Document Image Binarization Technique for Degraded Document Images

Supriya Lokhande
Student, M.E (VLSI &ES)
D.Y.Patil College of Engineering,
Ambi

N.A.Dawande
Professor,
Dept. of E&TC
D.Y.Patil College of Engineering, Ambi

ABSTRACT
Document image binarization is a vital pre-processing technique for document image analysis that segments text from badly degraded document images. In this paper, we propose a robust document image binarization technique that is based on the concept of adaptive image contrast. The adaptive image contrast which is formed by combining local image contrast and the local image gradient makes it tolerant to text and background variation caused by different types of document degradations. In the proposed technique the adaptive contrast map is binarized and text stroke edge pixels are detected using Canny's algorithm. The document text is further segmented by a local threshold that is assessed in light of the intensities of detected text stroke edge pixels within a local window. The above mentioned process has been rehashed by combining adaptive image contrast with Sobel's Edge detection technique and Total Variation Edge Detection technique respectively A comparison between these techniques is then made on the basis of Peak-signal to Noise Ratio and Mean Square Error values. These methods have been tested on images suffering from different types of degradations .It has been found out that adaptive image contrast used with Canny's edge detection technique gives the best results.

General Terms
Document image analysis, bimodal pattern, edge detection, segmentation.

Keywords
Adaptive image contrast, document image processing, degraded document image binarization**.**

1. INTRODUCTION
Document image binarization is performed in the preprocessing stage for document analysis. [1] It intends to segment foreground text from the background text. As illustrated in Fig.1 historical documents suffer from bleed through effect where the ink from the other side seeps through the front.

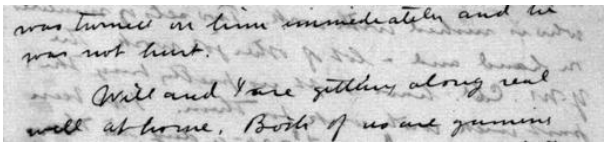

Fig 1.Degraded document image taken from DIBCO dataset

Fig.2 (a) shows a document image having a complex background and Fig 2(b) shows a document having small intensity variation within the document background but large intensity variation within the text strokes.

(a) (b)

Fig 2(a)-(b).Degraded document image taken from DIBCO dataset

Thus, due to different kinds of degradations, thresholding of degraded document images is a very challenging task. This paper presents a document image binarization technique that is simple, robust and involves minimum amount of parameter tuning. The rest of the paper is organized as follows. In section 2; a review of the state of art current binarization techniques has been provided. The mathematical model of the proposed technique has been described in Section 3.The implementation methodology has been described in detail in Section 4.The results have been described in Section 5.The performance evaluation of the techniques has been made on the basis of peak-signal to noise ratio and mean square error value in Section 6. Finally; conclusions are presented in Section 7.

2. RELATED WORK
In Otsu's method [8] cluster-based image thresholding, has been used for the reduction of a gray level image to a binary image. This algorithm tries to reduce combined spread (intra class variance) by assuming that the image contains two classes of pixels. It assumes that an image follows a bimodal histogram i.e. it contains foreground and background pixels. It then calculates the optimum threshold separating the two classes to ensure that its combined spread is minimal. This method gives acceptable results when the pixels in each class are close to each other. The limitations of this method are that many degraded document images do not have a clear bimodal pattern. Also another limitation is that minimization of intra class variances maximizes between class scatter. [3] Niblack's algorithm [4] calculates a pixel wise threshold by sliding a rectangular window over the gray level image The threshold *T* is computed by using the mean *m* and standard deviation *s*, for all the pixels within the window, and this threshold is denoted as:

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

where *k* is a constant, which determines how much of the total print object edge is retained, and has a value between 0 and 1. The value of *k* and the size *SW* of the sliding window defines the quality of binarization [9].The limitation of Niblack's method is that the resulting binary image suffers from a great amount of background noise especially in areas without text. [10].Another approach for document images binarization has

been adopted by Sauvola [5]. In this method the page is considered as a collection of subcomponents such as text, background and picture. To define a threshold for each pixel of the background and pictures a soft decision method is used. To define a threshold for each pixel of textual and line drawing areas a text binarization method is used. Finally the results of these algorithms are combined. [5].Although this method solves the problem posed by Niblack's approach but in many cases the characters become extremely thinned and broken.[10] In Bernsen's method [6] the local image contrast is defined as follows

$$C(i, j) = I_{max}(i, j) - I_{min}(i, j) \quad (2)$$

where C (i, j) denotes the contrast of an image pixel (i, j).

