhance both high contrast and low contrast pixels Laplacian operator is used. Laplacian is the second order derivative operator; it produces noise for complex background.

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

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

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

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

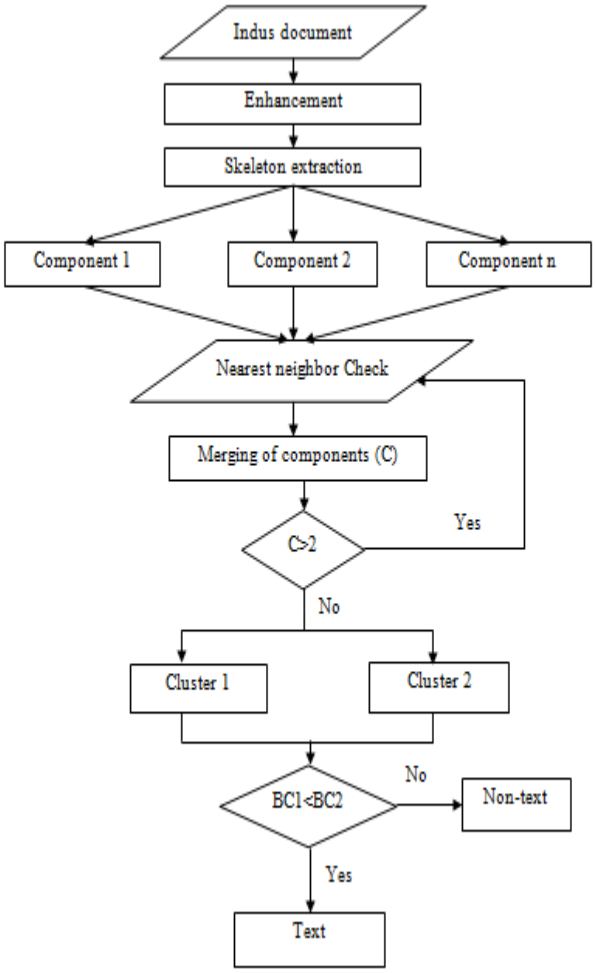

Fig.3.Flow diagram of the proposed system

Canny operator is widely known as the optimal detector. To obtain an enhanced image, canny operator performs multiple steps. Gaussian filter is used to discard noise from an image and it produces smoothened image. Then, the intensity gradient and direction of the smoothened image is calculated. Non-maximum suppression is applied to remove the unwanted pixels. Hysteresis thresholding is performed along edges using two thresholds, minimum and maximum. If a pixel gradient is higher than the maximum threshold, then the pixel will be marked as edge. Otherwise the pixel will be discarded. And if the pixel gradient is between the two thresholds, then only the pixel that is connected above the maximum threshold is noted as edge.

Pruning text components

Due to the complex document images, non-text components may be enhanced during enhancement step. Skeletonization process also creates disconnections between these components. To avoid disconnections, smoothing is performed using morphological operation. From the smoothed image, we can observe the connected components without disconnections. Then the proposed method places the bounding box for each component present in the smoothed image. If the bounding box of components overlaps with the bounding boxes of other components, it will be merged and considered as the single component. From the skeleton of the image, we can observe that the non-text components i.e. animal like pictures have more cursive branches compared to text components. If the number of branch is larger than a certain threshold, we can remove that as non-text components.

Text line extraction

The grouping of text and non-text components is required to extract the text lines from the image. Nearest neighbor clustering method is used to group the nearest components.
For each component in the image, nearest neighbor component is determined using Euclidean distance measurement. If the component gives minimum distance, it will be considered for grouping. This process repeats until we get two clusters for the whole image. Since it is a two class problem (text and non-text) and the average space between two regions is larger than the two components. If the distance between two regions does not satisfy a certain threshold, there may be a chance of getting more than two clusters. This is a rare case for Indus documents because it contains only text lines with animal like picture. It can be illustrated mathematically as follows.

