Improving Degraded Document Images Using Binarization Technique

Sayali Shukla, Ashwini Sonawane, Vrushali Topale, Pooja Tiwari

Abstract: Image segmentation is a set of segments that collectively cover the entire image, or a set of contours extracted from the image. In the process of improving degraded document images segmentation is one of the difficult task due to background and foreground variation. This paper presents a new approach for enhancement of degraded documents. It consists of an adaptive image contrast based document image binarization technique that is tolerant to different type of document degradation such as uneven illumination document smear involving smudging of text, seeping of ink to the other side of page, degradation of paper ink due to aging etc. The images i.e. scanned copies of these degraded documents are provided as an input to the system. They are processed to get the finest improved document so that the contents are visible readable. Contrast image construction can be constructed using local image gradient and local image contrast. Further edge estimation algorithm is used to identify the text stroke edge pixels .The text within the document is further segmented by a thresholding technique which is based on the height and width of letter size present in degraded document image. It works for different format of degraded document images. The method has been tested on Document Image Binarization Contest (DIBCO) experiments on Bickley diary dataset, consists of several challenging degraded document images.

Keywords: Binarization, Adaptive Image Contrast, Local Image Contrast, Local Image Gradient, Detection of Text Stroke Edges, Pixel Classification, Thresholding.

1 INTRODUCTION

To analyze the document, its image is binarized before processing it. It is nothing but segmenting the document background & the foreground text. For the confirmation of document image processing task an accurate document image binarization technique is a must. After years of studies in document image binarization, the thresholding of degraded document images is still found to be a challenging task because of the high inter/intra variation between the text stroke and the document background across various document images. The stroke width, stroke brightness, stroke connection, and document background vary in the handwritten text within the degraded documents. Moreover, bleed through degradation is observed in historical documents by variety of imaging outputs. For most of the existing techniques many kinds of document degradations, it is still an unsolved problem of degraded document image binarization due to the document thresholding error. A document image binarization technique presented in this paper is an extended version of an existing local maximum minimum method [5].

- *Author Sayali Shukla currently pursuing batchers degree program in computer engineering in Pune University, Country - India, PH - 09420701209. E-mail: sayalishukla100@gmail.com*
- *Co-Author Ashwini Sonawane currently pursuing batchers degree program in computer engineering in Pune University, Country - India, PH-08080261461. E-mail: ashut61@gmail.com*
- *Co-Author Vrushali Topale currently pursuing batchers degree program in computer engineering in Pune University, Country – India, PH- 08806497235. E-mail: vrushali.topale024@gmail.com*
- *Co-Author Pooja Tiwari currently pursuing batchers degree program in computer engineering in Pune University, Country - India, PH- 08552929293. E-mail: tiwari.pooja113@gmail.com*

The method can handle different degraded document images with least number of parameters, making it simple & robust. It uses the adaptive image contrast which is a combination of local image contrast & local image gradient. Thus it is capable of tolerating the text & background variation induced by different types of document degradation. The organization of the rest of the paper is stated as follows. Section II first describes the current document section III. Section V reports the experimental results to demonstrate the superior performance of our framework. Finally, the conclusions are presented.

2 DETAILS EXPERIMENTAL

2.1 Sub-block classification and thresholding

The three feature vectors described below were used to test the local regions and classify them into three types: heavy strokes, faint strokes or background. Typical examples of these three types of regions are shown in Fig.2.The background of a document does not contain any useful content information. A background area typically has lower values of edge strength and variance. A background which is totally noise-free also has a small mean-gradient value. Faint stroke areas contain faint strokes, which are very difficult to detect from the background. This kind of area typically has a medium value of edge strength and mean gradient but less variance. Heavy stroke areas have strong edge strength, more variance and larger mean-gradient value. The proposed weighted gradient thresholding method is applied to the different classes of sub block.