I_{max} (i, j) and I_{min} (i, j) denote the maximum and minimum intensities within a local neighborhood window of (i, j) respectively. The pixel will be classified into text or background by comparing with the mean of I_{max} (i, j), and I_{min} (i, j).If the local contrast C(i, j) is smaller than the threshold then the pixel is set as background and vice-versa. This method is simple. But the limitation of this method is that it does not work properly on degraded document images with a complex background [2]. Local maximum minimum method [7] is an improvement over Bernsen's method and handles the documents with a complex background well. In this method the local image contrast introduces a normalization factor. This normalization factor compensates for the image variation within the document background. Here the local image contrast is evaluated as follows:-

$$C(i, j) = \frac{I_{max}(i, j) - I_{min}(i, j)}{I_{max}(i, j) + I_{min}(i, j) + \epsilon} \quad (3)$$

where ϵ is a positive but infinitely small number that is added in case the local maximum is equal to 0. [2] In particular, the numerator (i.e. the difference between the local maximum and the local minimum) captures the local image difference that is similar to the traditional image gradient. The denominator acts as a normalization factor that lowers the effect of the image contrast and brightness variation. For image pixels within bright regions around the text stroke boundary, the denominator is large, which neutralizes the large numerator and accordingly results in a relatively low image contrast. But for image pixels within dark regions around the text stroke boundary, the denominator is small, which compensates the small numerator and accordingly results in a relatively high image contrast. [7] As a result, the contrasts of image pixels (lying around the text stroke boundary) within both bright and dark document regions converge close to each other and this facilitates the detection of high contrast image pixels lying around the text stroke boundary. [7] The limitation of this method is that, it cannot handle document images with bright text having bright background properly. A weak contrast is calculated for image pixels having bright text stroke edges and which lie within bright regions. A large denominator and a small numerator are produced for documents having bright text stroke edges and which lie within bright regions. This problem is known as over-normalization problem. The proposed technique overcomes the over-normalization problem by assigning weights to image contrast and image gradient. The mathematical model for the same has been discussed in the next section.

3. MATHEMATICAL MODEL FOR ADAPTIVE IMAGE CONTRAST

The adaptive image contrast solves the problem of over-normalization posed by the Local Maximum Minimum Method [7]. The image gradient gives better results for documents that have a uniform background. But it identifies many non-stroke edges from the document background. To extract solely the stroke edges properly, the image gradient must be normalized to compensate the image variation among the document background [1].The local image contrast evaluated with the help of equation (3) suffers from over-normalization problem as discussed in Section 2.To overcome these problems the local image contrast and the local image gradient have been combined and the equation of the adaptive image contrast has been derived [1].

$$C(i,j) = \alpha C(i,j) + (1-\alpha)(I_{max}(i,j) - I_{min}(i,j)) \quad (4)$$

where C(i, j) denotes the local contrast in Equation 2 and $(I_{max}(i, j) - I_{min}(i, j))$ refers to the local image gradient that is normalized to [0, 1]. The weight α between local contrast and local gradient is controlled based on the document image statistical information. [1] When the document image will have a significant intensity variation α will be assigned a higher weight .Otherwise the local image gradient will be assigned a high weight. The value of α is calculated as follows:

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

where Std denotes the document image intensity standard deviation and γ is a pre-defined parameter. γ can be selected from [0, ∞] [1].We are going to set the value of γ to 1.Local image contrast will play a major role when γ has a small value. The local image gradient will play a major role whenγ has a large value.

4. IMPLEMENTATION METHODOLOGY

Fig 3.shows the block diagram of the proposed system.

