2009 10th International Conference on Document Analysis and Recognition

Feature Based Binarization of Document Images Degraded by Uneven Light Condition

Jung Gap Kuk
Seoul National University
School of Electrical Engineering
Seoul 151-744, Korea
Email : jg-kuk@ispl.snu.ac.kr

Nam Ik Cho
Seoul National University
School of Electrical Engineering
Seoul 151-744, Korea
Email : nicho@snu.ac.kr

Abstract

This paper proposes a document image binarization method, which is especially robust to the images degraded by uneven light condition, such as the camera captured document images. A descriptor that captures the regional properties around a given pixel is first defined for this purpose. For each pixel, the descriptor is defined as a vector composed of filter responses with varying length. This descriptor is shown to give highly discriminating pattern with respect to the background region, text region, and near text region. Of course there are misclassified pixels, which are then relabeled using an energy optimization methodbased on graph cut method. For this, we devise an appropriate energy function that leads to clear and correct binarization. The proposed descriptor is also used for the skew detection, and thus correcting the skewed documents.

1. Introduction

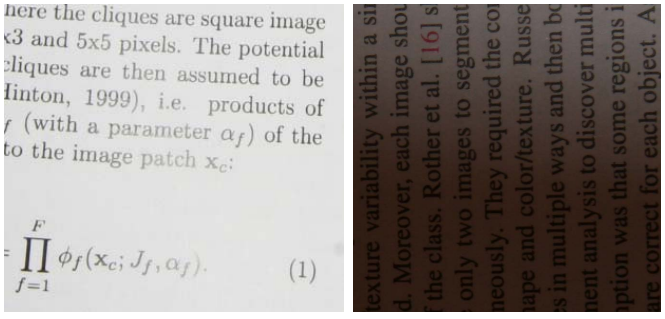
The resolution of very cheap digital camera nowadays is very high enough to be used as a capturing device for documents instead of the flatbed scanners. However, one of the most serious problems when capturing a document image by a digital camera is the uncontrolled light condition. While the flatbed scanner scans a document with the material contacted tightly on a bed with well controlled lights, a document image captured by a digital camera suffers from unevenness of light and geometric distortion. When binarizing a document image for optical character recognition (OCR) or other purposes, the most serious problem would be the uneven brightness as shown in the example of Fig. 1. It can be seen that text gray levels are brighter around the word "patch" at the center in Fig. 1 (a) and shadow gets gradually darker as it goes to the bottom in Fig. 1 (b). The degradation is inevitable and the image processing of this document image produces poor results. Hence we need to take into account the light condition when binarizing document images, specially the ones captured by digital camera.

There have been many adaptive algorithms for the binarization of a document image. To the best of our knowledge all the algorithms are basically the thresholding algorithms that depend on the choice of thresholds :

$$f_p = \begin{cases} 1 & \text{if } \psi_p > T_p \\ 0 & \text{otherwise,} \end{cases}$$

where f_p is a binary random variable which specifies an assignment of *foreground* or *background* to a pixel p, ψ is an information extracted from an observed image y, and T is a threshold surface. It is noted that *foreground* denotes text region and *background* denotes the region except for the text region.

Most of binarization algorithms define ψ as a pixel intensity [4, 5, 6, 2]. In this literature, how to determine a threshold surface is a key issue since the algorithm should be able to adaptively handle the local variability. The well-known



(a) (b)

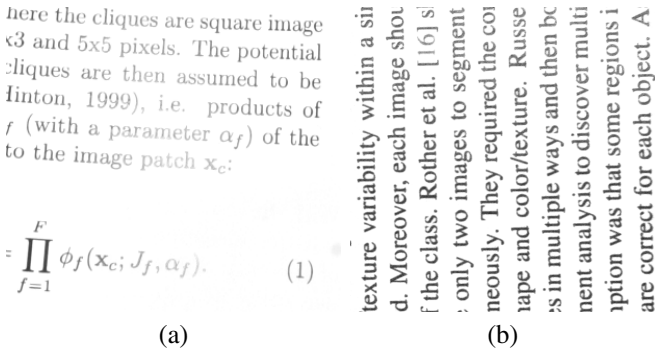
Figure 1. Observed images with uneven light distribution

