RESEARCH ARTICLE OPEN ACCESS

Text line Segmentation of Curved Document Images

Anusree.M[*], Dhanya.M.Dhanalakshmy[**]
[*](Department of Computer Science, Amrita Vishwa Vidhyapeetham, Coimbatore -641 112)
[**] (Department of Computer Science, Amrita Vishwa Vidhyapeetham, Coimbatore -641 112)

ABSTRACT
Document image analysis has been widely used in historical and heritage studies, education and digital library. Document image analytical techniques are mainly used for improving the human readability and the OCR quality of the document. During the digitization, camera captured images contain warped document due perspective and geometric distortions. The main difficulty is text line detection in the document. Many algorithms had been proposed to address the problem of printed document text line detection, but they failed to extract text lines in curved document. This paper describes a segmentation technique that detects the curled text line in camera captured document images.

Keywords **–** Curved document images, Gradient vector flow, Gray scale image, Optical character recognition, Rectification.

I. Introduction

Digitization of document is the most commonly used method for increasing the quality and compatibility of the document. Now a days, compared to scanned document, camera captured document has been widely used for digitization. Most of the camera captured documents such as degraded books and newspapers suffer curling of text lines due to perspective distortion. There is a huge amount of document images in libraries and in various national archives that have not been exploited electronically. If these documents need to be electronically available, the first and major step will be document segmentation into text lines. Before text line segmentation, preprocessing steps are carried out. Thresholding is used for the binarization that converts a gray-scale (8 bit per pixel) document image into a binary document image (1 bit per pixel). It can be divided into global thresholding and local thresholding. For flattening or straightening the document, first stage is to segment the text line from the document images. This is the most difficult step in rectification method. There are many methods in text line extraction in the typical text line extraction stage. But in the case of the curled document image, it is difficult to use usual algorithms such as projection profile, x cut, run length smearing etc. Here we are proposing method text line segmentation for curved document images.

II. Literature survey

There are many methods to extract text lines from single-oriented documents. Shijian Lu et al. [1] proposed a method for implementing the text line segmentation of curved document. The rectification of curled text lines has been carried out by using partition based approaches. Partition based approaches include baseline detection, x line, and vertical stroke identification. Here x line and baseline have been estimated using least square fitting of the character tip point. Documents are partitioned into multiple quadrilateral patches. Finally target rectangle has been constructed. It is done using aspect ratio and number of the characters enclosed within the partitioned block. This method work on Latin-based document and the authors classified the document in to six categories based on shape such as character span, character ascender and descender. Compared with the existing algorithms, this method doesn't need any special hardware or camera calibration. But the main disadvantage of this system is that it will not work on low resolution images. This method is not suitable for handwritten and historical documents.

Syed Saqib Bukhari et al. [2] presented a new algorithm for text line segmentation of curled document images. Here curled text-line segmentation is performed by adapting active contour (snake) which has been used for de-wrapping and improving the OCR quality as well as the readability. Snakes are used for estimating the local pair of x line and base line. This techniques uses x line and base line estimation using connected component analysis and this is applied to the monocular single camera captured document images. The overlapping x line and baseline are considered as segmented text lines. Here coupled snakes model has been used instead of snake model. The snake moves towards targeted object under the influence of internal energy and external energy. The internal and external energy has been calculated from image content and snake point

respectively. After finding the external energy, the evolving and deformation of the snakes have to be carried out using gradient vector flow. The main advantage of this method is that it is less sensitive to high degree of curl and skew, while the challenge is that it produces many number of over-and under-segmentation errors.

Partha Pratim Roy et al. [3] proposed text line extraction using foreground and background information in the document image. For the extraction of curved document images, water reservoir technique has been used. Using positional information and the size, each individual component is grouped together as a cluster. In the second stage, the potential region has been found and individual lines are estimated using candidate region. The main advantage of this method is to determine the document containing multi-oriented and curled lines. The steps include initial character clustering, grouping cluster, candidate point extraction followed by extension of cluster group. The efficiency of the algorithm depends upon the binarization result. So it does not work with the colored document. But it is independent of font style, rotation, scale etc..

Vaibhav Gavali et al. [4] describe the algorithm that is independent of the text and the color of the document. The paper describes a technique which extracts text lines from a curved gray scale images. The gray scale image has been taken as input; binarized and discrete haar wavelet is applied followed by thresholding to remove some non-text region in the document. The morphological operation has been used for each band to connect disconnected characters. At last, 3 dilated bands are combined using logical AND operator to get segmented text lines. Here 2D haar wavelets are used for edge detection which decomposes the input document in to LL, LH, HL, and HH. The detailed information of the text lines are presents in the LH, HL, HH and this output is threshold using dynamic threshold.