Let C= {C1, C2, C3, Cn} be the finite set of components. Let Ci be the candidate components considered for merging. The distance between a candidate component and another component is determined by calculating the boundary values of the components. Cj is the set of all the other components excluding Ci. The distance between the two components is used for merging. Let (X1, Y1) and (X2, Y2) be the extreme coordinates of two components facing each other, and the distance between the two components is calculated by,

$$\text{Euclidean distance}_{X1,\,X2} = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2} \quad (7)$$

The proximity of closeness defined by the minimum distance criteria is obtained by,

$$Cn = \text{Min}\ \{d(a, b): a \in Ci, b \in Cj_{=1,\,n} - Ci\} \quad (8)$$

We merge two nearest components by,

$$C_{new} = R_{C_i} U R_{C_j} U C_n \quad (9)$$

Text components have fewer branches compared to non-text components. Let NB_{C1} be the number of branches in cluster 1 and NB_{C2} be the number of branches in cluster 2.If $NB_{C1} < NB_{C2}$ then C1 can be considered as the text cluster and C2 can be considered as the non-text cluster. If $NB_{C2} <$ NB then C2 can be treated as text cluster and C1 can be treated as non-text cluster.

IV. RESULTS AND DISCUSSION

The experiments are performed on windows 7 processor with an Intel(R) Core™ i3-2328M Machine with 2.20 GHz CPU and 2GB RAM. All the programs are written and

compiled on MATLAB version 7.14.0.739 (R2012a).
The sample historical document image is shown in figure 4.

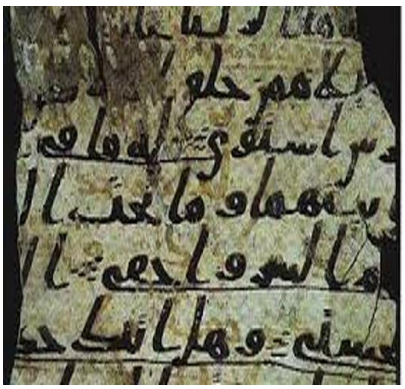

Fig 4. Input document image

The gray scale image of input document image is shown in figure 5.

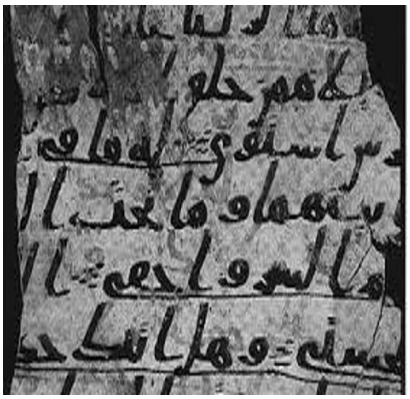

Fig 5. Grayscale image of input document

The sobel edge detected image is shown in figure 6.

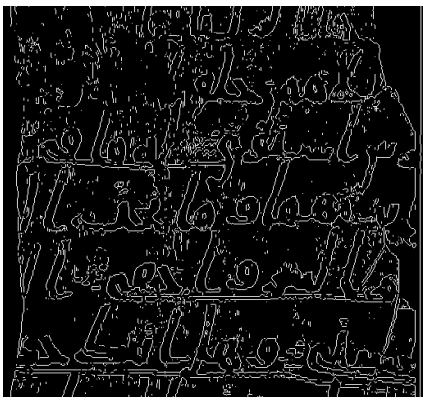

Fig 6. Sobel gradient image

The Laplacian edge detected image is shown in figure 7.

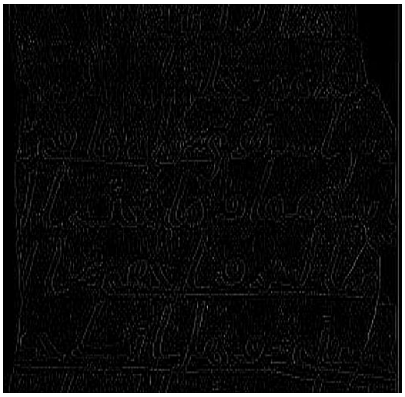

Fig 7. Laplacian image

The Canny edge detected image is shown in figure 8.