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.157

Niblack algorithm adjusts the threshold level according to the statistics of pixels in a local patch [5]. The Palumbo algorithm in [6] extends the support region of statistics to the neighboring patches to have an adaptive threshold surface. Recently, a method combining these two algorithms has also been proposed [4]. Also, a quantile linear algorithm based on the Niblack algorithm is proposed in [3], which is the collection of the gradients of all edge pairs in multiscale Laplacian domain. The global threshold is determined from this information, which best discriminates character edges from the other kind of edges. In retinex based algorithm [7] , the *lightness* image is defined as an information function and the global threshold is set experimentally.

Even though these algorithms give satisfying results, they often fail to handle uneven light condition when working with severely degraded images. To tackle this problem explicitly, we need to find a feature that best discriminates *foreground* from *background* even if their difference is small. For this, we apply several different size filters on the *lightness* image. Although simple thresholding on the *lightness* image fails to binarize severely degraded images [7], it can be a good initial point for the proposed algorithm. Upon *lightness* image, the responses of filters, specially mean filters in this paper, work as a good descriptor for determining whether the current pixel belongs to *foreground* or *background*. By developing the proper measure and decision rule, the pixels are classified into 3 categories - text, near text and background regions, and misclassified pixels are relabeled by graph cut method.

This paper is organized as follows. We briefly review the retinex algorithm to introduce the *lightness* image in Sec.2 and the proposed binarization method on the *lightness* image is shown in Sec.3. In Sec.4, we compare the proposed algorithm with the retinex method, and extend our method to skew detection. Finally conclusions are given in Sec.5.



(a) (b)

Figure 2. *Lightness* **images of Fig. 1**

2 Retinex algorithm

Retinex theory is originated from the observation of a biological phenomenon that human visual system (HVS) is able to catch the consistent color of objects regardless of illumination conditions. For example, HVS perceives the color of white paper under red light as a white not a red. To mimic this HVS in the computer vision system, the retinex algorithm is designed to decompose the image into the illumination and the reflectance components. In the retinex algorithm it is assumed that image $I(x, y)$ is formulated as the multiplication of the illumination $L(x, y)$ and the reflectance $R(x, y)$, *i.e.,* $I(x, y) = L(x, y)R(x, y)$ and the reflectance component is obtained by the ratio as

$$\tilde{R}(x, y) = \frac{I(x, y)}{\tilde{L}(x, y)} \tag{1}$$

where $\tilde{L}(x, y)$ denotes the estimation of $L(x, y)$. The estimated reflectance $\tilde{R}(x, y)$ in eq. (1) is called *lightness* in the literatures of retinex theory.

In a binarization algorithm that exploits the retinex theory [7], the illumination component is estimated by gaussian filtering with large kernel under the assumption that illumination component varies smoothly throughout the whole image and the *lightness* is obtained by eq. (1). Fig. 2 shows the *lightness* images of Fig. 1.

3 Proposed binarization method

3.1 Classifying pixels into text, near text and background regions

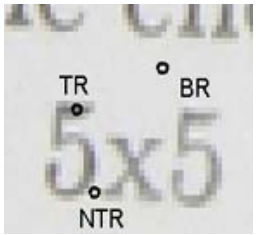
On the *lightness* image calculated in Sec. 2, we define a descriptor for each pixel, which is composed of several mean filter responses. Let $\gamma_p(n)$ be the response of a mean filter with support size of $2n + 1$ to the pixels centered at p. Also, let ρ_p a descriptor vector $\{\gamma_p(0), \gamma_p(1), \ldots, \gamma_p(M - 1)\}$ for the pixel p. Then the descriptor has three different patterns according to local variability as in Fig. 3. First, filter responses in the text region (*TR*) usually increase as the size of support becomes larger as in Fig. 3 (b), because the region of support includes higher level pixels as filter size increases. Second, the filter responses in near text region (*NTR*) in Fig. 3 (c) decrease because the region of support includes more low level pixels as the filter size gets larger. Lastly, filter responses in the background region (*BR*) tend to be flat because the region in support includes only high level pixels even if the size of support gets larger. With this observation we define a function $\Gamma : p \rightarrow \{TR, NTR, BR\}$ to determine the region to which the pixel p belongs. The first element $\gamma_p(0)$ in the descriptor and the variation of $\gamma_p(n)$ with respect to n play