G. Louloudis et al. [5] used a block-Based Hough Transform for text line extraction. This method is used for unconstrained handwritten documents. In the first step, preprocessing for enhancement of image, connected component extraction and average character height estimation were carried out. In binarization, adaptive thresholding method has been used. From the result all connected component has been identified and area of the bounding box is calculated. The proposed approach consists of partitioning the connected component space into three subsets, and the splitting of connected components into equally spaced blocks each of them voting in the Hough domain. This proposed method is applied to handwritten document which are in the Latin format. The output of the document is post-processed to reduce more than one line which corresponds to same Hough domain.

III. Proposed method

The architecture of the system is shown in the figure 1. In the first stage, input color image is converted to gray scale and preprocessed. The next stage determines edges. At last, curved text lines are segmented using morphological operation.

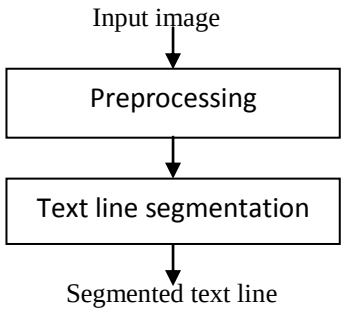

Fig.1.Proposed system architecture

3.1. Pre-processing

Pre-processing is the first step for document image analysis. Here pre-processing include binarization and noise reduction.

3.1.1. Binarization

In an OCR, one of the main processing stages is binarization of document images. Binarization of a text image should give the foreground text in black and noisy background in white. There are different thresholding methods already exist; they don't give exact results for all types of document images. Few algorithms might work better for one type of documents while they might give poor results for other types where there is a complex background, extremely low intensity variation. Depending upon the document the output of threshold will vary, hence in this project different methods were tried for getting better result.

3.1.1.1. Otsu's method

Otsu's method is one of the famous global thresholding methods. This method tries to find the threshold *T* which separates the gray-level histogram in an optimal way into two segments. Here the calculation of inter-classes or intra-classes variances is based on the normalized histogram of the image H= [h0...h255] where Σ hi=1. In Otsu algorithm, result has noise in the form of background being detected as foreground.

3.1.1.2. Niblack's algorithm

Niblack's algorithm calculates a pixel-wise threshold by sliding a rectangular window over the gray level image. The computation of threshold is based on the local mean *m* and the standard deviation *s* of all the pixels in the window and is given by the equation 1 below

$$T = m + k * s \qquad (1)$$

Here *m* is the average value of the pixels *pi*, and *k* is fixed to -0.2. Advantage of Niblack is that it always identifies the text regions correctly as foreground but on the other hand tends to produce a large amount of binarization noise in non-text regions also.

3.1.1.3. Savola algorithm

Savola algorithm claims to improve Niblack's method by computing the threshold using the dynamic range of image gray-value standard deviation.The binarization is given by equation (2)

$$T = m(1 - k)(1 - \frac{\sigma}{R}) \qquad (2)$$

Where *k* is se to 0.5 and *R* varies from 0 to 255. This method outperforms Niblack's algorithm in images where the text pixels have near 0 gray-values and the background pixels have near 255 gray-values. However, in images the gray values of text and non-text pixels are close to each other results degrade significantly.

Thresholding is used to segment an image by setting all pixels whose intensity values are above a threshold to a foreground value and all the remaining pixels to a background value. The traditional thresholding operator uses a global threshold for entire images; this method changes the threshold dynamically over the image. it can accommodate changing lighting conditions in the image, such as those occurring as a result of a strong illumination gradient or shadows.

3.1.1.4. Adaptive thresholding

Adaptive thresholding takes a grayscale or color image as input and it produces a binary image representing the text lines in the image. The pixel value is above threshold, it is set to the foreground value; otherwise it assumes the background value. Local adaptive thresholding will select an individual threshold for each pixel based on the range of intensity values in its local neighborhood pixels. This allows for binarization of a document whose global intensity histogram doesn't contain distinctive peaks. The algorithm involves following steps:

1) Convolve the image with a suitable statistical operator.
2) Subtract the original from the convolved image.
3) Threshold the difference image with *C*.
4) Invert the threshold image.

Out of the four methods described above, adaptive binarization algorithm was found to give better results for our dataset.

3.1.2 Noise Removal

Morphological image processing is a set of operations related to the shape or morphology of features in an image. Here median filter is mainly used for marginal noise reduction. The median filter is mainly used for removal of salt and pepper noise. It replaces value of the pixel by the median of the intensity values in the neighborhood of that pixel.

