THRESHOLD CORRECTION OF DOCUMENT IMAGE BINARIZATION FOR TEXT EXTRACTION

Hiroshi Tanaka, Yusaku Fujii and Yoshinobu Hotta
Fujitsu, 4-1-1 Kamikodanaka, Nakahara-ku, Kawasaki, Kanagawa 211-8588, Japan

Keywords: Adaptive binarization, Text extraction, Thresholding, Otsu binarization, Threshold correction, Background noise, Niblack, Image resolution.

Abstract: In this paper, a simple threshold correction method for document image binarization for text extraction is presented. This method enhances the binary image of characters, which is often adversely influenced by neighboring strong pixels or background noise. The threshold correction method is based on a similar method applied to ruled-line extraction presented by the author, and is claimed to be effective to text extraction. The author also reveals the relationship between effectiveness of the method and the image resolution.

1 INTRODUCTION

One of the most important objectives of document image binarization is to extract text images from the document background. In a simple document model, each object is considered to be placed on the flat surface of the document background. According to this model, binarization can be considered as a two-class discrimination problem for determining a global threshold (Otsu, 1979). However, complicated document images require adaptive binarization, in which the local threshold is calculated for each pixel. Such images have complex designs, which cannot be expressed using two classes; further, they could be severely degraded.

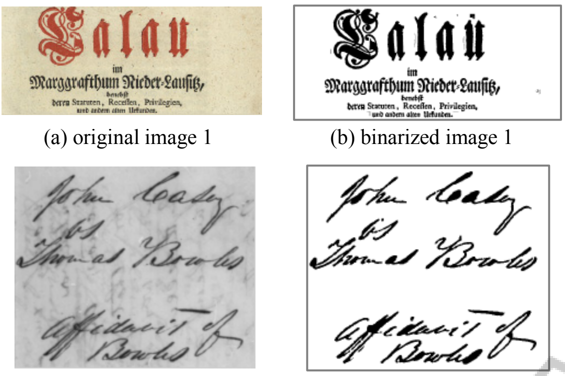
In the past, various adaptive binarization methods have been proposed. Trier (Trier and Jain, 1995) compared several binarization methods on the bases of thier character recognition accuracies and concluded that Niblack's method (Niblack, 1986) yields the best result when the noise reduction technique is applied. Sauvola (Sauvola et. al., 1997) modified Niblack's method using region analysis, in which textual and nontextual regions were separated from each other. Sauvola's method has been the most popular binarization method for document images. These methods assume that pixels can be classified into two classes among local neighbors.

In the recent years, we can also find a lot of newly invented binarization methods that may overcome some problems of conventional methods. For example, DIBCO 2009, the Document Image Binarization Contest held in ICDAR 2009, is a good collection of the latest document binarization methods (Gatos et. al., 2009). Although there are great methods proposed in DIBCO 2009, most of them focus on binarizing much degraded images such as historical documents depending on the image quality used in the contest (Fig. 1), and then they require much computing cost.

Our document recognition system recognizes binarized text images obtained by an adaptive binarization method based on Niblack's method (Kamada and Fujimoto, 1999). As described later, adaptive binarization methods, including those developed by Niblack and Sauvola, have a problem. Because these methods are based on the assumption that local neighbors can be classified into two classes, some pixels that have three or more pixel classes in each local area are often dropped off. This results in broken shapes of character images (Fig. 2) and causes errors in character recognition. We solve this problem by correcting the binarization threshold with respect to the neighboring threshold surface. This technique was once applied to ruled-line extraction (Tanaka, 2009) and is also proved to be effective to text extraction.

In Section 2, we describe the problems of conventional methods and our solutions. In Section 3, we present experimental results. Finally in Section 4, we conclude the paper.

Tanaka H., Fujii Y. and Hotta Y..
THRESHOLD CORRECTION OF DOCUMENT IMAGE BINARIZATION FOR TEXT EXTRACTION.
DOI: 10.5220/0003396503870391
In *Proceedings of the International Conference on Computer Vision Theory and Applications* (VISAPP-2011), pages 387-391
ISBN: 978-989-8425-47-8



(a) original image 1

(b) binarized image 1

(c) original image 2

(d) binarized image 2

Figure 1: Sample images used in DIBCO 2009. (copied from Gatos et al., 2009)

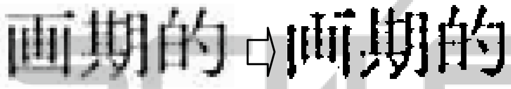


Figure 2: Broken text image.

2 BINARIZATION ALGORITHM

In our algorithm (Fig. 3), initial thresholds are calculated for each pixel, and the pixels area initially classified as foreground or background pixels. The initial classification is used in the background determination process to determine whether each pixel can be a foreground pixel. For foreground pixels, binarization thresholds are obtained by correcting initial thresholds.