Table 1. Functionality of $\Gamma(\rho_p)$

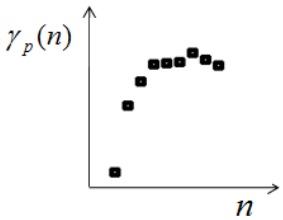
$\Gamma(\rho_p)$	condition
BR	$\nu_p \leq \mathcal{C}_\nu$
TR	$\nu_p > \mathcal{C}_\nu, \gamma_p(0) - \gamma_p^{min} < \mathcal{R}_\gamma$
NTR	$\nu_p > \mathcal{C}_\nu, \gamma_p^{max} - \gamma_p(0) \leq \mathcal{R}_\gamma$

an important role in the classification process. The background region is almost flat and has low variation, while other regions have high variations. Also, when the variation is high, text region and near text region can be distinguished by $\gamma_p(0)$ that is, the high variation with high $\gamma_p(0)$ states that the pixel belongs to the near text region and high variation with low $\gamma_p(0)$ means that the pixel belongs to text region. $\gamma_p(0)$ is called the prime element by its importance. In this paper the variation is measured by the difference of minima and maxima in ρ_p as $\nu_p = \gamma_p^{max} - \gamma_p^{min}$, and $\Gamma(\rho_p)$ is finally determined using $\gamma_p(0)$ and ν_p as summarized in Table 1, where $\mathcal{C}_\nu$ and $\mathcal{R}_\gamma$ denote the thresholds for the variation and the strength of prime element respectively. As can be seen in Table 1 the strength of prime element is defined as the difference with γ_p^{min} (or γ_p^{max}), which measures how close to γ_p^{min} (or γ_p^{max}) the prime element is. It is noted that the uniqueness condition that guarantees the pixel to belong to only one region should be satisfied as $\mathcal{C}_\nu/2 = \mathcal{R}_\gamma$.

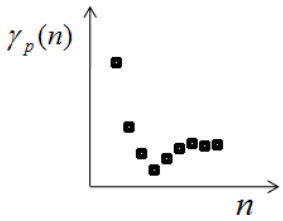
As a result of the mapping by $\Gamma(\rho_p)$ in Table 1, three regions are obtained but with noise as in Fig.4. To be specific, the pixels in *BR* are frequently misclassified into *TR* or *NTR* as seen in Fig. 4 (b) and (c), because captured image is itself noisy due to lossy compression, high ISO and so on. To remove the noise in *BR*, relabeling process based on energy minimization is applied to complete text region (*CTR*) which covers both *TR* and *NTR* as shown in Fig. 4 (d).



