RESEARCH ARTICLE OPEN ACCESS

Optical Character Recognition from Degraded Document Images

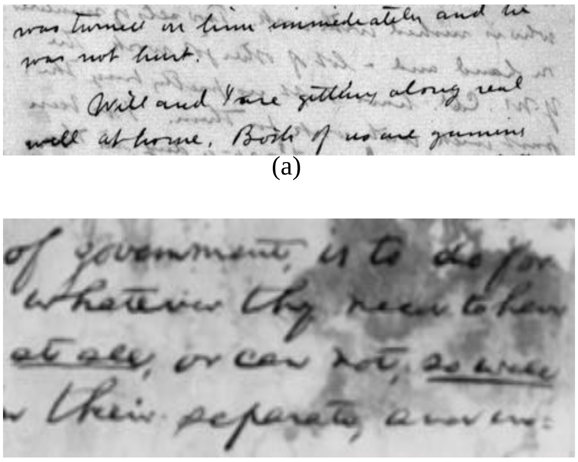
[1]P.R.Nisha Beevi, [2]Mr.S.Mohammed Nuhuman
[1]II M.E CSE Student, [2]Assistant Professor
Department of Computer Science & Engineering, National College of Engineering Maruthakulam, Tirunelveli-627 151 [1]nishabtech15@gmail.com,[2]nuhu24@gmail.com

Abstract
Segmentation of the text from badly degraded document images is very challenging tasks due to the high inter/intra variation between the document background and the foreground text of different types of document images. In this paper, a novel document image binarization technique is used to addresses the issues in the degraded document images by using adaptive image contrast. The adaptive image contrast is a combination of the local image contrast and the local image gradient that is tolerant to text and background variations caused by different types of document degradations. The adaptive contrast map is first constructed for an input degraded document image. Then the contrast map is then binarized and combined with the Canny's edge map to identify the text stroke edge pixels. The document text is further segmented by a local threshold that is estimated based on the intensities of detected text stroke edge pixels within a local window. Then apply the vertical scanning to find how many lines in the binary document image. After then apply the horizontal scanning to find how many characters in the image. Then the character is recognized using discrete wavelet transform.

***Index Terms*—**Adaptive image contrast, document analysis, document image processing, degraded document image binarization, pixel classification, character recognition.

I. INTRODUCTION

Document Image Binarization is performed in the pre-processing stage for document analysis and it aims to segment the foreground text from the document background. 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.



was turned in time immediately and he
was not hurt.
Will and I are getting along real
well at home. Both of us are gaining

(a)

of government, is to do for
whatever they need to have
at all, or can not, so well
in their separate, and indiv-

(b)

Fig.1.Degraded document image examples

As illustrated in Fig. 1, the handwritten text within the degraded documents often shows a certain amount of variation in terms of the stroke width, stroke brightness, stroke connection, and document background. In addition, historical documents are often degraded by the bleedthrough as illustrated in Fig. 1(a) where the ink of the other side seeps through to the front. In addition, historical documents are often degraded by different types of imaging artifacts as illustrated in Fig. 1(b). These different types of document degradations tend to induce the document thresholding error.

The binarization technique makes the use of the adaptive image contrast that combines the local image contrast and the local image gradient adaptively and that is tolerant to the text and background variation caused by different types of document degradations. Character recognition process is used to recognize the characters from the handwritten text using the discrete wavelet transform.

where C(i, j) denotes the contrast of an image pixel (i, j), Imax(i, j) and Imin(i, j) denote the maximum and minimum intensities within a local neighborhood windows of (i, j), respectively. If the local contrast C(i, j) is smaller than a threshold, the pixel is set as background directly. Otherwise it will be classified into text or background by comparing with the mean of Imax(i, j) and Imin(i, j).