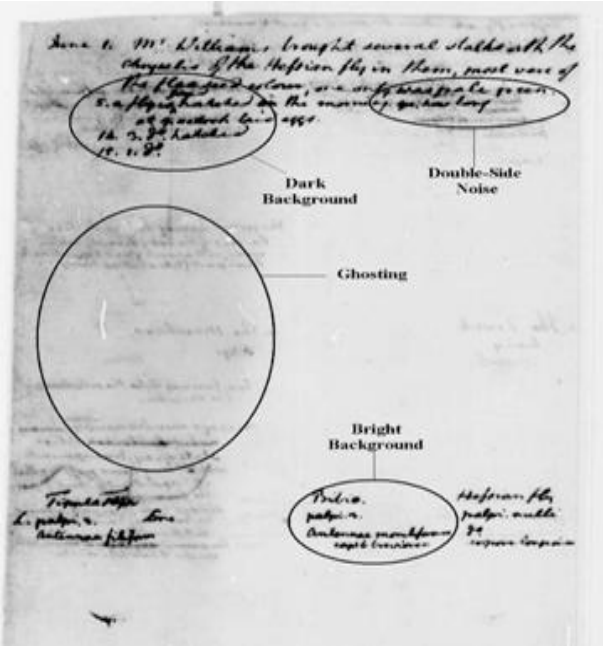

Fig.1. Example of typical historical document image

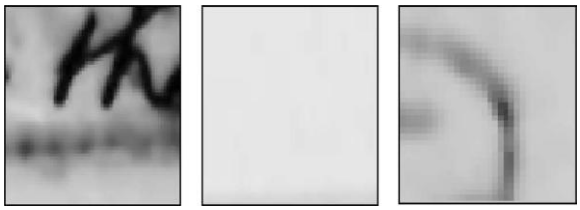

Fig.2. Examples of sub-regions containing (a) heavy strokes, (b) background (no strokes) and (c) faint strokes.

2.2 Faint handwritten image

2.2.1 Enhancement

Enhancement of faint strokes is necessary for further processing. To avoid the enhancement of noise, a Wiener filter was first applied. The enhancement can be divided into two steps.

1. Use 3x3 windows to enhance the image by finding the maximum and minimum grey value in the window.
2. Mini =min (elements in the window)
   Maxi = max (elements in the window)

Compare 'pixel – mini' and 'maxi – pixel', where 'pixel' is the pixel-value. If the former is greater, the 'pixel' is closer to the highest grey value than the lowest value in this window; hence the value of 'pixel' is set to the highest grey value ('pixel'='maxi'). If the former is smaller, then the value of 'pixel' is set to the lowest grey value ('pixel'='mini').

2.2.2 Thresholding

A new weighted method based on mean gradient direction is proposed for thresholding [6]-[10] faint strokes. Handwritten English or Western-style scripts normally contain strokes written in several directions.

2.2.3 Background area

This area is simply set to white (grey-scale value 255).

3 PROPOSED METHOD

Given a degraded document image, an adaptive contrast map is constructed and the text stroke edges are then detected. The text is then segmented based on the local threshold that is estimated from the detected text stroke edge pixels. After post processing is applied to improve the document binarization quality.

Fig. 3.. Five degraded document image examples (a)–(d) are taken from DIBCO series datasets and (e) is taken from Bickley diary dataset.

3.1 Contrast Image Construction

The purpose of the contrast image construction is to detect the stroke edge pixels of the document text properly. The constructed contrast image consists of a clear bi-modal pattern. It can be used to detect the text stroke edges of the document images that have a uniform document background. While, it often detects many non stroke edges from the background of degraded document that perhaps contains certain image variations because of uneven lighting, noise, bleed-through etc. For proper extraction of only the stroke edges, the image gradient needs to be normalized to compensate the variation in the image within the document background. The local contrast evaluated by the local image maximum and minimum is used to suppress the background variation as described in below Equation. 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 is a normalization factor that suppresses the image variation within the document background. For pixels of image within bright regions, it will generate a greater normalization factor to neutralize the numerator and accordingly result in a relatively low image contrast. For the pixels of image within dark regions, it will generate a small denominator and accordingly result in a relatively high image contrast.