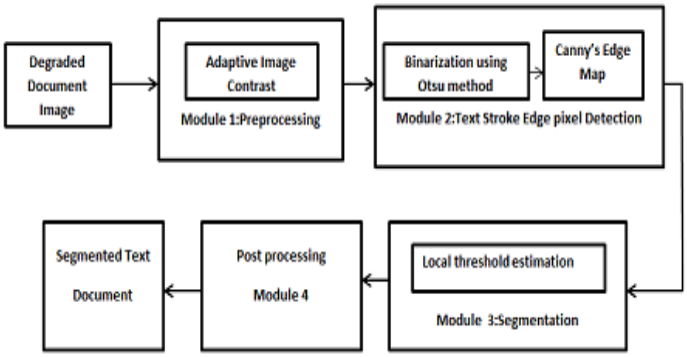

Fig 3.Block diagram of the proposed system.

The proposed system consists of four modules:

1. **Preprocessing Module**: In this module the adaptive image contrast is calculated with the help of equation (4) discussed in Section 3.
2. **Text Stroke Edge Detection Module**: In the text stroke edge detection module the text

stroke edge pixel candidates are detected by using Otsu method and Canny's edge detection algorithm.

3. **Segmentation Module:** In this the local threshold has been detected using Local Threshold Estimation algorithm discussed in section 4.1.
4. **Post Processing Module:** Finally, the restored image is produced using the post processing algorithm discussed in section 4.2.

4.1 Segmentation Module

The edge width estimation algorithm [1] calculates the adjacent pixel detection. It generates a histogram in which the distance between two adjacent edge pixels (which denotes two sides edge of a stroke) is plotted along the x-axis and the frequency of occurrence of this distance is denoted along the y-axis. The text width estimation algorithm has been described below [1]

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*.

The neighborhood window is selected as two times the edge width so that it contains the stroke edge pixels.

The document image text can thus be extracted based on the detected text stroke edge pixels as follows:

$$R(x,y) = \begin{cases} 1 & I(x,y) \le E_{mean} + \frac{E_{std}}{2} \\ 0 & otherwise \end{cases} \quad (5)$$

where E_{mean} and E_{std} are the mean and standard deviation of the intensity of the detected text stroke edge pixels within a neighborhood window *W*, respectively [1].

4.2 Post Processing Module

After the results are derived from Equation 5, the binarization results are further obtained using post processing algorithm described below [1].

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 *Edg*.

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 $B_{f.}$

4.3 Contributions for the development of the proposed system

The following subsections describe the modifications in the proposed system.

4.3.1 Use of Sobel's Edge Detection Method

To test the robustness of the system using Canny's edge detection method, the entire process of document image binarization discussed in the previous sections has been rehashed using Sobel's edge detection algorithm in place of Canny's edge detection algorithm in the test stroke edge pixel detection module. The Fig.4 shows the modified system.

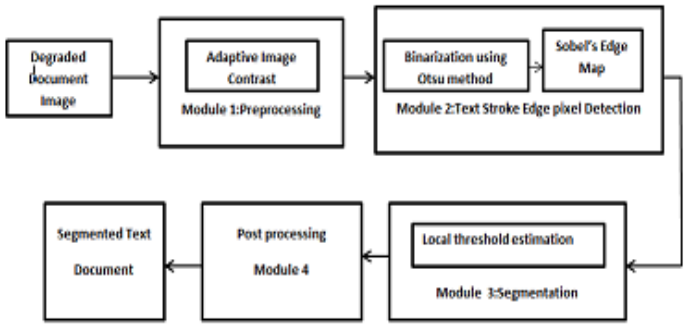

Fig.4 Block diagram of proposed system using Canny's edge detection method.

Sobel's edge detection technique is simple and requires fewer computations than Canny's edge detection method. However it is less accurate and more sensitive to noise than Canny's edge detection method.

4.3.2 Use of Total Variation Edge Detection Method:

The Sobel's edge detection technique has been replaced with total variation edge detection technique.Fig.5 shows the block diagram of the proposed system using total variation edge detection method.

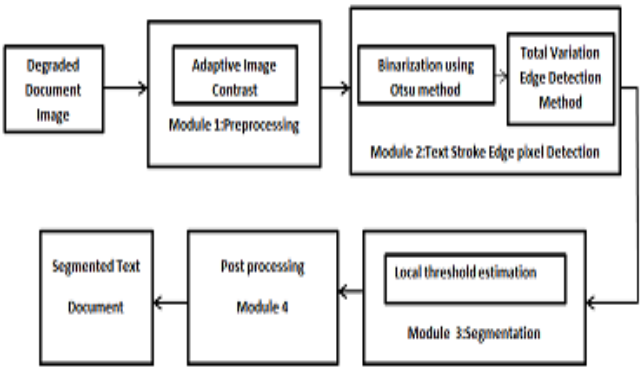

Fig.5 Block diagram of the proposed system using total variation edge detection method.

