Binarization of Badly Illuminated Document Images through Shading Estimation and Compensation

Shijian Lu, Chew Lim Tan
Department of Computer Science, School of Computing
National University of Singapore, Kent Ridge, 117543, Singapore
{lusj, tancl@comp.nus.edu.sg}

Abstract

This paper presents a document image binarization technique that segments text from badly illuminated document images. Based on the observations that text documents normally lie over a planar or smoothly curved surface and have a uniformly colored background, badly illuminated document images are binarized by using a smoothing polynomial surface, which estimates the shading variation and compensates the shading degradation based on the estimated shading variation. Badly illuminated document images are accordingly binarized through the global thresholding of the compensated document images. Compared with the reported methods, the proposed technique is tolerant to the variations in text size and document contrast. At the same time, it is much faster and able to produce a binary text image with little background noise.

1. Introduction

Document image binarization aims to segment a document image into two categories, namely, the blank background and the foreground text. Though it is often implemented in the preprocessing stage, its performance may be crucial to the success of the ensuing tasks such as document layout analysis and optical character recognition (OCR). As an increasing number of documents are being and will be digitalized, a fast and efficient document binarization technique is required to facilitate the management of the digitalized document images fast and efficiently.

Document images often suffer from various types of shading degradation. For example, as the thick bound volumes cannot be flattened physically, shading nearly always exists throughout the spine regions of scanned document shown in Figure 1(a). In addition, for documents captured by a digital camera, the shading degradation illustrated in Figure 1(c) becomes even more common because camera

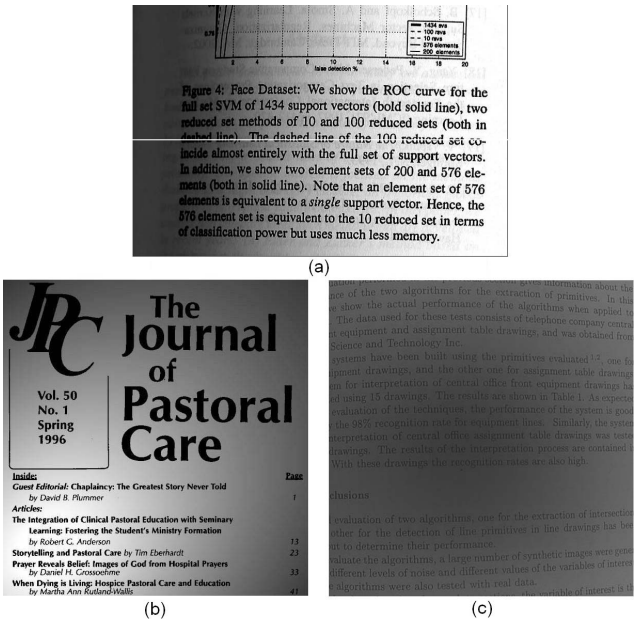


Figure 1. Three badly illuminated documents.

documents are more susceptible to the lighting variation. For badly illuminated documents, adaptive thresholding, which calculates a local threshold for each image pixel, is the traditional approach for the document binarization.

A large number of window-based adaptive thresholding techniques have been reported [8] in the literature. As a typical representative, Niblack's method [4] estimates the local threshold by using the local mean *m* and the standard variation *s*, computed within a small neighborhood window of each pixel as follows:

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

where *k* is a user defined parameter and it normally lies between -1 and 0.

As evaluated in [8], Niblack's method normally outperforms others in term of the binarization speed and the binarization efficiency. However, similar to other window based thresholding techniques, his method has a few limitations including the low thresholding speed, the sensitivity to the window size [1], and the presence of a large amount of noise in the background areas. To suppress the background noise, Sauvola *et al.* [6] later modify the formula in Equation (1) and propose a new thresholding formula as follows:

$$T = m \cdot (1 + k \cdot (\frac{s}{R} - 1)) \tag{2}$$

where parameter *R* refers to the dynamic range of the standard deviation and *k* instead takes a positive value between 0 and 1. The new thresholding formula reduces the background noise greatly, but it requires the knowledge of document contrast to set the parameter *R* properly.