$$c(i,j) = Imax(i,j) - Imin(i,j) \ldots\ldots\ldots\ldots\ldots (1)$$

$$c(i,j) = \frac{Imax(i,j) - Imin(i,j)}{Imax(i,j) + Imin(i,j) + \epsilon} \ldots\ldots\ldots (2)$$

$$Ca(i,j) = \propto C(i,j) + (1 - \propto)(Imax(i,j) - Imin(i,j)) \ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

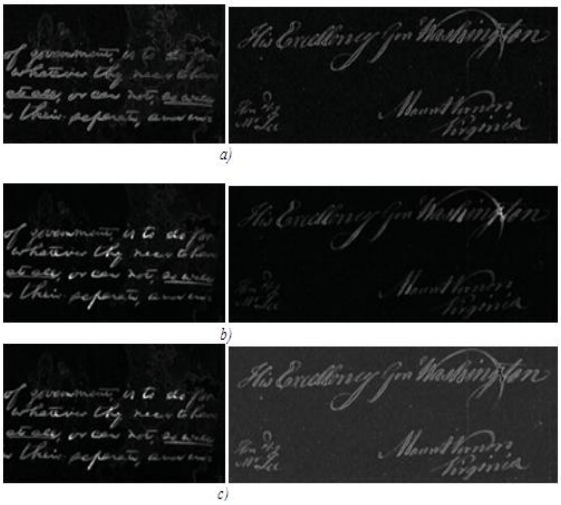

Fig.4. Contrast Images constructed using (a) local image gradient, (b) local image contrast, and (c) our proposed method of the sample document images in Fig. 3 (b) and (d)

3.2 Text stroke edge pixel detection

We get the stroke edge pixels of the document text properly from contrast image construction. The constructed contrast image consist a clear bi-modal pattern [5]. The local image gradient is evaluated by the difference between the maximum and minimum intensity in a local window, the pixels at both the sides of the text stroke will be selected as the high contrast pixels. Binary map is then constructed. In the combined map, we keep only pixels that appear within both the high contrast image pixel map and canny edge map. Accurate extraction of the text stroke edge pixels is helped out by this combination.

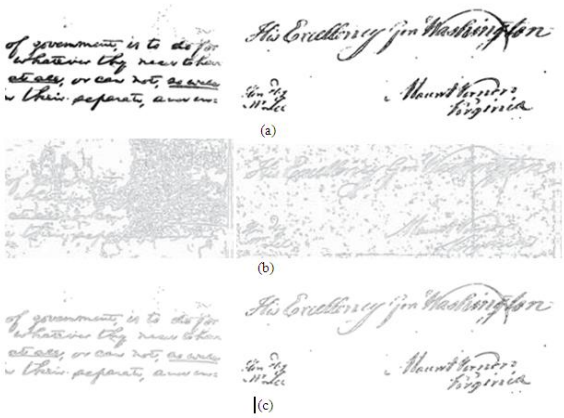

Fig.5. (a) Binary contrast maps, (b) canny edge maps, and their (c) combine edge maps of the sample document images in Fig. 3(b) and (d), respectively.

3.3 Local threshold estimation

Subsequent extraction of the text from the document background pixels is carried out once the high contrast stroke edge pixels are detected properly. Two characteristics can be observed from different kinds of document images First, the text pixels are close to the detected text stroke edge pixels. Second, the distinct intensity difference between the high contrast stroke edge pixels and the surrounding background pixels. The text within the document image can therefore be extracted 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 is as follows:

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 Edge 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. Utilize the most frequently occurring distance as the estimated stroke edge width EW.

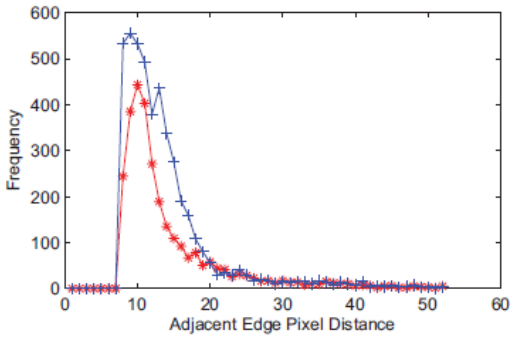

Fig. 4. Histogram of the distance between adjacent edge pixels. The "+++"line denotes the histogram of the image in Fig. 3(b). The "***" line denotes the histogram of the image in Fig. 3(d).