The total variation edge detection method finds the correct places of edges, and tests wider area around the pixel. But this technique malfunctions at the corners, curves and places in the document image where the gray level intensity function varies. The results obtained in Section 5.prove that Canny's edge map gives the highest value for Peak-Signal to Noise ratio and lowest value of mean square error.

5. RESULTS AND DISCUSSIONS

The sample document in Fig.6. has a complex document background .The Fig 7. shows the input which is a degraded document image

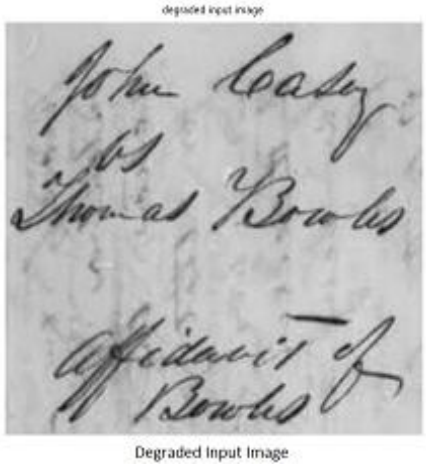

Fig.7 Degraded Input image

The Fig.8 shows the gradient image obtained using Equation 3

Fig.8.Gradient image

Fig.9 shows the Contrast image calculated using Equation 2.

Fig 9. Contrast Image

Since the document background of Fig.7 is complex the visual quality of the contrast image shown in Fig.9. is far better than the visual quality of the gradient shown in Fig.8.

Fig.10 shows the adaptive image contrast calculated using Equation 4.

Fig.10 Adaptive Image Contrast

As shown in the Fig.10, adaptive image contrast gives the best visual quality as compared to the gradient image and contrast image.

Fig 11 shows the results obtained by thresholding the image using Otsu's method.

Fig.11 Otsu's thresholded image

Fig.12 shows Canny's edge map

Fig.12 Canny's edge Map

It should noted that Canny's edge map detects a large number of edges.

Fig.13 shows the image obtained after combining Otsu with combined with Canny.

Fig.13. Otsu combined with Canny.

Fig.14 .shows the histogram of the image calculated using edge width estimation algorithm in subsection 4.1

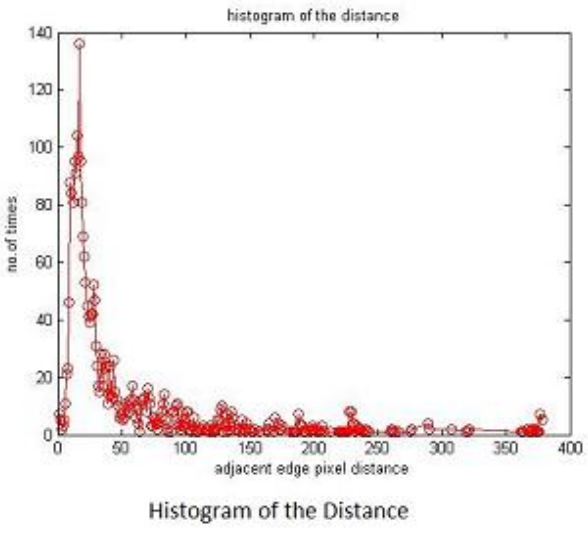

Fig.14.Histogram of the distance.

The edge width is 25 units and the frequency of occurrence of this distance is 139.

Fig.15 (a) shows the sub-binarization image obtained using equation 5.Fig.15 (b) shows the classification image obtained by using steps 1-6 of post processing algorithm.

Fig.15 (a) Sub-binarization image (b) Classification Image

Fig.16 shows the restored image.

Fig.16 Restored image.

Fig.17 shows the image obtained by using Sobel's edge detection algorithm.

Fig.17 Sobel's edge map

It detects fewer edges as compared to Canny's edge map

Fig.18 shows Otsu method combine with Sobel's method.

