BINARIZATION TECHNIQUE FOR DEGRADED DOCUMENT IMAGE

Rucha Vivek Patil[1], Pratiksha Satish Sarwade[2], Manjusha Bhaskar Pandit[3], Renuka Kalyan Pisal[4]

Prof. C.V.Longani[5]

[1] BE Computer, SRES' COE Kopargaon, SPPU, Maharashtra, India
[2] BE Computer, SRES' COE Kopargaon, SPPU, Maharashtra, India
[3] BE Computer, SRES' COE Kopargaon, SPPU, Maharashtra, India
[4] BE Computer, SRES' COE Kopargaon, SPPU, Maharashtra, India
[5]ME Computer, SRES' COE Kopargaon, SPPU, Maharashtra, India

ABSTRACT

Now a days we can easily store, duplicate and backup our data which is in digital form. But sometimes the old documents may contain important information .They may get degraded after long span of time. Many of these degraded documents may have their foreground mixed with background. To read text from such degraded documents is very challenging. We present binarized documentation technique to segment text from such badly degraded documents. In the presented technique first we generate adaptive image contrast map is by giving degraded document images as input later we detect the text stroke edge pixel. Text of the document is segmented by using local threshold estimation then further the post processing is applied to improve binarization quality of document. This technique is effective, easy and involves minimum parameters. The output of this technique will produce a clear and binarized image. This technique will be presented on android platform and will be easily available to user as android application.

Keyword: - *Binarized image, adaptive image contrast, degraded document image, post processing, and android.*

1. INTRODUCTION

Document Image Binarization aims in portioning the text in two parts mainly foreground text and back ground text. It is usually performed in the Preprocessing stage for document analysis. Binarization technique which is having high performance and taking less time is important for the ensuing document image processing tasks such as optical character recognition (OCR).

Document image binarization has been studied for many years, there are many thresholding techniques which are available but still the thresholding of degraded document images is still an unsolved problem due to the high inter/intra variation between the text stroke and the document background across different document images. The handwritten text within the degraded documents shows a certain amount of difference in terms of the stroke width, stroke brightness, stroke connection, and document background. Also historical documents are often degraded by the bleed through where the ink of the other side seeps through to the front. Historical documents also get degraded by different types of imaging artifacts. Thus the level of image contrast and amount of degradations tend leads to the document thresholding error and make degraded document image binarization a big challenge.

This paper presents a document binarization technique that extends previous local maximum-minimum method [3] and the method used in the latest DIBCO 2011. The proposed method is simple, robust. It can of handle different type of degraded document images also it require minimum parameters. It makes use of the adaptive image contrast that combines the local image contrast and the local image gradient adaptively and therefore is tolerant to the text and background variation caused by different types of document degradations. This paper is arranged as following sections: Section 2. describes Literature Review,Section 3 describes limitation of exiting system. Section 4 describes system overview document and Section 5 presents conclusion.

2. LITERATURE SURVEY

1. Otsu (1979): Otsu is the most efficient global thresholding method and was presented in 1979. It basically looks as the bar graph of an image. It considers the values of pixel and the property to obtain segments. It looks for the regions inside the segments that we want to segment. This method is not utilized in case of non-uniform background. Still, to get the rid of problems with non-uniform background, Otsu can be used in segments by using a moving window. Moving window migrates from different regions and then thresholds are evaluated for every region. In case of overlapping window an average of thresholds is calculated.

S.Lu, B.Su and C.L.Tan (2010)[8]: S.Lu,B.Su and C.L.Tan invented a document binarization method that uses text stroke edge information and the document background surface . It adaptively uses an iterative polynomial smoothing technique. The stroke edges are then identified on the local image variation within the document image which is compensated by using estimated document background surface. In last stage the local threshold is calculated.

J. Bernsen (1986) [7]: Bernsens method calculates the local threshold. It makes use of the mean value of the maximum and minimum intensities of pixels which are inside a window. When contrast is in large amount this local threshold works accurately.

J.Sauvola and M.Pietikainen (2000) [6]: They proposed a local thresholding algorithm. They computed Local threshold by calculating local standard deviation and local mean. Sauvola’s method is an substitution for Niblacks method which is guaranteed performance on documents that are having uneven brightness, light texture and huge amount variations in intensities.

J. Kittler and J. Illingworth(1985) [5] :They proposed an algorithm which is used when there are distinct object from background in grayscale images .Conditional probabilities of background and object class density functions are considered to be normal distributions. This algorithm solves the problem of Gaussian density fitting with minimum error and works by considering the difference of Gaussian density function as unequal. by using a histogram is used for classification of error for the fusion of two Gaussians. Kitler technique finds the optimal threshold at which the probability of error classification is minimum. This technique gives proper result only if the object and background are differentiable in terms of grey levels.

3. EXISTING SYSTEM LIMITATION

The exiting document binarization method has a few drawbacks they are as follows:
1. When we deal with ink –bleeding in this method the segmentation of text is easy when the text strokes on back-side are weaker as compared to front side of text. But when the back – side text are dark as compared to front side text the existing system cannot classify text stroke correctly.
2. The existing system was dependent on high contrast pixel on large scale. So it has high chance of introducing error if background of degraded document consist of certain amount pixel that have high contrast and also dense.