A novel document image binarization method [2] by using the local image contrast that is evaluated as follows:

II. RELATED WORK

Document images often suffer from different types of degradation that renders the document image binarization a challenging task. The binarization technique first estimates a document background surface through an iterative polynomial smoothing procedure. Different types of document degradation are then compensated by using the estimated document background surface. The text stroke edge is further detected from the compensated document image by using image gradient [1]. Many thresholding techniques [3] have been reported for document image binarization. Adaptive thresholding, which estimates a local threshold for each document image pixel, is often a better approach to deal with different variations within degraded document images. Other approaches are including background subtraction [1], [4], texture analysis, recursive method [5], [6], decomposition method [7], contour completion, Markov Random Field [8], matched wavelet [9], Laplacian energy user assistance [10] and combination of binarization techniques. If word boundaries are known, we can force the model to interpret a given word image to return the highest scoring word [11], [12], [15].

The local image contrast and the local image gradient are very useful features for segmenting the text from the document background because the document text usually has certain image contrast to the neighbouring document background. They are very effective and have been used in many document image binarization techniques [2], [8], [9]. The local contrast is defined as follows:

$$C(i,j) = I\max(i,j) - I\min(i,j) \qquad (1)$$

where C(i, j) denotes the contrast of an image pixel (i, j), Imax(i, j) and Imin(i, j) denote the maximum and minimum intensities within a local neighborhood windows of (i, j), respectively. If the local contrast C(i, j) is smaller than a threshold, the pixel is set as background directly. Otherwise it will be classified into text or background by comparing with the mean of Imax(i, j) and Imin(i, j).
A novel document image binarization method [2] by using the local image contrast that is evaluated as follows:

$$C(i,j) = \frac{[I_{max}(i,j) - I_{min}(i,j)]}{I_{max}(i,j) + I_{min}(i,j) + \epsilon} \qquad (2)$$

where is a positive but infinitely small number that is added in case the local maximum is equal to 0. Compared with Equation 1, the local image contrast in Equation 2 introduces a normalization factor (the denominator) to compensate the image variation within the document background.

III. PROPOSED METHOD

This section describes about the document image binarization techniques. Given a degraded document image, an adaptive contrast map is first constructed and the text stroke edges are then detected through the combination of the binarized adaptive contrast map and the canny edge map. The text is then segmented based on the local threshold that is estimated from the detected text stroke edge pixels. Some post-processing is further applied to improve the document binarization quality. After then vertical and the horizontal scanning is used to find the lines and characters. Then the discrete wavelet transform is used to recognize the characters.

A. Contrast Image Construction

The image gradient has been widely used for edge detection and it can be used to detect the text stroke edges of the document images effectively that have a uniform document background. In earlier method [2], The local contrast evaluated by the local image maximum and minimum is used to suppress the background variation as described in Equation 2.

However, the image contrast in Equation 2 has one typical limitation that it may not handle document images with the bright text properly. This is because a weak contrast will be calculated for stroke edges of the bright text where the denominator in Equation 2 will be large but the numerator will be small. To overcome this over-normalization problem, combine the local image contrast with the local image gradient and derive an adaptive local image contrast as follows:

$$C_a(i,j) = \alpha C(i,j) + (1-\alpha)\big(I_{max}(i,j) - I_{min}(i,j)\big) \qquad (3)$$

Where *C(i, j)* denotes the local contrast in Equation 2. The *α* is the weight between local contrast and local gradient. Ideally, the image contrast will be assigned with a high weight (i.e. large *α*) when the document image has significant intensity variation. So that the binarization technique depends more on the local image contrast that can capture the intensity variation well and hence produce good results. Otherwise, the local image gradient will be assigned with a high weight. The binarization technique relies more on image gradient.

Model the mapping from the document image intensity variation to *α* by a power function as follows:

$$\alpha = \left(\frac{Std}{128}\right)^{\gamma} \qquad (4)$$

where *Std* denotes the document image intensity standard deviation, and *γ* is a pre-defined parameter. The power function has a nice property in that it monotonically and smoothly increases from 0 to 1. *γ* can be selected from [0,∞], where the power function becomes a linear function when $\gamma = 1$. The setting of parameter γ will be discussed in Section IV.