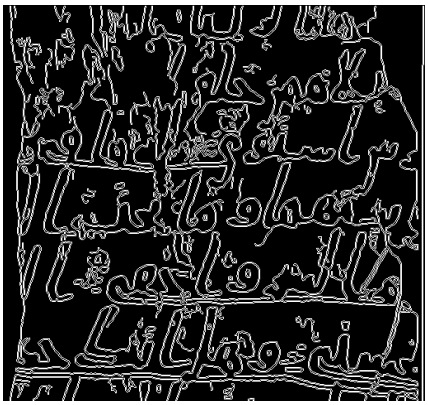

Fig 8. Canny edge detected image

The skeleton of the enhanced image is shown in figure 9.

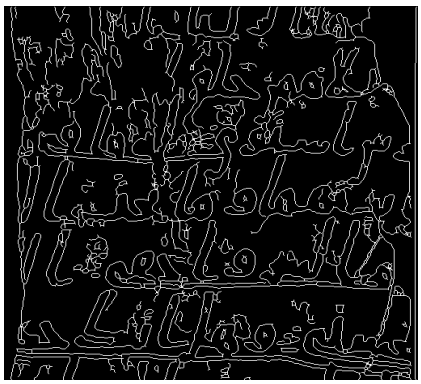

Fig 9. Skeletonized image

The bounding box of each connected component is shown in figure 10.

Fig 10. Overlapping bounding boxes

The extracted text components are shown in figure 11.

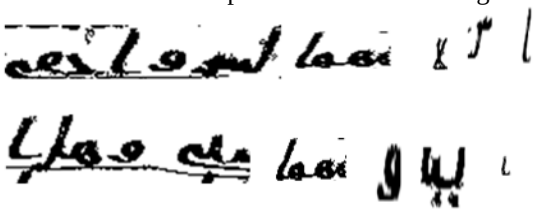

Fig 11.Extracted text components

V.CONCLUSION

We have proposed a new method for extracting text lines from degraded historical document images. The canny operator is used to enhance the low contrast pixels present in the image. Then, the skeleton of the enhanced image is obtained to preserve the structures of the edge components. The characters are identified and grouped using clustering method. The iterative clustering process is used to extract the text components effectively.

REFERENCES

[1] Omar A.Lu CC. "Text line extraction for historical Document image Using steerable directional filters". Proc ICALIP 2014:pp 312-317.
[2] Gatos B. LouloodisG, Stamotopoulos N. "Segmentation of historical Hand written documents into text zones and text lines". proc. ICFHR 2014:pp.464-469.
[3] Kleber F, Diem M, Sablatnig R. "Robust skew estimation of Handwritten and printed documents based on gray value images". Proc. ICPR 2014:pp. 3020-3025.
[4] Garz A,Fischer A,Bunke H,Ingold R.a Binarization-free clustering approach to segment curved text lines in historical manuscripts.Proc ICDAR 2013:pp.1290-1294.
[5] Rabaev I,Biller O,El-Sana J,Kedem k,Dinstein I.Text line detection in corrupted and damaged historical manuscripts.Proc ICDAR 2013:pp. 812-816.
[6] Diem M,Kleber F,Sablatnig R.text line detection for heterogeneous documents.Proc ICDAR 2013:pp.743-747.
[7] Bukhari SS,Shafait F,Breuel TM,Script-independent handwritten text lines segmentation using active contours.Proc ICDAR 2009:pp.446-450.
[8] Murthy KS,Kumar GH,Shivakumara P,Ranganatha PR.Nearest Neighbor clustering approach for line and character segmentation in Epigraphical scripts.Proc ICCS 2004.
[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:pp.95-99.
[10] Soumya A,Kumar GH.preprocessing of camera captured inscriptions and segmentation of handwritten Kannada text.IJARCCE 2014:1(5):pp.6794-6803.
[11] Gaurav SM,Nandish C.A survey and analysis of segmentation,feature extraction and classification in OCR system.IJAR 2015:5(1):pp. 24-26.
[12] Thakur P, Azam A. Edge detection through integrated morphological gradient and fuzzy logic approach. IJSETR 2015:4(5):pp.1613-1616.. Mach. Learn., vol. 25, no. 6, Jun 2003, pp. 713–724.