In [3], a local polynomial surface is used to binarize fingerprints by fitting the polynomial surface to pixels within a small sliding window. In [7], Seeger et al. propose to binarize camera documents by interpolating the background surface. In this paper, we globalize the local polynomial surface and binarize badly illuminated document images by fitting a polynomial surface to pixels within a whole document image. We utilize polynomial surfaces for document binarization based on two observations. Firstly, physical documents in real scene normally lie over a planar or smoothly curved surface. Therefore, the illumination normally varies smoothly along the document surface. Secondly, text documents generally have a uniformly colored (typical in white) background such as blank document margins. As a result, the luminance variation in these blank regions can be assumed to be caused by the illumination variation.

In the proposed technique, badly illuminated document images are binarized through the estimation and compensation of the shading degradation. In particular, two round of smoothing are implemented where the first aims to detect the blank background roughly and the second further estimated the shading variation. The process finally produces a roughly uniformly illuminated document image that can be binarized by using some global thresholding technique. Compared with the reported methods, the proposed technique has a few advantages. Firstly, it is fast. Though it is a little slower than global thresholding techniques, it is much faster than most adaptive thresholding technique [4, 6]. Secondly, it is robust and tolerant to the variations in text size and document contrast. Thirdly, it is efficient and able to produce a binary image with little background noise.

2. Proposed Method

This section presents the proposed document binarization technique. In particular, we will divide this section into three subsections, which deal with the description of the polynomial surface smoothing, the estimation of the document shading degradation, and the binarization of badly illuminated document images, respectively.

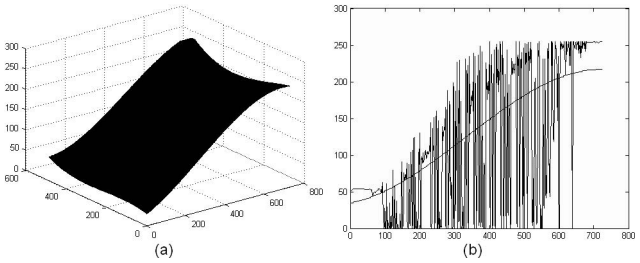


Figure 2. (a) The first round cubic polynomial surface PS_f fitted to the documents in Figure 1(a); (b) the pixel intensity and the PS_f along the horizontal scan line in Figure 1(a).

2.1 Polynomial Surface Smoothing

Smoothing is a process by which signals are averaged over their neighbors. For a series of equally spaced signals $[y_{x_1}\ y_{x_2} \cdots y_{x_n}]$, the smoothing normally replaces each signal $y_{x_k}, k = 1 \cdots n$, by y'_{x_k}, a linear combination of the y_{x_k} and the signals within a small neighborhood window. Take the mean filter as an example. It computes the y'_{x_k} as the average of signals within a neighborhood window.

Similar to the mean filter, the polynomial smoothing (also named Savitzky-Golay filter [2]) aims to approximate the underlying function. The fundamental idea is to fit a least square polynomial or polynomial surface to the signals within a neighborhood window. The smoothed signal y'_{x_k} is accordingly estimated as the value of the fitted polynomial or polynomial surface at the point x_k. The equation below gives a polynomial surface of degree d:

$$f(x, y) = \sum_{i+j=0}^{d} a_{i,j} x^i y^j \tag{3}$$

where x, y refers to the coordinate of the image pixel at $G(x, y)$. $A, a_{i,j}, i + j = 0 \cdots d$ gives the coefficients of the polynomial surface, which can be estimated as follows:

$$A = (X^T \cdot X)^{-1} \cdot X^T \cdot I \tag{4}$$

where I gives the intensity of image pixels and the matrix X is constructed as follows:

$$X = \begin{vmatrix} 1 & x_0 & y_0 & x_0^2 & x_0 y_0 & \cdots & y_0^3 \\ 1 & x_1 & y_1 & x_1^2 & x_1 y_1 & \cdots & y_1^3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & y_n & x_n^2 & x_n y_n & \cdots & y_n^3 \end{vmatrix}$$

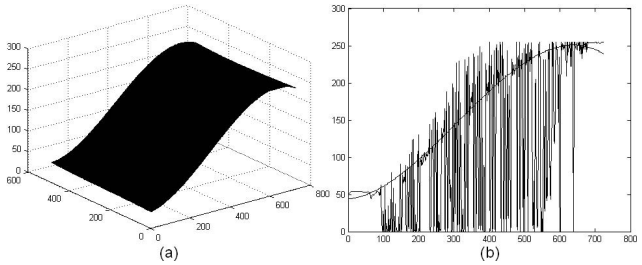


Figure 3. (a) The second round polynomial surface PS_s of order three fitted to the background pixels of the documents in Figure 1(a); (b) the pixel intensity and the PS_s along the scan line in Figure 1(a).