3.2. Text line detection

Optical character recognition (OCR) or text recognition is an important area in commercial software development and academic research. The application areas of digital document processing are office and library automation, publishing houses, communication technology, banking sector, historical and many other domain areas. A large number of techniques have been proposed to address the problem of document text extraction. Text line detection is a critical stage towards unconstrained handwritten document recognition. It refers to the segmentation of a document page image into distinct entities, the text lines. The challenges in problems that appear in this stage are the difference in the skew angle between lines on the page, overlapping words and adjacent lines touching.

3.2.1. Canny Edge Detection

To detect the edges in the document canny edge detector has been used .The main advantages this method are it has low error rate, edge point is localized and accurate result is obtained on the single edge point pixel. The canny detector algorithm involves four steps:-

1. Smoothing the input image using guasssain filter.
2. Find out gradient magnitude and angle
3. Nonmaxima suppreeion to the previous result.
4. Apply double thresholding to identify the edges.

The image smoothening is carried out by using circular 2D Gaussian function which is denoted by equation 3:-

$$G(x.y) = e\frac{(x^2 + y^2)}{2\sigma^2} \qquad (3)$$

Then smoothening is carried out by convolving input image with the gradient function.

$$f(x, y) = G(x, y) * f(x.y) \qquad (4)$$

3.2.2. Connected component determination

Connected component analysis is used to find connected region in the color image as well as in the

binary images. There may be character or broken letters. Each region which is in maximum is called a connected component.

3.2.3 Closing

The closing operation defined as, image I by a structuring element *T* is a dilation followed by erosion as given in equation (5)

$$I \bullet T = (I \oplus T_{rot}) \ominus T_{rot} \qquad (5)$$

The dilation operation uses a structuring element for probing and expanding the shapes contained in the input image, and the erosion will removes pixels on object boundaries. The closing and opening operation has mainly used for removing morphological noise. Which means opening remove small objects and closing removes small holes in a given document.

3.2.4. Y-histogram projection

Y-Histogram is performed on the previous results. The main aim of y-histogram projection is to find all possible line segments in the images

IV. Experimental result

The experiments are performed in camera captured document of textbook, which include 30 images of same viewing angle is considered, which does not contain broken or missing characters.Figure 2 shows input image. Figure 3 shows the canny edge detected result. Figure 4 and figure 5 shows connected components and text area respectively.