Fig.18 Otsu combined with Sobel

Fig.19 shows the histogram of the image calculated using edge width estimation algorithm in subsection 4.1

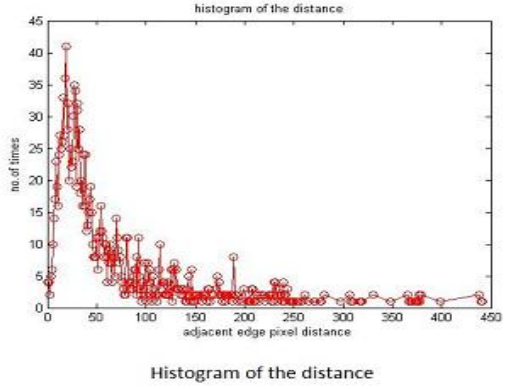

Fig.19 Histogram of the distance.

The edge width is 25 units and the frequency of occurrence of this distance is 42.

Fig 20(a) shows the sub-binarization image obtained using equation 5.Fig.20 (b) shows the classification image obtained by using steps 1-6 of post processing algorithm.

Fig.20 (a) Sub-binarization image (b) Classification Image

Fig 21 shows the restored image.

Fig.21 Restored Image

Fig.22 shows the Otsu method combine with Sobel's method.

Fig.22 Otsu combined with total variation

Fig.23 shows the histogram of the image calculated using edge width estimation algorithm in subsection 4.1

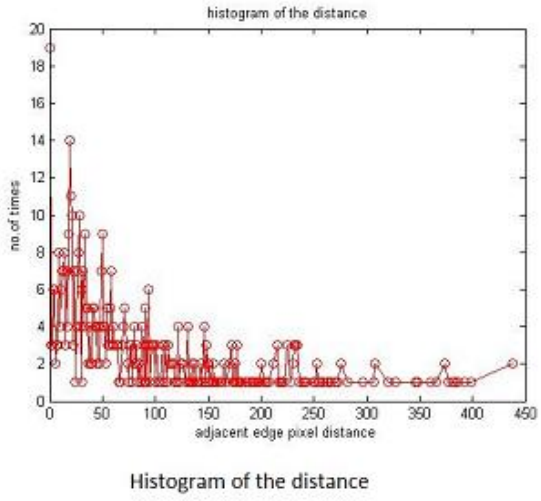

Fig.23 Histogram of the distance.

The edge width is 15 units and the frequency of occurrence of this distance is 14.

Fig 24 shows the restored image.

Fig.24 Restored Image

Similar results have been obtained for adaptive thresholded image

Fig.25 shows the restored image obtained for adaptive thresholded by using Canny's edge detection technique.

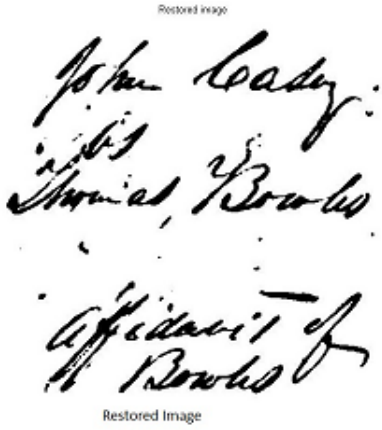

Fig.25 Restored image obtained for adaptive thresholded by using Canny's edge detection technique.

Fig.26 shows the restored image obtained for adaptive thresholded by using Sobel's's edge detection technique.

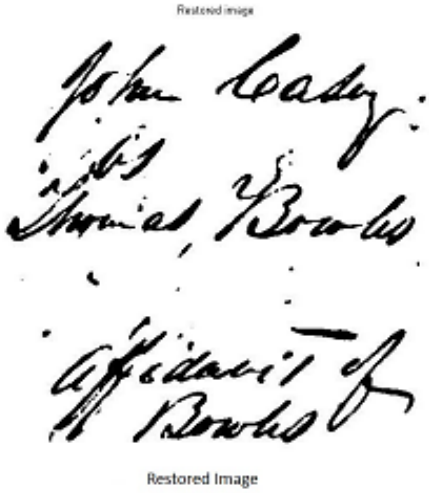

Fig.26. Restored image obtained for adaptive thresholded by using Sobel's edge detection technique.

Fig.27 shows the restored image obtained for adaptive thresholded by using Total variation edge detection technique.