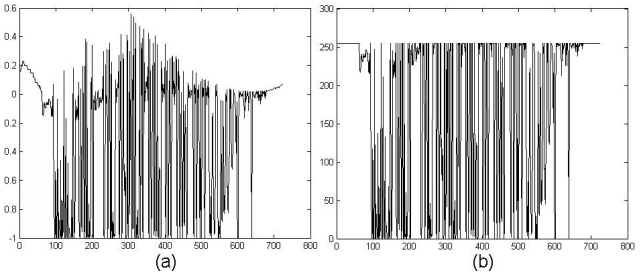


Figure 4. (a) Normalized document pixels along the horizontal scan line in Figure 1(a); (b) compensated pixel intensity along the horizontal scan line in Figure 1(a).

where n refers to the number of image pixels within the studied document image.

The polynomial surface smoothing is originally a local operator, which fits a polynomial or polynomial surface to the data within a small neighborhood window. In this paper, we globalize the local polynomial smoothing operator and fit a polynomial surface to the intensity of pixels within a whole document image. We aim to estimate the shading variation, which will then be utilized for shading compensation and global thresholding of the compensated document images as detailed in the following sections.

2.2 Shading Estimation

In the proposed method, we first estimate the shading variation by two rounds of polynomial surface smoothing process. In particular, the first round aims to detect the blank background by fitting a polynomial surface PS_f to the intensity of pixels within the whole document image. The second round then estimates the shading variation by fitting a polynomial surface PS_s to the intensity of pixels that have been classified to the background.

For the badly illuminated document image in Figure 1(a), Figure 2(a) shows the first round polynomial surface PS_f that is fitted by using all document pixels. To visualize the estimation process, we scan the document in Figure 1(a) horizontally in the middle position (highlighted by a white line in Figure 1(a)). Figure 2(b) shows the intensity of pixels along that horizontal scan line and the fitted smoothing surface PS_f along that scan lines. As Figure 2(b) shows, the polynomial surface PS_f traces the shading variation properly. However, as inked document pixels (typical black) "pull" the PS_f down, PS_f is generally lower than the background luminance shown in Figure 2(b).

After the fitting of PS_f, we classify document pixels into two categories, namely, an inked category that contains the pixels of text and other dark document components and a background category containing document pixels within the blank regions. In particular, we classify the pixels with intensity far below the PS_f to the inked category and the remaining to the background category as follows:

$$INK = \{G(x,y) : PS_f(x,y) - I(x,y) > K\} \quad (5)$$

where $I(x,y)$ refers to the intensity of document pixel at $G(x,y)$ and $PS_f(x,y)$ refers to the value of the first round polynomial surface at (x, y). The threshold K is roughly estimated as the mean of the difference between PS_f and the intensity of document pixels whose intensity is smaller than PS_f as follows:

$$K = \frac{1}{N}\sum PS_f(x,y) - I(x,y)$$
$$\forall G(x,y) : I(x,y) - PS_f(x,y) < 0 \quad (6)$$

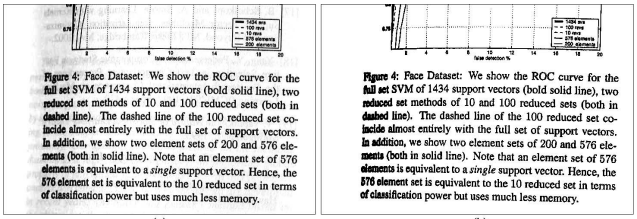
where N gives the number of document pixels whose intensity is smaller than the value of PS_f.

The second round polynomial surface is further fitted by using document pixels that are classified to the background category. For the badly illuminated document in Figure 1(a), Figure 3(a) shows the second round polynomial surface PS_s. Figure 3(b) further shows the values of the smoothing surface PS_s along the horizontal scan line highlighted in Figure 1(a). As Figure 3(b) shows, the PS_s traces the shading variation more accurately.