3.4 Post Processing

After deriving the initial binarization result from above the method that binarization result can further is improved as described in below Post processing procedure algorithm. Require: The Input Document Image I , Initial Binary Result B and Corresponding Binary Text Stroke Edge Image Edge

Ensure: The Final Binary Result B

1. Look for all the connect components of the stroke edge pixels in Edge
2. Eliminate those pixels that are not connected 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 and background) then
6. Allot the pixel with lower intensity to foreground class (text), and the other to background class.
7. end if
8. end for
9. Eliminate single-pixel artifacts along the text stroke boundaries after the document thresholding.
10. 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). A single pixel out of the pixel pair is therefore labeled to the other category if both of the two pixels belong to the same class. At last, certain numbers of single-pixel artifacts along the text stroke boundaries are filtered out by using several logical operators.

4 EXPERIMENTS AND DISCUSSION

To demonstrate the effectiveness and robustness of our method, we first analyze the performance of the proposed technique on public datasets for parameter selection. The proposed technique is then tested and compared competition datasets: DIBCO 2009 dataset [1], H-DIBCO 2010 dataset[3], and DIBCO 2011 dataset[2]. Further we evaluate over a very challenging Bickley diary dataset which is written 100 years ago. The binarization performance are evaluated by using F-Measure, pseudo F-Measure, Peak Signal to Noise Ratio (PSNR), Negative Rate Metric (NRM), Misclassification Penalty Metric (MPM), Distance Reciprocal Distortion (DRD) and rank score that are adopted from DIBCO 2009, H-DIBCO 2010 and DIBCO 2011[1]- [3], no all of the metrics are applied on every images, due to lack of ground truth data in some dataset.

A Parameter Selection

In the first experiment, we apply different γ to obtain different power functions and test their performance under the DIBCO 2009 and H-DIBCO 2010 datasets. The γ increases from 2−10 to 210 exponentially and monotonically. Another parameter, i.e., the local window size W, is tested in the second experiment on the DIBCO 2009, H-DIBCO 2010 and DIBCO 2011 datasets. W is closely related to the stroke width EW. Generally, a larger local window size will help to reduce the classification error that is often induced by the lack of edge pixels within the local neighborhood window. In addition, the performance of the proposed method becomes stable when the local window size is larger than 2EW consistently on the three datasets. W can therefore be set around 2EW because a larger local neighborhood window will increase the computational load significantly.

B Testing on Competition Datasets

In this experiment, we quantitatively compare our proposed method with other state-of-the-art techniques on DIBCO 2009, H-DIBCO 2010 and DIBCO 2011 datasets. These methods include Otsu's method (OTSU) [12], Sauvola's method (SAUV) [18], Niblack's method (NIBL) [19], Bernsen's method (BERN) [14], Gatos et al.'s method (GATO) [21], and our previous methods (LMM[5] , BE] [4]). The three datasets are composed of the same series of document images that suffer from several common document degradations such as smear, smudge, bleed-through and low contrast. The DIBCO 2009 dataset contains ten testing images that consist of five degraded handwritten documents and five degraded printed documents. The H-DIBCO 2010 dataset consists of ten degraded handwritten documents. The DIBCO 2011 dataset contains eight degraded handwritten documents and eight degraded printed documents. In total, we have 36 degraded document images with ground truth.