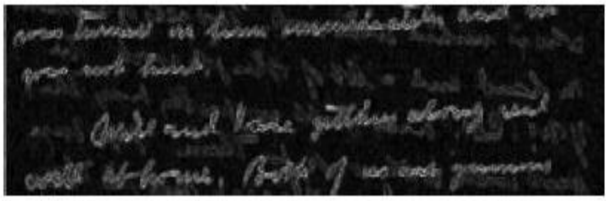


Fig.2. Contrast Images constructed using proposed method of the sample document images in Fig. 1(a)

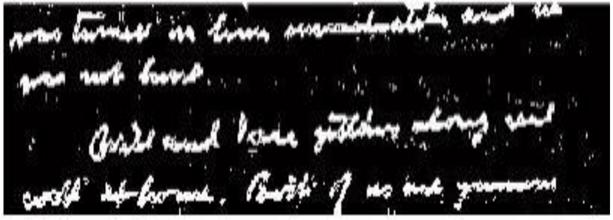
Fig. 2 shows the contrast map of the sample document image in Fig. 1 (a) that are created by using local image gradient, local image contrast [2] and adaptive local image contrast method in Equation 3, respectively.

As a comparison, the adaptive combination of the local image contrast and the local image gradient in Equation 3 can produce proper contrast maps for document images with different types of degradation.

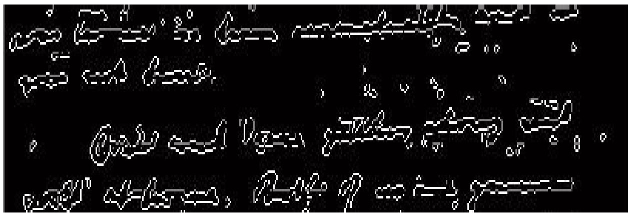
B. Text Stroke Edge Pixel Detection

The purpose of the contrast image construction is to detect the stroke edge pixels of the document text properly. Detection of the text stroke edge pixels by using the Otsu's global thresholding method. For the contrast images in Fig. 2, Fig. 3(a) shows a binary map by Otsu's algorithm that extracts the stroke edge pixels properly.

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. In addition, canny edge detector uses two adaptive thresholds and is more tolerant to different imaging artifacts such as shading. It should be noted that Canny's edge detector by itself often extracts a large amount of non-stroke edges as illustrated in Fig. 3(b) without tuning the parameter manually.



(a)



(b)

Fig.3. (a) Binary contrast map by Otsu's algorithm, (b) canny edge map.

In the combined map, keep only the 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 accurately.

C. Local Threshold Estimation

The text can then be extracted from the document background pixels once the high contrast stroke edge pixels are detected properly. The *EW*, which can be estimated from the detected stroke edges [shown in Fig. 3(b)] as stated in Algorithm 1.

```
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.
```

Since do not need a precise stroke width, just calculate the most frequently distance between two adjacent edge pixels (which denotes two sides edge of a stroke) in horizontal direction and use it as the estimated stroke width. After that a histogram is constructed that records the frequency of the distance between two adjacent candidate pixels. The stroke edge width *EW* can then be approximately estimated by using the most frequently occurring distances of the adjacent edge pixels.

D. Post-Processing

Once the initial binarization result is derived from the local threshold estimation, the binarization result can be further improved by incorporating certain domain knowledge as described in Algorithm 2.

```
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
Find out all the connect components of the stroke edge pixels in Edg.
Remove those pixels that do not connect with other pixels.
for Each remaining edge pixels (i, j): do
    Get its neighborhood pairs :(i − 1, j) and (i+ 1, j); (i, j − 1) and (i, j + 1)
    if The pixels in the same pairs belong to the same class (both text or background) then
        Assign the pixel with lower intensity to foreground class (text), and the other to background class.
    end if
end for
Remove single-pixel artifacts [4] along the text stroke boundaries after the document thresholding.
Store the new binary result to B f.
```

First, the isolated foreground pixels that do not connect with other foreground pixels are filtered out to make the edge pixel set precisely. Second, the neighborhood pixel pair that lies on symmetric sides of a text stroke edge pixel should belong to different classes (i.e., either the document background or the foreground text).