2.3 Document Binarization

Based on the shading variation estimated in the last subsection, badly illuminated document images can be compensated through a normalized process as follows:

$$N(x,y) = \begin{cases} \frac{I(x,y) - PS_s(x,y)}{PS_s(x,y)} & \text{if } PS_s(x,y) > 0 \\ -1 & \text{Otherwise} \end{cases} \quad (7)$$



(a) (b)

Figure 5. (a) Shading compensation of the document image in Figure 1(a); (b) global thresholding of document in Figure 5(a).

where $I(x, y)$ refers to the intensity of the pixel at $G(x, y)$ and $PS_s(x, y)$ denotes the value of the polynomial surface PS_s at the point (x, y). It should be noted that $I(x, y)$ normally goes close to zero when $PS_s(x, y)$ goes below zero. Under such cirsumstance, we assign $N(x, y)$ to -1, which will then be transformed to 0 as described below.

For document pixels within the background category, the normalized pixel intensity is normally close to zero because $I(x, y)$ is close to $PS_s(x, y)$. But for document pixels within the inked category, $N(x, y)$ goes close to -1 because the intensity of inked pixels is normally close to 0. The compensated document can therefore be determined as:

$$G(x, y) = 255 \cdot \big(1 + N(x, y)\big) \tag{8}$$

As $N(x, y)$ in Equation (7) normally lies around $[-1\ 0]$, the transformation in Equation (8) above converts the intensity from $[-1\ 0]$ to [0 255] roughly.

For the horizontal scan line in Figure 1(a), Figure 4(a) shows the normalized pixel values $N(x, y)$ in Equation (7) and Figure 4(b) shows the compensated pixel intensity. As Figure 4 shows, the shading variation has been roughly compensated. Figure 5(a) then shows the compensated document image. Finally, the badly illuminated document image in Figure 1(a) can be binarized through the global thresholding of the compensated document in Figure 5(a). Figure 5(b) shows the resultant binary document image.

3. Experiments and Discussions

The proposed method has been tested by using 30 badly illuminated document images. Besides the shading degradation, some test documents are also degraded by the variations in the document contrast and character size illustrated in Figures 1(b) and 1(c). In the experiments, we compare our method with three well-known ones, namely, Otsu's method [5], Sauvola's method [6], and Niblack's method [4]. In particular, we set the window size of Niblacks and Sauvola's methods at 20×20. The parameter k in Equations (1) and (2) is set to -0.2 and 0.5. The parameter R in

(a) (b) (c) (d)

Figure 6. Binarization results of the document image in Figure 1(b) by using Otsu's in (a), Niblack's in (b), Sauvola's in (c), and our proposed document binarization method in (d).

Equation (2) is set at 128 as recommended in [6]. The binarization results in Figures 6(d) and 7(d) show the superior performance of the proposed thresholding technique.

The speed of the proposed thresholding method is first tested. Experiments show that though the proposed technique is a little slower than Otsu's method, it is much faster (around 10 times) than Niblack's and Sauvola's window based thresholding techniques. The higher speed can be explained by the polynomial surface smoothing process, which fits the polynomial surface to document pixels and accordingly avoids the calculation of the intensity mean and variance for each document pixel. The speed of the proposed technique can be further improved through the data reduction, which fit the polynomial surface to the regularly sampled instead of all document pixels.

Besides, the proposed technique is robust and tolerant to the variation in text size and document contrast. As a typical limitation, window-based adaptive thresholding techniques require a manual tuning process to get an optimal window size. In particular, if the window is too big, the computational load increases dramatically. Otherwise, the interior of characters of large size shown in Figures 6(b) and 6(c)