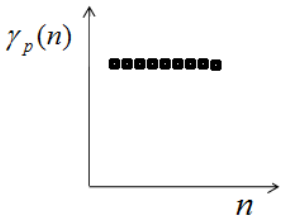
(a) Patch



(b) Filter responses in *TR*

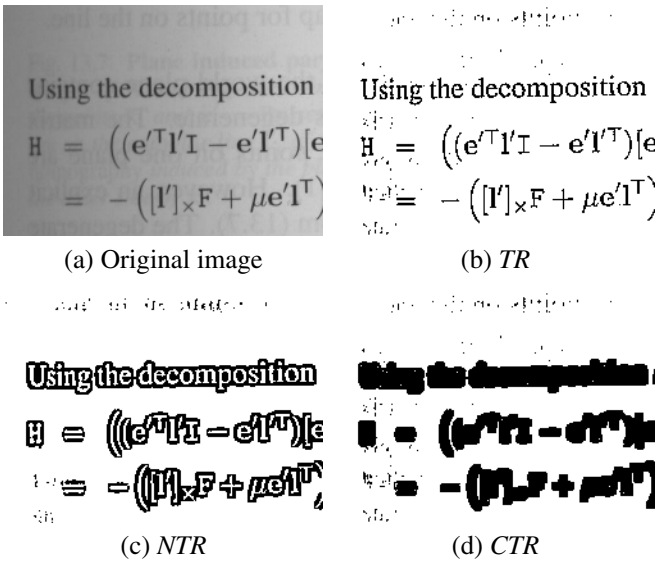


(c) Filter responses in *NTR*



(d) Filter responses in *BR*

Figure 3. Filter responses in various regions. Each dot denotes the value of $\gamma_p(n)$



(a) Original image (b) *TR*

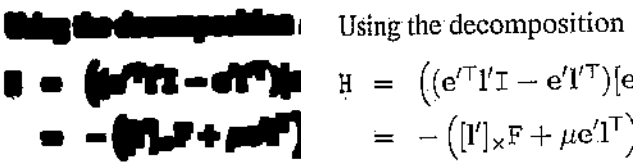
(c) *NTR* (d) *CTR*

Figure 4. Classified results

3.2 Relabeling process based on discrete minimization

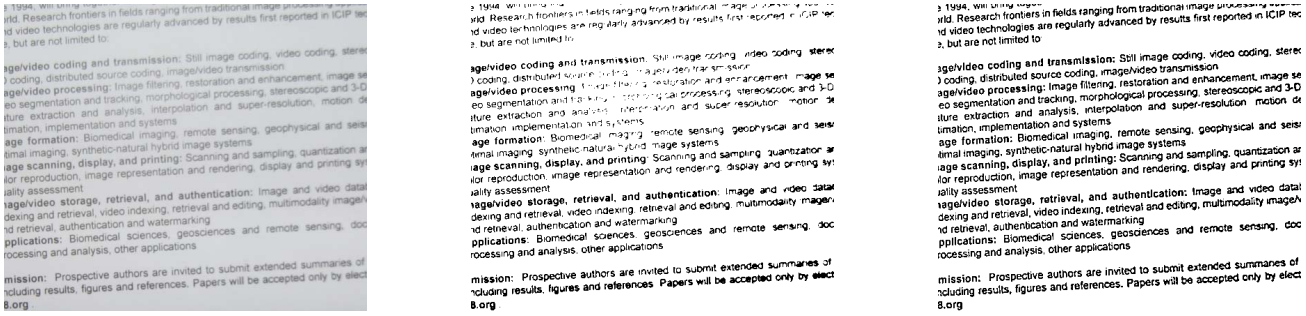
The problem of removing the noise can be seen as a labeling problem where noisy pixels in *TR* or *NTR* are relabeled to be in *BR*. Hence we adopt a discrete minimization technique, specially graph cut method. To work with a graph cut method, we need to formulate the energy function that leads to a successful labeling result when minimized [8]. In general, the energy, for which the graph cut can achieve the global solution, is defined by

$$E(f) = \sum_{p \in V} U_p(f_p) + \lambda \sum_{\{p,q\} \in E} w_{p,q} \delta_{p,q}(f_p, f_q) \quad (2)$$



(a) *CTR* after relabeling (b) *TR* after relabeling

Figure 5. Relabeled results of Fig. 4



(a) Original image (b) Retinex algorithm (c) Proposed algorithm

Figure 6. Experimental result

where V denotes a set of all pixels, E denotes a set of all pairs of neighboring pixels and λ is for adjusting the extent of smoothness. Also, $f_p \in \{CTR, BR\}$, U_p is a unary energy which measures how likely the pixel conforms to the label f_p and $w_{p,q}\delta_{p,q}(f_p, f_q)$ denotes pairwise energy which encodes discontinuity preserving smoothness between neighboring pixels. $\delta_{p,q}(f_p, f_q)$ is an indicator function whose value is 1 when $f_p \neq f_q$ and 0 when $f_p = f_q$ and $w_{p,q}$ denotes data-driven weight. In eq. (2) $U_p(f_p)$ and $w_{p,q}$ have to be designed for our relabeling process.