TABLE I

EVALUATION RESULTS OF THE DATASET OF DIBCO 2009

Methods	F-Measure (%)	PSNR	NRM ($\times 10^{-2}$)	MPM ($\times 10^{-3}$)	Rank Score
OTSU [12]	78.72	15.34	5.77	13.3	196
SAUV [18]	85.41	16.39	6.94	3.2	177
NIBL [19]	55.82	9.89	16.4	61.5	251
BERN [14]	52.48	8.89	14.29	113.8	313
GATO [21]	85.25	16.5	10	0.7	176
LMM [5]	91.06	18.5	7	0.3	126
BE [4]	91.24	18.6	4.31	0.55	101
Proposed Method	93.5	19.65	3.74	0.43	100

TABLE II

EVALUATION RESULTS OF THE DATASET OF H-DIBCO 2010

Methods	F-Measure (%)	Pseudo F-Measure (%)	PSNR	NRM ($\times 10^{-2}$)	MPM ($\times 10^{-3}$)	Rank Score
OTSU [12]	85.27	90.83	17.51	9.77	1.35	188
SAUV [18]	75.3	84.22	15.96	16.31	1.96	225
NIBL [19]	74.1	85.4	15.73	19.06	1.06	263
BERN [14]	41.3	44.4	8.57	21.18	115.98	244
GATO [21]	71.99	74.35	15.12	21.89	0.41	284
LMM [5]	85.49	92.64	17.83	11.46	0.37	216
BE [4]	86.41	88.25	18.14	9.06	1.11	202
Proposed Method	92.03	94.85	20.12	6.14	0.25	178

TABLE III
EVALUATION RESULTS OF THE DATASET OF DIBCO 2011

Methods	F-Measure (%)	PSNR	DRD	MPM	Rank Score
OTSU [12]	82.22	15.77	8.72	15.64	412
SAUV [18]	82.54	15.78	8.09	9.20	403
NIBL [19]	68.52	12.76	28.31	26.38	362
BERN [14]	47.28	7.92	82.28	136.54	664
GATO [21]	82.11	16.04	5.42	7.13	353
LMM [5]	85.56	16.75	6.02	6.42	516
BE [4]	81.67	15.59	11.24	11.40	376
LELO [45]	80.86	16.13	104.48	64.43	252
SNUS	85.2	17.16	15.66	9.07	279
HOWE [36]	88.74	17.84	5.37	8.64	299
Proposed	87.8	17.56	**4.84**	**5.17**	307

5 EXPERIMENTAL RESULTS

The evaluation measures are adapted from the DIBCO report including F-measure, peak signal-to-noise ratio (PSNR), negative rate metric (NRM), and misclassification penalty metric (MPM). In particular, the F-measure is defined as follows:

$$FM = \frac{2 * RC * PR}{RC + PR}$$

where RC and PR refer the binarization recall and the binarization precision, respectively. This metric measures how well an algorithm can retrieve the desire pixels. The PSNR is defined as follows:

$$PSNR = 10\log(\frac{C^2}{MSE})$$

where MSE denotes the mean square error and C is a constant and can be set at 1. This metric measures how close the result image to the ground truth image. The NRM is defined as follows:

$$NRM = \frac{\frac{Nfn}{Nfn + Ntp} + \frac{Nfp}{Nfp + Ntn}}{2}$$

where Ntp, Nfp, Ntn, Nfn denote the number of true positives, false positives, true negatives, and false negatives respectively. This metric measures pixel mismatch rate between the ground truth image and result image. The MPM is defined as follows:

$$MPM = \frac{\sum_{i=1}^{Nfn} d_{fn}^{i} + \sum_{j=1}^{Nfp} d_{fp}^{j}}{2D}$$

Where di FN and d j fp denote the distance of the ith false negative and the j th false positive pixel from the contour of the ground truth segmentation. The normalization factor D is the sum over all the pixel-to-contour distances of the ground truth object. This metric measures how well the result image represents the contour of ground truth image

6 DISCUSSION

Our proposed method involves several parameters, from most of which can be automatically calculated based on the statistics of the input document image. This method makes our proposed technique more stable and easy to read for document images with variety of degradation. It gives superior performance by explaining the several factors. First combination of the local image contrast and the local image gradient that help to suppress the background variation and avoid the over-normalization of document images. Second, the combination with edge map helps to produce a precise text stroke edge map. Then, the proposed method makes use of the text stroke edges that help to extract the foreground text from the document background accurately. But the performance on Bickley diary dataset and some images of DIBCO contests still needs to be improved.