E. Line Detection using Vertical Scanning

After the binary image is get from the post-processing apply the vertical scanning to find how many lines in the binary document image.

The vertical scanning is described in Algorithm 3.

```
Algorithm 3 Vertical Scanning
Require: The Binary Result B f

Ensure: The Number of lines in the binary image
first get the binary image.
Then traverse from top to bottom of the image.
Through traversing find the white pixel's starting x position and starting y position.
Save the x and y position.
After completion of traversing count the number of starting and ending points.
Then divide the total count value by 2 to find the total number of lines in the given input image.
Finally crop the line from the image based on the starting and the ending point of the x and y position.
```

After the vertical scanning number of lines in the binary image is detected and then the line is cropped from the image.

F. Character Detection using Horizontal Scanning

The cropped line is get from the vertical scanning. The cropped line is used to find the number of characters in the image using the horizontal scanning. The horizontal scanning is described in the Algorithm 4.

```
Algorithm 4 Horizontal Scanning

Require: The Cropped line of the image from vertical scanning

Ensure: The Number of Characters in the binary image
first get the line image which is get from the vertical scanning process.
Then traverse from left to right of the image.
Through traversing find the white pixel's starting x position and starting y position.
Save the x and y position.
After completion of traversing count the number of starting and ending points.
Then divide the total count value by 2 to find the total no of character in the given input image.
Finally crop the character from the image based on the starting and the ending point of the x and y position.
```

After getting the number of characters from the image, crop the characters based on their starting and ending point of the x and y position.

G. Character Recognition using Discrete Wavelet Transform (DWT)

Number of characters is estimated from the document image using the horizontal scanning. Then the characters are recognized using the Discrete Wavelet Transform. The DWT algorithm is described in the Algorithm 5.

```
Algorithm 5 Character Recognition using DWT

Require: The Cropped line of the image from vertical scanning

Ensure: The Number of Characters in the binary image
1:  First get the training image folder which contains the image.
2:  Then apply the DWT on to the image, DWT split the image into four parts such LL,LH,HL and HH(The L means Low pass filtered image, H means High pass filtered image).
3:  Then find the average value of all the sub images.
4:  And then store the average values.
5:  Then get the input image which is to be get from the horizontal scanning.
6:  Then apply DWT on to the image.
7:  And then find the average value of these images.
8:  Finally find the difference value of this average value with the trained value images.
9:  Then find the minimum value to find the corresponding character.
10: Put these characters to the document.
```

After recognizing characters using the DWT, put that characters in the document.

IV. EXPERIMENTS AND DISCUSSION

The binarization performances are evaluated by using F-Measure, Peak Signal to Noise Ratio (PSNR), Negative Rate Metric (NRM), and Misclassification Penalty Metric (MPM).

A. Parameter Selection

The γ increases from 2^{-10} to 2^{10}. In particular, α is close to 1 when γ is small and the local image contrast C dominates the adaptive image contrast C_a in Equation 3. On the other hand, C_a is mainly influenced by local image gradient when γ is large. At the same time, the variation of α for different document images increases when γ is close to 1.

B. Performance Evaluation

In this experiment, quantitatively compare proposed method with other techniques. These methods include Otsu's method (OTSU), Sauvola's method (SAUV), Niblack's method (NIBL), Bernsen's method (BERN), Gatos et al.'s method (GATO). The document images that suffer from several common document degradations such as smear, smudge, bleed-through and low contrast.

Methods	F-Measure value	PSNR value	NRM value	MPM value	Rank Score value
OTSU	78.72	15.34	5.77	13.3	196
SAUV	85.41	16.39	6.94	3.2	177
NIBL	55.82	9.89	16.4	61.5	251
BERN	52.48	8.89	14.29	113.8	313
GATO	85.25	16.5	10	0.7	176
Proposed	93.5	19.65	3.74	0.43	100

TABLE: Performance Evaluation

Table shows, Adaptive Image Contrast method achieves the highest scores in F-Measure, PSNR, and NRM and its MPM. This means that proposed method produces a higher overall precision and preserves the text strokes better.