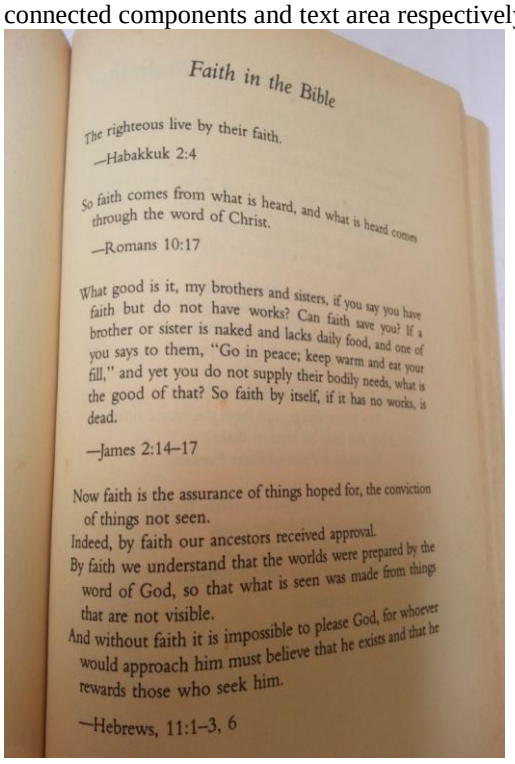

Fig.2.input image

Fig.3.canny edge detection

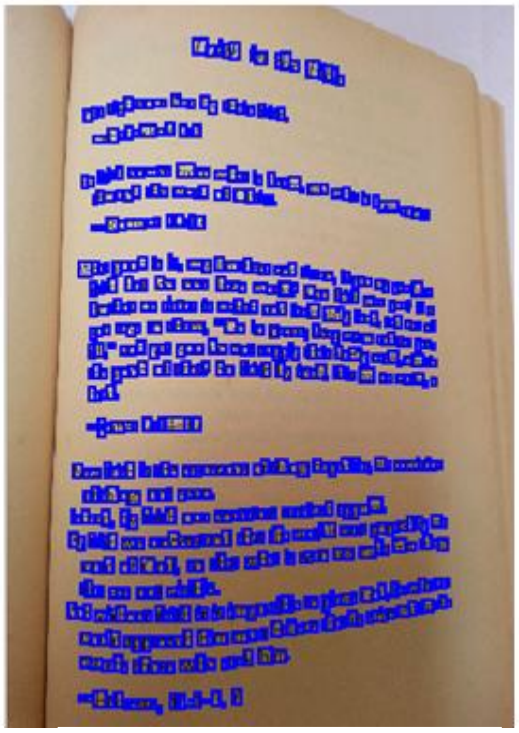

Fig.4.connected components

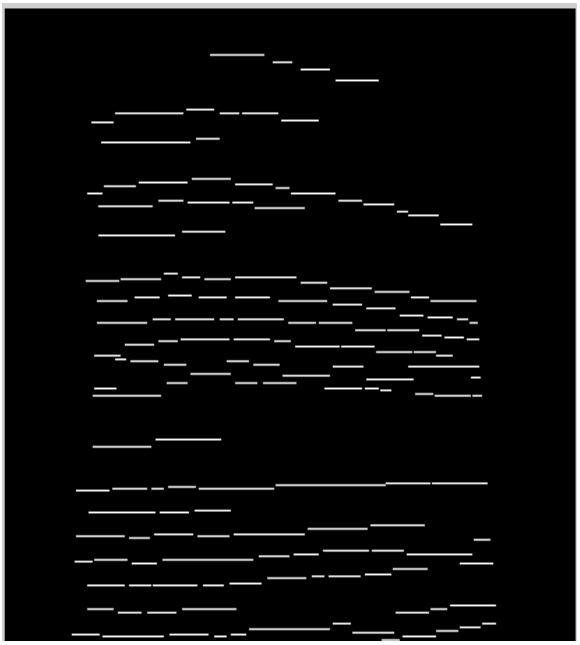

Fig.5.Segmented Text line

V. Conclusion

In this paper, we propose a model for extracting text line from a warped document captured by a camera. The proposed method is applied on grayscale document images and is based on an adaptive document image binarization and the text line detection. In preprocessing, adaptive thresholding method has been used which statistically examine the intensity values of the local neighborhood of each pixel. This method is useful when the text image contain strong illumination-gradient, then use of the average may prove to be effective in eliminating the ill influence. Text detection or extraction from the documents is the next step before segmentation. Here curled text lines are detected based on histogram and projection. Compare to previous method it is less sensitive to higher level of skew and the orientation.Future work focuses on flattening of the text line which are extracted by previous method. Rectification is method is mainly used for straighten the document one by one. Also we will concentrate on reducing the perspective and geometric distortions.

References

[1] Shijian Lu, Ben M. Chen , C.C. Ko ,A partition approach for the restoration of camera images of planar and curled document,*Conference on Image and Vision Computing* , 24, 2006, 837-848.

[2] Syed Saqib Bukhari, Faisal Shafait, Thomas.M. Breuel,Coupled snakelets for curled text-line segmentation from warped document images**,** *10th Int. Conf. on Document Analysis and Recognition,* 2011,33-53

[3] Partha Pratim Roy ·, Umapada Pal ·, Josep Lladós,Text line extraction in graphical documents using background and foreground information, *International Journal on Document Analysis and Recognition*,15, 2011, 227-241.

[4] Vaibhav, Gavali.Multi-Oriented and Text Line Extraction from documents, *International Journal of Computer Science and Mobile Computing*, 2,2013, 285-293

[5] G. Louloudis, B. Gatos, I. Pratikakis, K. Halatsis,A Block-Based Hough Transform Mapping for Text Line Detection in Handwritten Documents, *science direct Image and Vision Computing* . vol. 24, no. 10, Oct. 2010, pp.837-848

[6] Syed Saqib Bukhari and Thomas M. Breuel,"Text line information extraction from Grayscale camera-captured document images". *IEEE Trans. On Pattern Analysis and Machine Intelligence*, vol. 26, Oct. 2011, pp. 1295-1306

[7] Shafait, F., Keysers, D., Breuel, and T.M,Performance evaluation and benchmarking of six page segmentation algorithms, *IEEE Trans.Pattern Anal. Mach. Intell.* Vol.**30** pp.941–954 (2008)

[8] Aivind Due Trier and Torfinn Taxt,Evaluation of Binarization Methods for Document Images, *IEEE transactions on pattern analysis and machine intelligence,* vol. 3, march 1995

[9] Irina Rabaev, Ofer Biller, Jihad El-Sana, Klara Kedem,Text Line Detection in Corrupted and Damaged Historical Manuscripts, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol 34, April 2012, pp.707-722.

[10] E. Kavallieratou, N. Fakotakis, G. Kokkinakis,"Skew angle estimation for printed and handwritten documents using the Wigner–Ville distribution,Image and Vision Computing" vol .20 , pp.813–824.

[11] U. Pal, B.B. Chaudhuri, "Multi-oriented text lines detection and their skew estimation", *Third Indian Conference on Computer Vision, Graphics and Image Processing*,, pp. 270–275