7 CONCLUSIONS

This paper presents an adaptive image contrast based document image binarization technique that is tolerant to variety of document degradation such as uneven illumination and document smear. The proposed technique is easy & robust, only few parameters are involved. It works for different kinds of degraded document images. It makes use of the local image contrast that is evaluated based on the local maximum and minimum. The proposed method has been tested on the various datasets. Experiments show that the proposed method outperforms most reported document binarization methods in term of the F-measure, pseudo F-measure, PSNR, NRM, MPM and DRD.

8 REFERENCES

[1]. B. Gatos, K. Ntirogiannis, and I. Pratikakis, "ICDAR 2009 document image binarization contest (DIBCO 2009)," in Proc. Int. Conf. Document Anal. Recognit., Jul. 2009, pp. 1375–1382.

[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]. I. Pratikakis, B. Gatos, and K. Ntirogiannis, "H-DIBCO 2010 handwritten document image binarization competition," in Proc. Int. Conf. Frontiers Handwrit. Recognit., Nov. 2010, pp. 727–732.

[4]. 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.

[5]. 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.

[6]. 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,” in Proc. Int. Conf. Document Anal. Recognit., vol. 13. 2003, pp. 859–864.

[7]. M. Sezgin and B. Sankur, “Survey over image thresholding techniques and quantitative performance evaluation,” J. Electron. Imag., vol. 13, no. 1, pp. 146–165, Jan. 2004.

[8]. O. D. Trier and A. K. Jain, “Goal-directed evaluation of binarization methods,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 17, no. 12, pp. 1191–1201, Dec. 1995.

[9]. O. D. Trier and T. Taxt, “Evaluation of binarization methods for document images,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 17, no. 3, pp. 312–315, Mar. 1995.

[10]. A. Brink, “Thresholding of digital images using two-dimensional Entropies,” Pattern Recognit., vol. 25, no. 8, pp. 803–808, 1992.

[11]. J. Kittler and J. Illingworth, “On threshold selection using clustering criteria,” IEEE Trans. Syst., Man, Cybern., vol. 15, no. 5,pp. 652–655, Sep.–Oct. 1985.

[12]. N. Otsu, “A threshold selection method from gray level histogram,” IEEE Trans. Syst., Man, Cybern., vol. 19, no. 1, pp. 62–66, Jan. 1979.

[13]. N. Papamarkos and B. Gatos, “A new approach for multithreshold selection,” Comput. Vis. Graph. Image Process., vol. 56, no. 5, pp. 357–370, 1994.

[14]. J. Bernsen, “Dynamic thresholding of gray-level images,” in Proc. Int. Conf. Pattern Recognit., Oct. 1986, pp. 1251–1255.

[15]. L. Eikvil, T. Taxt, and K. Moen, “A fast adaptive method for binarization of document images,” in Proc. Int. Conf. Document Anal. Recognit., Sep. 1991, pp. 435–443.

[16]. I.-K. Kim, D.-W. Jung and R.-H. Park, “Document image binarization based on topographic analysis using a water flow model,” Pattern Recognit., vol. 35, no. 1, pp. 265–277, 2002.

[17]. J. Parker, C. Jennings, and A. Salkauskas, “Thresholding using an illumination model,” in Proc. Int. Conf. Doc. Anal. Recognit., Oct. 1993, pp. 270–273.

[18]. J. Sauvola and M. Pietikainen, “Adaptive document image binarization,” Pattern Recognit., vol. 33, no. 2, pp. 225–236, 2000. SU et al.:

[19]. W. Niblack, an Introduction to Digital Image Processing. Englewood Cliffs, NJ: Prentice-Hall, 1986.

[20]. J.-D. Yang, Y.-S. Chen, and W.-H. Hsu, “Adaptive thresholding algorithm and its hardware implementation,” Pattern Recognit. Lett., vol. 15,no. 2, pp. 141–150, 1994.