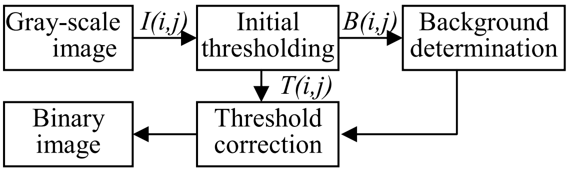


Figure 3: Binarization algorithm.

In the following subsections, the basic concept of the adaptive binarization method, which is used to calculate initial thresholds, is described. Background noise reduction, which is a modification of Eikvil's method (Eikvil, 1991), will be described. Finally, the method for the correction of binarization thresholds is presented.

2.1 Initial Thresholding

The threshold values for each pixel are calculated in the initial thresholding step using Niblack's method and the threshold surface is formed by the threshold values. The threshold for a pixel is obtained using the neighboring pixel values *I (i, j)*, reverse values of brightness, in the $w \times w$ local areas. The initial threshold T is calculated using the following formulas:

$$m = \sum_i^w \sum_j^w I(i,j) / w^2 \qquad (1)$$

$$\delta^2 = \sum_i^w \sum_j^w \{I(i,j) - m\}^2 / w^2 \qquad (2)$$

$$T = m + k\delta \qquad (3)$$

A binarized image is obtained by comparing each pixel value with its associated threshold (Fig. 4). The value $k\delta$ in formula 3 is used to shift the threshold value so as to reduce background noise (Fig.5).

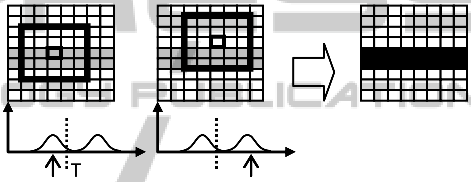
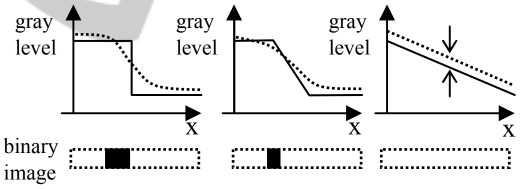


Figure 4: Thresholding by Niblack's method.



(a) step boundary (b) slope boundary (c) flat slope

Figure 5: Binarization example.

2.2 Background Noise Reduction

When a background surface has small fluctuations, many tiny noises will be extracted from the background because the value $k\delta$ is so small that even minute noises cannot be suppressed (Fig. 6). To reduce these noises, the value of background flatness is calculated, and noises in the flat region are eliminated.

We calculate the flatness value on the bases of Eikvil's method, which is described as follows. In the neighboring $w \times w$ local areas, each pixel can be classified according to the initial threshold (formula 3), and the mean values of each class, i.e., μ_1 and μ_2, can be obtained. The flatness value F of the central pixel of the local area is defined as

$F = |\mu_1 - \mu_2| - d$, where d is a predefined parameter, and the pixel is considered to be flat when F is negative. For reference, Bernsen (Bernsen, 1986) has defined F as follows:

$$F = \max_{i,j}^{w}\{I(i,j)\} - \min_{i,j}^{w}\{I(i,j)\} - d \qquad (4)$$

And Sauvola has proposed a flatness measure called "transient difference" to distinguish between flat and nonflat regions.

Although Eikvil's method is effective for most images, it encounters problems when binarizing low contrast images. Fig. 7 is an example of a test in which we applied Eikvil's method to extract ruled-lines. As shown in Fig. 7, some of the ruled-line pixels are dropped with the normal d parameter (e.g., 10), but there should be a lot of noise pixels with the lower parameter d (e.g., 2). We had proposed a new method using adaptive parameter to improve ruled-line binary image (Tanaka, 2009), in which the parameter d is set to lower value only when the mask shown in Fig. 8 detects straightness feature.

Although the adaptive parameter improves the ruled-line images, it is not always effective for text images because the straightness feature is not essential for text images. Then we adopt another noise remove method which removes small CCAs as noise pixels (Sauvola et.al., 1997). It reduces some of the remaining noise pixels as shown in Fig. 9.

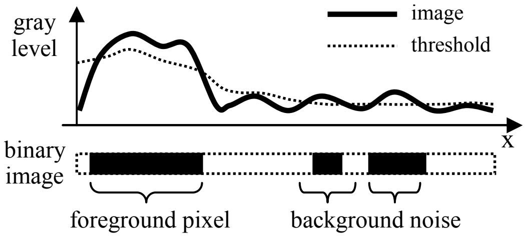
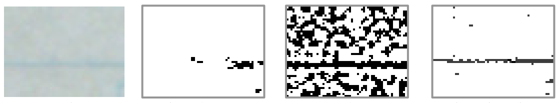


Figure 6: Background noise.