4. SYSTEM OVERVIEW

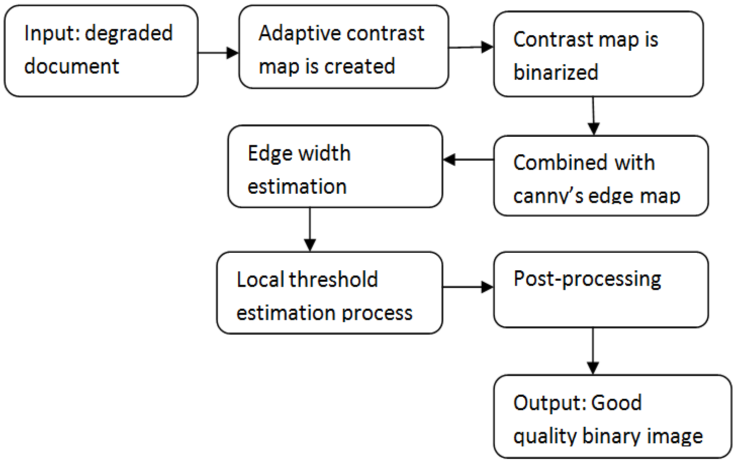
Input: degraded document
Adaptive contrast map is created
Contrast map is binarized
Combined with canny's edge map
Edge width estimation
Local threshold estimation process
Post-processing
Output: Good quality binary image

Fig -1 : System Overview

The proposed system is stated in Fig -1. The input of system will be degraded document image. This image will go through the basic image FIR filter to reduce noise. Then it has basic 4 modules as stated below:
A. Contrast image construction
B. Text Stroke Edge Pixel Detection
C. Local Threshold Estimation
D. Post-Processing

A. Contrast Image Construction

Adaptive image contrast is a combination of local image gradient and local image contrast. These are important factors which are used for segmentation of text from the document background . We do so because document text usually has certain image contrast with respect to the neighbouring document background. We are going to use the following formula to determine adaptive local image contrast. This formula removes the drawback of over normalization from existing system.

$Ca (i,j) = \alpha C(i, j)+(1 - \alpha) (Imax(i; j)- Imin(i; j))$.
Where,
1. $C(i, j)$: local contrast of image of pixel (i,j).
2. $(Imax(i; j)- Imin(i; j))$: local image gradient.
3. α : weight between local contrast and local gradient.
The above equation results in proper contrast maps of document images with different types of degradation. Ideally, the image contrast i.e. α will be assigned with a high weight especially when the document image has significant variation in intensity. So that the presented binarization technique depends more on the local image contrast that can capture the intensity variation well and hence it produces better results. The purpose of the contrast image construction is to detect the stroke edge pixels of the document text properly.

B. Text Stroke Edge Pixel Detection

As the local image contrast and the local image gradient are evaluated by the difference between the maximum and minimum intensity in a local window, the pixels at both sides of the text stroke will be selected as the high contrast pixels. The binary map can be further improved through the combination with the edges by Canny's edge detector because Canny's edge detector has a good localization property. In the combined map, we keep only pixels that appear within both the high contrast image pixel map and canny edge map. The combination helps to extract the text stroke edge pixels with less error and more accurately .

C. Local threshold estimation

We first detect the high stroke edge pixel properly and then carry out the extraction of text from document background pixels. There are two characteristics which have been observed by analyzing different kinds of

document images:
1) The detected text stroke pixels and text pixel are closed to each other.
2) There is difference between intensity of the high contrast stroke edge pixels and the surrounding background pixels.
So we can extract the text of document image which is based on the detected text stroke edge pixels. In this we are calculating the mean value. Here we are using the Edge width estimation algorithm as stated below:

Algorithm 1 Edge Width Estimation

```
Require: The Input Document Image $I$ and Corresponding Binary Text Stroke Edge Image $Edg$
Ensure: The Estimated Text Stroke Edge Width $EW$
1: Get the width and height of $I$
2: for Each Row $i = 1$ to height in $Edg$ do
3:     Scan from left to right to find edge pixels that meet the following criteria:
       a) its label is 0 (background);
       b) the next pixel is labeled as 1(edge).
4:     Examine the intensities in $I$ of those pixels selected in Step 3, and remove those pixels that have a lower intensity than the following pixel next to it in the same row of $I$ .
5:     Match the remaining adjacent pixels in the same row into pairs, and calculate the distance between the two pixels in pair.
6: end for
7: Construct a histogram of those calculated distances.
8: Use the most frequently occurring distance as the estimated stroke edge width $EW$.
```

D. Post Processing

After deriving the initial binarization result from above the method that binarization result can is be enhanced further as described in below Post processing procedure algorithm. It is stated as below:

Algorithm 2 Post-Processing Procedure

```
Require: The Input Document Image $I$ , Initial Binary Result $B$ and Corresponding Binary Text Stroke Edge Image $Edg$
Ensure: The Final Binary Result $B f$
1: Find out all the connect components of the stroke edge pixels in $Edge$.
2: Remove those pixels that do not connect with other pixels.
3: for Each remaining edge pixels $(i, j )$: do
4:     Get its neighborhood pairs: $(i − 1, j )$ and $(i + 1, j )$; $(i, j − 1)$ and $(i, j + 1)$
5:     if The pixels in the same pairs belong to the same class (both text or background) then
6:         Assign the pixel with lower intensity to foreground class (text), and the other to background class.
7:     end if
8: end for
9: Remove single-pixel artifacts [4] along the text stroke boundaries after the document thresholding.
10: Store the new binary result to $B f$ .
```

4.1 System flow diagram

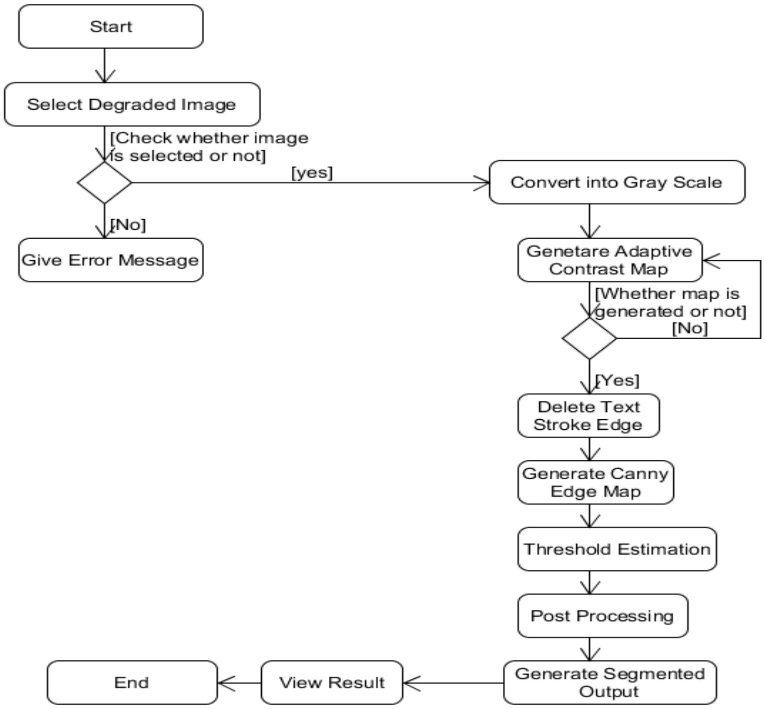

Fig-2 System Flow Diagram

Fig -2 shows the system flow diagram. It gives us detailed information about the Process which will lead to Binarized document image.

5. CONCLUSION

This paper presents a easy and efficient binarization technique which is capable of binarizing different types of document that are degraded due to various reason like ink seepage, ancient document, document distortion. This technique removes the drawback of existing system i.e. over normalization. The output image that will be obtained by this technique will be clear and binarized. It will prove very useful as it will be available as android application.

ACKNOWLEDGEMENT

We are extremely thankful to our guide, Prof. C. V. Logani, guidance assisted us to complete the task successfully. This experience will always steer us to do our work perfectly and professionally. We also extend our gratitude to Prof. D. B. Kshirsagar (H.O.D. Computer Department) and Dr. A. B. Pawar (Project Co-ordinator) who has provided facilities to explore and were always encouraged us. We are thankful to all the teachers and staff of the Department of Computer Engineering, Sanjivani Rural Education Society’s, College Of Engineering, Kopargaon for their co-operation and moral support.

REFERENCES

[1] Bolan Su, Shijian Lu, and Chew Lim Tan, “Robust Document Image Binarization Technique for Degraded Document Images”,IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 22, NO. 4, APRIL 2013.
[2] S. Lu, B. Su, and C. L. Tan, “Document image binarization using background estimation and stroke edges”, Int. J. Document Anal. Recognit.,vol. 13, no. 4, pp. 303314, Dec. 2010**.**
[3] B. Su, S. Lu, and C. L. Tan, “Binarization of historical handwritten document images using local maximum and minimum filter”,in Proc.Int. Workshop Document Anal. Syst., Jun. 2010, pp. 159166.
[4] N. Otsu, “A threshold selection method from gray level histogram”,IEEE Trans. Syst., Man, Cybern., vol. 19, no. 1, pp. 6266, Jan. 1979.
[5] J. Kittler and J. Illingworth, “On threshold selection using clustering criteria IEEE Trans. Syst., Man, Cybern., vol. 15, no. 5,pp. 652655, Sep.Oct. 1985.

[6] J. Sauvola and M. Pietikainen, “Adaptive document image binarization”Pattern Recognit., vol. 33, no. 2, pp. 225236, 2000.
[7] J. Bernsen, “Dynamic thresholding of gray-level images in Proc. Int. Conf. Pattern Recognit., Oct. 1986, pp. 12511255.
[8] B. Su, S. Lu, and C. L. Tan, “A self-training learning document binarization framework, in Proc. Int. Conf. Pattern Recognit., Aug. 2010, pp. 31873190.