First the unary energy U_p in eq. (2) should be able to penalize the randomly distributed pixels *i.e.*, isolated pixels with the label of *CTR* in *BR* since the noisy pixels opt to be scattered randomly over the image as shown in Fig. 4 (d). The densities of *CTR* and *BR* at each pixel p can be a good clue for the measurement of isolation and thus U_p is designed to have small energy for high density by

$$U_p(f_p) = -\log \mathcal{N}_{f_p} \tag{3}$$

where $\mathcal{N}_{f_p}$ denotes the frequency of the pixels with label f_p in the 3×3 window centered at p and measures the density.

And let f_p^0 denote the initial label configuration after the mapping of $\Gamma(\rho_p)$ but has the label from the set $\{CTR, BR\}$. The densities of neighboring pixels $\mathcal{N}_{f_p^0}$ and $\mathcal{N}_{f_q^0}$ at the boundary between *CTR* and *BR* may be equally distributed and $w_{p,q}$ is thus designed to encode discontinuity preserving smoothness by

$$\begin{aligned} w_{p,q} &= 1 - \delta(f_p^0, f_q^0) \\ &+ \delta(f_p^0, f_q^0)(\mathcal{N}_{f_p^0} - \mathcal{N}_{f_q^0})^2. \end{aligned} \tag{4}$$

Neighboring pixels in eq. (4) are highly weighted to smooth the labels of neighboring pixels when the initial labels are same or when the densities are not equally distributed at the boundary. However the weight gets close to zero when the densities at the boundary are similarly distributed, which allows the neighboring pixels to have different labels. Incorporating the unary energy in (3) and the pairwise energy in (4) into (2) completes the energy function and the minimization based on graph cut method gives the clear result as shown in 5 (a) and (b).

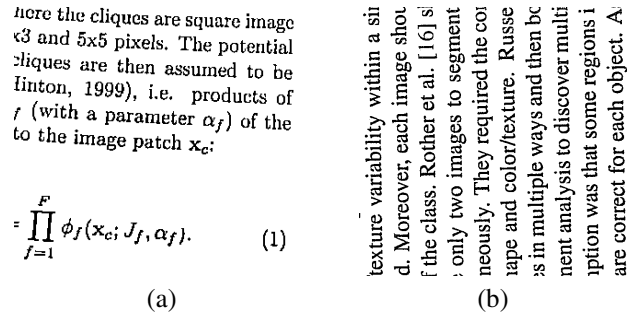
4 Experimental results

4.1 Binarization

In this section, we verify the proposed binarization algorithm on images degraded by uneven light condition as in Fig. 6 (a). The result of the proposed algorithm is compared with a retinex-based binarization algorithm [7]. As can be seen in Fig. 6 (a) we have both faint and clear text regions and the retinex algorithm fails to binarize it as in Fig. 6 (b). Compared to the retinex algorithm, the proposed method gives robust result over whole image. The results of proposed method on Fig. 1 are also shown in Fig. 7.

4.2 Skew detection

In this section we simply extend our method to skew detection based on the observation that the *CTR* gives useful information on the connectivity of the word. Hence



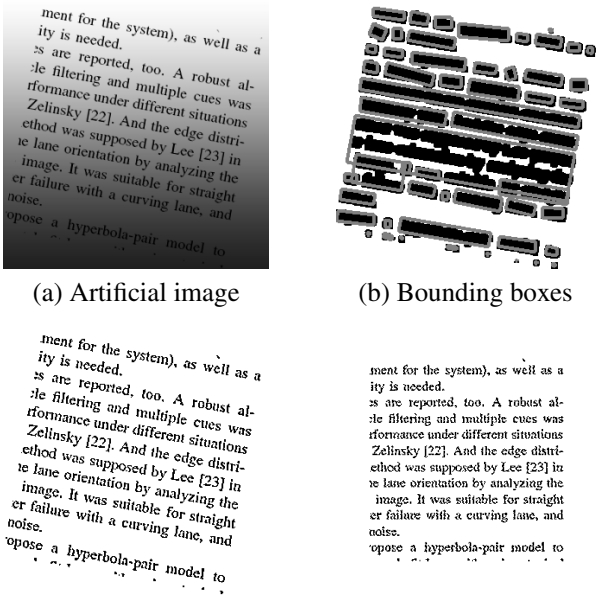
(a) (b)