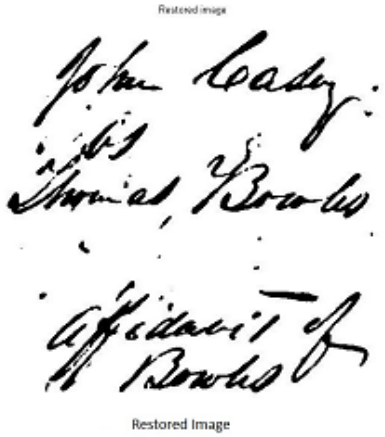

Fig.27. Restored image obtained for adaptive thresholded by using Total Variation edge detection technique.

Fig.28 shows the value of PSNR and MSE obtained for the methods obtained above.

	PSNR	MSE
otsu_canny	58.0385	0.1021
otsu_sobel	58.0385	0.1021
otsu_tv	58.0630	0.1016
adapt_canny	59.2753	0.0768
adapt_sobel	59.2752	0.0768
adapt_tv	59.2752	0.0768

Fig.28 Comparative analysis of results.

It can be clearly observed that adaptive thresholded image combined with Canny gives the best results

6. PERFORMANCE EVALUATION.

The performance of the above methods has been evaluated for Fig. 7 with the help of PSNR and MSE values [2].A higher value of PSNR indicates that the algorithm under consideration has enhanced a degraded image to more closely resemble the original image.A low value of MSE indicates that the error by which the original image differs from degraded image is less.

Table 1.shows the comparison of the above images on the basis of PSNR and MSE values

Table 1.Comparison of various methods

Sr.No	Method	PSNR in dB	MSE
1	Otsu [8]	9.9601	0.1009
2	Niblack [4]	58.571151	0.090358
3	Sauvola [5]	58.507285	0.091696

4	Proposed Otsu Canny	58.0385	0.1021
5	Proposed Otsu Sobel	58.0385	0.1021
6	Proposed Otsu Total Variation	58.0630	0.1016
7	Proposed Adapt Canny	58.2753	0.0768
8	Proposed Adapt Sobel	58.2752	0.0768
9	Proposed Adapt Total Variation	58.2752	0.0768

From the table it can be clearly observed that adaptive thresholded image combined with Canny gives the best results

7. CONCLUSION

The proposed technique is simple and robust. Since this method combines the image contrast and the image gradient it can handle all types of degradations in the document images**.** It overcomes the problem of over -normalization. The adaptive thresholded image combined with Canny gives superior performance in terms of Peak-Signal to Noise ratio and Mean Square Error when compared with existing techniques which were proposed by Otsu, Niblack and Sauvola. The performance of the proposed system can be compared with existing systems on the basis of Misclassification Penalty Metric-Measure and Distortion Reduction Matrix.

8. 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] I. Pratikakis, B. Gatos, and K. Ntirogiannis, ‘‘ICDAR 2011 document image binarization contest (DIBCO 2011),’’ in *Proc. Int. Conf. Document Anal. Recognit.*, Sep. 2011, pp. 1506--1510.

[3] M. Sezgin and B. Sankur, ‘‘Survey over image thresholding techniques and quantitative performance evaluation,’’ *J. Electron. Imag*, vol. 13,

[4] W. Niblack, *An Introduction to Digital Image Processing*. Englewood Cliffs, NJ: Prentice --Hall, 1986 no 1, pp. 146-165, Jan 2004

[5] J. Sauvola and M. Pietikainen, ‘‘Adaptive document image binarization,’’ *Pattern Recognit.*, vol. 33, no. 2, pp. 225--236, 2000.

[6] J. Bernsen, ‘‘Dynamic thresholding of gray-level images,’’ in *Proc. Int. Conf. Pattern Recognit.*, Oct. 1986, pp. 1251--1255.

[7] 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.

[8] N. Otsu, ‘‘A threshold selection method from gray level histograms,’’ *IEEE Trans. Syst. Man Cybern.* SMC-9, 62–66 ~1979.

[9] B.Gatos, I. Pratikakis, S.Perantonis, ‘An Adaptive Binarization Technique for Low Quality Historical Documents ‘, Springer-Verlag pp102-113, 2004

[10] B. Gatos, I. Pratikakis, and S. Perantonis, ‘‘Adaptive degraded document image binarization,’’ *Pattern Recognit.*, vol. 39, no. 3, pp 317--327, 2006