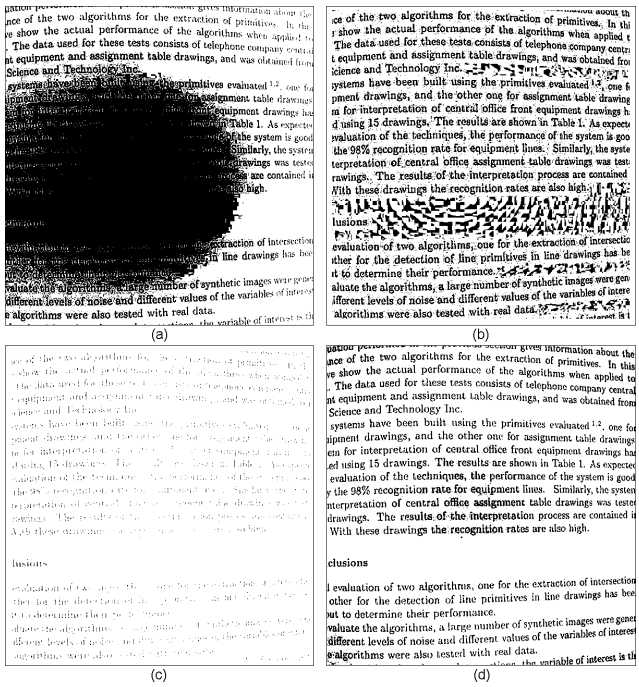


Figure 7. Binarization results of the document image in Figure 1(c) by using Otsu's in (a), Niblack's in (b), Sauvola's in (c), and our proposed document binarization method in (d).

may be classified into the background incorrectly. On the other hand, some window-based method such as Sauvola's requires the knowledge the document contrast to set the parameter R properly. Otherwise, text may also be classified to the background shown in Figure 7(c).

Furthermore, the proposed method is more efficient than most reported ones. Experiments show that the character segmentation rate of the proposed method reaches up to 91.33%, which is much higher than Otsu's and Sauvola's (61.58% and 78.34%). The low segmentation rate of the Otsu's method can be explained by the fact that it classifies a large amount of shaded text into the background (shown in Figures 6(a) and 7(a)). Sauvola's segmentation rate is low because text within some document of low contrast is also classified into the background (shown Figure 7(c)). Niblack's method can achieve a 90.29% segmentation rate, but it produces a large amount of background noise (shown in Figures 6(b) and 7(b)), which will impede the ensuing document processing tasks such as OCR.

Though the proposed technique is fast, robust, and efficient, a few limitations exist. Firstly, we set the order of the polynomial surface at 3 and experiments show that such polynomial surface can estimate most smooth shading variation properly. However, for documents suffering from more complex shading variation, the polynomial surface of order 3 may not model the shading variation correctly. Secondly, for images with a large amount of uniformly colored background such as text documents, maps, and engineering drawings, the proposed two rounds of smoothing process is capable of estimation the shading variation properly in most cases. But for images with some non-text components such as textured graphics, more than two rounds of smoothing process may be required to estimate the shading variation properly. We will study these two issues in our future work.

4. Conclusion

This paper presents a document binarization technique that segments text from badly illuminated document images fast and efficiently. A smoothing polynomial surface is utilized for the shading estimation and compensation, which produce a roughly uniformly illuminated document image that can be binarized by some global thresholding techniques. Experiments show that the proposed technique is fast, tolerant to the variations in text size and document contrast, and able to binarize badly illuminated document image with little background noise.

5 Acknowledgement

This research is supported by the Agency for Science, Technology and Research (A*STAR), Singapore, under grant no. 0421010085.

References

[1] M. L. Feng and Y. P. Tan. Adaptive binarization method for document image analysis. *IEEE International Conference on Multimedia and Expo*, 1:339– 342, June 2004.

[2] R. W. Hamming. *Digital Filter*. Prentice-Hall, Englewood Cliffs, New Jersey, 1983.

[3] M. Krzysztof, M. Preda, and M. Axel. Dynamic threshold using polynomial surface regression with application to the binarization of fingerprints. *Proceedings of the SPIE*, 5779:94–104, 2005.

[4] W. Niblack. *An Introduction to Digital Image Processing*. Prentice-Hall, Englewood Cliffs, New Jersey, 1986.

[5] N. Otsu. A threshold selection method from graylevel histogram. *IEEE Transactions on System, Man, Cybernetics*, 19(1):62–66, January 1978.

[6] J. Sauvola and M. Pietikainen. Adaptive document image binarization. *Pattern Recognition*, 33(2):225–236, January 2000.

[7] M. Seeger and C. Dance. Binarising camera images for ocr. *Proceedings of ICDAR*, pages 54–58, 2001.

[8] O. D. Trier and T. Taxt. Evaluation of binarization methods for document images. *IEEE Transaction on Pattern Analysis and MachineIntelligence*, 17(3):312–315, March 1995.