Figure 7. Results of proposed method on Fig. 1

the analysis of each connected component in *CTR* enables to detect the document skew. In this paper we adopt and modify skew detection algorithm in [9] which uses minimum area bounding rectangle. First we find the angle π_c which has minimum area α_c among all candidates for each connected component c and calculate the confidence $\kappa_c = \alpha_c R_c$ where R_c denotes the aspect ratio of the rectangle. By its definition the confidence is getting higher as the area becomes larger and the aspect ratio gets larger. And finally, the skew angle is calculated by

$$\pi_c = \sum_{c \in N_c} \alpha_c \kappa'_c \tag{5}$$

where N_c is the number of connected components in *CTR* and κ'_c is normalized confidence. To verify the algorithm we have performed experiments on the artificial image which is skewed by 10 degrees and degraded by gradation as in Fig. 8 (a). As can be seen in Fig. 8 (b), each connected component in *CTR* is bounded by rectangular box and the skewed angle is calculated to be 9.86 degrees by eq. (5). Hence the binarized image in Fig. 8 (c) can be corrected as in Fig. 8 (d).



(a) Artificial image (b) Bounding boxes

(c) *TR* (d) Corrected *TR*

Figure 8. Skew detection

5 Conclusions

We have proposed an algorithm for the robust binarization of document images, especially the ones degraded by uneven light condition. For this purpose we define a descriptor that measures the variability of pixel values around a given pixel. The descriptor is a vector that is composed of the responses of several filters around a given pixel. The descriptor gives much information for classifying each pixel into the background region, text region and near text region. After this classification, the noisy pixels are relabeled using the graph cut energy minimization method. An energy function for this scheme has been designed with the focus on generating visually pleasing results. It has also been shown that the proposed descriptor can be used for the skew detection, and thus for the correction of skewed images.

References

[1] Z. Shi and V. Govindaraju, "Historical document image enhancement using background light intensity normalization," *ICDAR*, pp. 473-476, 2004.

[2] M. Ramirez, E. Tapia, M. block and R. Rojas, "Quantile linear algorithm for robust binarization of digitalized letters," *ICDAR*, pp. 1158-1162, 2007.

[3] Y. Li, C. Suen and M. Cheriet, "A threshold selection method mased on multiscale and graylevel co-occurence matrix analysis," *ICDAR*, pp. 575-579, 2005.

[4] Y. Xi, Y. Chen and Q. Liao, "A novel binarization system for degraded document images," *ICDAR*, pp. 287-291, 2007.

[5] W. Niblack, "An introduction to image processing," pp. 115-116, Prentice Hall, Englewood Cliffs, NJ, 1986.

[6] P. W. Palumbo, P. Swaminathan and S. N. Srihari, "Document image binarization : Evaluation of algorithms", *Proc. SPIE*, pp. 278-286, 1986.

[7] M. Pilu and S. Pollard, "A light-weight text image processing method for handheld embedded cameras", *bmvc*, Mar., 2002.

[8] Y. Y. Boykov and M. P. Jolly, "interactive graph cuts for optimal boundary & region segmentation of objects in N-D images," *Int. Conf. on Computer Vision*, vol. 1, pp. 105-112, July, 2001.

[9] R. Safabakhsh and S. Khadivi, "Document skew detection using minimum-area bounding retangle," *Int. Conf. on Information Techonology:Coding and Computing*, pp.253-258, March, 2000.