(a) gray image (b) d=10 (c) d=2 (d) adaptive

Figure 7: Background decision adaptation.

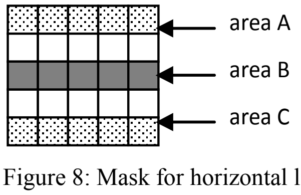
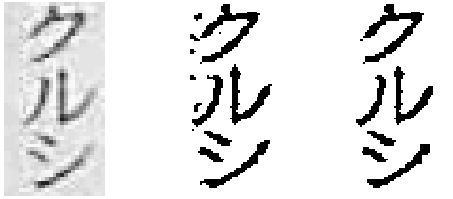


Figure 8: Mask for horizontal line.



(a) gray image (b) lower 'd' value (c) remove noise

Figure 9: Small CCA removal.

2.3 Threshold Correction

As adaptive binarization methods, such as those developed by Niblack and Sauvola, assume that a local area contains pixels of two classes, local areas containing more than two classes of pixels often encounter classification failures. Fig. 10 shows the effect of neighboring pixels on the binarization threshold. When a local area contains a third class of pixels (black pixels in the figure), the threshold shifts because of the change in the mean value m (see formula 3). This causes lacking pixels in the text images in Fig. 2.

The effect of the third class of pixels can be prevented if the local area is shifted away from this class as shown in Fig. 11. This shift is equivalent to using the binarization threshold of the shifted position at a shift width S from the current pixel. We define the threshold conversion using a shift operation as follows.

$$T'(i,j) = \min_{k,l=-S}^{S}\{T(i+k, j+l)\} \qquad (5)$$

Fig. 12 explains the detail and a modification of threshold correction. Fig. 12(a) is an ideal binary image which has two classes (black and white) of pixels. By scanning, it may be converted to an image which has three or more pixel values such as Fig. 12(b) because of quantization errors. Fig.12 (b) also expresses the neighboring area in which threshold values are used to correct the threshold of current position (Formula 5). As we think that the pixels in the correction area should have similar pixel values, the correction area is narrowed as shown in Fig. 12(c). Using this modification, threshold correction works like Fig. 12(d),(e).

By the modification, formula 5 is also modified as follows, where e is a predefined parameter and we set a value 20 for e in the experiment.

$$\begin{gathered} T'(i,j) = \min_{k,l=-S}^{S}\{T(i+k, j+l)\} \\ when\ |I(i+k, j+l) - I(i,j)| < e \end{gathered} \qquad (6)$$

By carrying out threshold conversion using formula 6, the result of binarization of the image shown in Fig. 13(a) changes from that shown in Fig. 13(b) to Fig. 13(c). Fig. 14 shows another example. Please note that *S* must be smaller than *w/2*, or the current pixel should be outside the shifted local window.

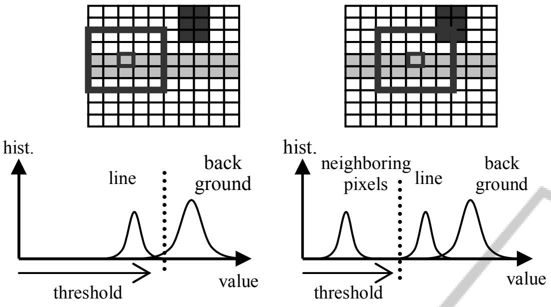


Figure 10: Affection of neighboring pixels.

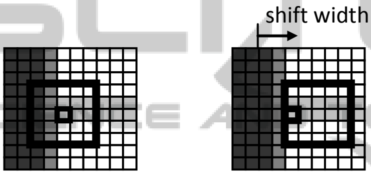
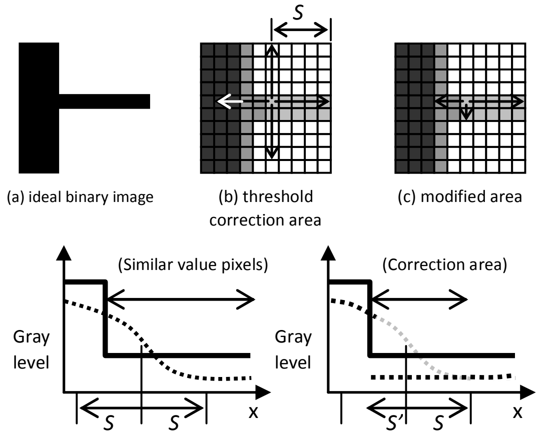


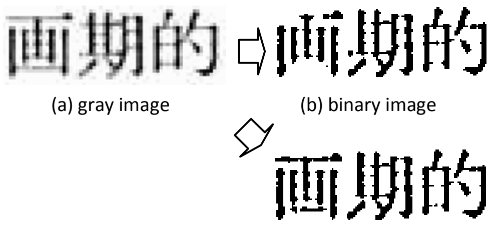
Figure 11: Shifted local window.