C. Discussion

As described in previous sections, the proposed method involves several parameters, most of which can be automatically estimated based on the statistics of the input document image. This makes our proposed technique more stable and easy-to-use for document images with different kinds of degradation. The superior performance of proposed method can be explained by several factors. First, the proposed method combines the local image contrast and the local image gradient that help to suppress the background variation and avoid the over-normalization of document images with less variation. Second, the combination with edge map helps to produce a precise text stroke edge map. Third, the proposed method makes use of the text stroke edges that help to extract the foreground text from the document background accurately.

V. CONCLUSION

This paper presents an adaptive image contrast based document image binarization technique that is tolerant to different types of document degradation such as uneven illumination and document smear. Moreover, it works for different kinds of degraded document images. The proposed technique makes use of the local image contrast that is evaluated based on the local maximum and minimum. Characters are recognized using the discrete wavelet transform. Experiments show that the proposed method outperforms most reported document binarization methods in term of the F measure, PSNR, NRM, and MPM.

REFERENCES

[1] 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. 303–314, Dec. 2010.

[2] 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. 159–166.

[3] G. Leedham, C. Yan, K. Takru, J. Hadi, N. Tan, and L. Mian, "Comparison of some thresholding algorithms for text/ background segmentation in difficult document images," inProc. Int. Conf. Document Anal. Recognit., vol. 13. 2003, pp. 859–864.

[4] B. Gatos, I. Pratikakis, and S. Perantonis, "Adaptive degraded document image binarization," Pattern Recognit., vol. 39, no. 3, pp. 317–327, 2006.

[5] M. Cheriet, J. N. Said, and C. Y. Suen, "A recursive thresholding technique for image segmentation," in Proc. IEEE Trans. Image Process., Jun. 1998, pp. 918–921.

[6] S. Kumar, R. Gupta, N. Khanna, S. Chaudhury, and S. D. Joshi, "Iterative multimodel subimagebinarization for handwritten character segmentation," IEEE Trans. Image Process., vol. 13, no. 9, pp. 1223–1230, Sep. 2004.

[7] Y. Chen and G. Leedham, "Decompose algorithm for thresholding degraded historical document images," IEE Proc. Vis., Image SignalProcess., vol. 152, no. 6, pp. 702–714, Dec. 2005.

[8] C. Wolf and D. Doermann, "Binarization of low quality text using a markov random field model," in Proc. Int. Conf. Pattern Recognit., 2002, pp. 160–163.

[9] S. Kumar, R. Gupta, N. Khanna, S. Chaudhury, and S. D. Joshi, "Text extraction and document image segmentation using matched wavelets and MRF model," IEEE Trans. Image Process., vol. 16, no. 8, pp. 2117–2128, Aug. 2007.

[10] H. Yi, M. S. Brown, and X. Dong, "User-assisted ink-bleed reduction," IEEE Trans. Image Process., vol. 19, no. 10, pp. 2646–2658, Oct. 2010.

[11] K. Wang and S. Belongie, "Word spotting in the wild," in Proc. European Conf. on Computer Vision, September 2010. [12]A. Mishra, K. Alahari, and C. V. Jawahar, "Top-down and bottomup cues for scene text recognition," in Proc. Conf. on Computer Vision and Pattern Recognition, pp. 2687–2694, 2012.

[13] H. Zhang, C. Liu, C. Yang, X. Ding, and K. Wang, "An improved scene text extraction method using conditional random field and optical character recognition," in Proc. Intl. Conf. on DocumentAnalysis and Recognition, pp. 708–712, 2011.

[14] T. Caesar, J. M. Gloger, and E. Mandler, "Estimating the baseline for written material," in Proc. Intl. Conf. on Document Analysis and Recognition, vol. 1, pp. 382–385, August 1995.

[15] C. Jacobs, P. Y. Simard, P. Viola, and J. Rinker, "Text recognition of low-resolution document images," in Proc. Intl. Conf. on Document Analysis and Recognition, pp.695–699, 2005.