(a) ideal binary image (b) threshold correction area (c) modified area

(d) threshold correction (e) modified area

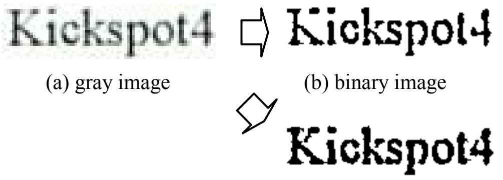
Figure 12: Modification of correction area.



(a) gray image (b) binary image

(c) binary image (improved)

Figure 13: Improved binary image.



(a) gray image (b) binary image

(c) binary image (improved)

Figure 14: Improved binary image (another example).

3 EVALUATION

We tested our method using a document image set consists of 17 kinds of document images with 4 resolutions (150, 200, 300 and 600 dpi); 68 images in total. Table 1 shows the result of character recognition with recall and precision rate. Although the absolute values do not imply anything because they depend on the test images, the results of the developed method are relatively more accurate than those of the conventional method. It is observed that almost one-third of the recognition errors are corrected using the proposed method.

In addition, the proposed method can be seen effective especially for lower resolution images. It is because the dropped pixels (Fig. 2) are mainly low contrast part of the text pixels, and are often found in low resolution images caused by quantization errors.

Table 1: Character recognition rates.

	conventional method		proposed method	
	recall	precision	recall	precision
150 dpi	94.91%	93.75%	97.21%	97.21%
200 dpi	97.05%	95.41%	98.37%	98.12%
300 dpi	97.94%	97.11%	98.01%	97.17%
600 dpi	97.98%	97.57%	98.23%	97.66%
average	96.97%	95.71%	97.96%	97.54%

4 CONCLUSIONS

In this paper, we have described a threshold correction method that can be used to improve the quality of text image binarization and lead to higher accuracy of character recognition.

This method was originally developed to enhance ruled-line extraction (Tanaka 2009), and is revealed to be effective for text extraction. In the case of ruled-lines, we used the straight-line feature of ruled-lines to distinguish line pixels from background noise. On the other hand, we use another noise reduction method to suppress noises for text binarization.

According to the experimental result, we have found that the proposed method is especially effective for lower resolution images. The reason of it can be explained that the pixel dropping problem is often caused by quantization errors which is mainly seen in lower resolution images. In addition, it can be considered that the same errors may occur even in high resolution images when character size is very small.

As the next step, we will investigate the relationship between the effectiveness of our method and the character sizes in detail. It is because we expect our method to improve the recognition accuracy for higher resolution images.

REFERENCES

Otsu, N., 1979. A Threshold Selection Method from Gray-level Histograms. In IEEE Trans. Systems, Man, and Cybernetics, vol.9, no.1, 1979, pp. 62-66.

Trier, O. D., and Jain, A. K., Goal-directed Evaluation of Binarization Methods. In *IEEE Trans. PAMI.* vol.17, no.12, Dec. 1995, pp.1191-1201.

Niblack, W., An Introduction to Digital Image Processing. *In Prentice Hall,* Englewood Cliffs, N. J., 1986, pp.115-116.

Sauvola, J., Seppanen, T., Haapakoski, S., and Pietikainen, M., Adaptive Document Binarization. *In Proc. 4th. ICDAR*, Ulm, Germany, Aug. 1997, pp.147-152.

Gatos, B., Ntirogiannis, K., and Pratikakis, I., ICDAR 2009 Document Image Binarization Contest (DIBCO 2009). *In Proc. 10th. ICDAR*, Barcelona, Spain, Jul. 2009, pp.1375-1382.

Kamada, H., and Fujimoto, K., High-speed, High-accuracy Binarization Method for Recognizing Text in Images of Low Spatial Resolutions. *In Proc. 5th. ICDAR*, Bangalore, India, Sep. 1999, pp.139-142.

Tanaka, H., Threshold Correction of Document Image Binarization for Ruled-line Extraction. *In Proc. 10th. ICDAR*, Barcelona, Spain, Jul. 2009, pp.541-545.

Eikvil, L., Taxt, T., and Moen, K., A Fast Adaptive Method for Binarization of Document Images. *In Proc. 1st. ICDAR*, Saint-Malo, France, Sep. 1991, pp.435-443.

Bernsen, J., Dynamic Thresholding of Gray-level Images. *In Proc. 8th. ICPR*, Paris, France, Oct. 1986, pp.1251-1255.