SPECIALIZED TEXT BINARIZATION TECHNIQUE FOR CAMERA-BASED DOCUMENT IMAGES

T Kasar, J Kumar and A G Ramakrishnan

Medical Intelligence and Language Engineering Laboratory
Department of Electrical Engineering, Indian Institute of Science
Bangalore, INDIA - 560 012
tkasar@ee.iisc.ernet.in, jayantmishra@gmail.com, ramikag@ee.iisc.ernet.in

ABSTRACT

Complex color documents with both graphics and text, where the text varies in color and size, call for specialized binarization techniques. We propose a novel method for binarization of color documents whereby the foreground text is output as black and the background as white regardless of the polarity of foreground and background shades. The method employs an edge-based connected component approach to determine text-like components and binarize them individually. The threshold for binarization and the logic for inverting the output are derived from the image data and do not require any manual tuning. Unlike existing binarization methods, our technique can handle documents with multi-colored texts with different background shades. The method is applicable to documents having text of widely varying sizes, usually not handled by local binarization methods. Experiments on a broad domain of target document types illustrate the effectiveness and adaptability of the method.

Index Terms— Binarization, Color documents, Camera-based document analysis

1. INTRODUCTION

In acquiring document images, there has been an increased use of cameras as an alternative to traditional flat-bed scanners and research towards camera based document analysis is growing [1]. Digital cameras are compact, easy to use, portable and offer a high-speed non-contact mechanism for image acquisition. The use of cameras has greatly eased document acquisition and has enabled human interaction with any type of document. It has several potential applications like licence plate recognition, road sign recognition, digital note taking, document archiving and wearable computing. At the same time, it has also presented us with much more challenging images for any recognition task. Traditional scanner-based document analysis systems fail against this new and promising acquisition mode. Camera images suffer from uneven lighting, low resolution, blur, and perspective distortion. Overcoming these challenges will help us tap the potential advantages of camera-based document analysis.

2. REVIEW OF EARLIER WORK

Binarization often precedes any document analysis and recognition procedures. It is critical to achieve robust binarization since any error introduced in this stage will affect the subsequent processing steps. The simplest and earliest method is the global thresholding technique that uses a single threshold to classify image pixels into foreground or background class. Global thresholding techniques are generally based on histogram analysis [2, 3]. They work well for images with well separated foreground and background intensities. However, most of the document images do not meet this condition and hence the application of global thresholding methods is limited. Camera-captured images often exhibit non-uniform brightness because it is difficult to control the imaging environment as much as we can with the scanner. As such, global binarization methods are not suitable for camera images. On the other hand, local methods use a dynamic threshold across the image according to the local information. These approaches are generally window-based and the local threshold for a pixel is computed from the gray values of the pixels within a window centred at that particular pixel. Niblack [4] proposed a binarization scheme where the threshold is derived from the local image statistics. The sample mean $\mu(x, y)$ and the standard deviation $\sigma(x, y)$ within a window W centred at the pixel location (*x,y*) are used to compute the threshold $\mathrm{T}(x, y)$ as follows:

$$\mathrm{T}(x, y) = \mu(x, y) - k\,\sigma(x, y), \quad k = 0.2 \qquad (1)$$

Trier and Jain [5] evaluated 11 popular local thresholding methods on scanned documents and reported that Niblack's method performs the best for optical character recognition (OCR). The method works well if the window encloses at least 1-2 characters. However, in homogeneous regions larger than the size of the window, the method produces a noisy

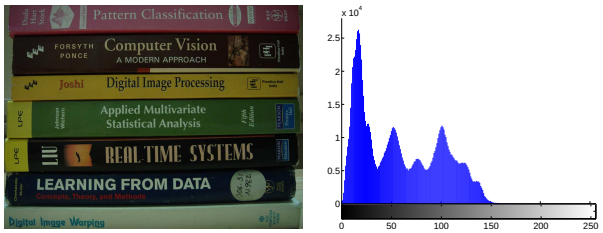


Fig. 1. An example image with multi-colored textual content and its gray level histogram. A conventional binarization technique, using a fixed foreground-background polarity, will treat some of the characters as background pixels leading to the loss of some textual information.

output since the expected sample variance becomes the background noise variance. Sauvola and Pietikainen [6] proposed an improved version of the Niblack's method by introducing a hypothesis that the gray values of the text are close to 0 (Black) while the background pixels are close to 255 (White). The threshold is computed with the dynamic range of standard deviation (R) which has the effect of amplifying the contribution of standard deviation in an adaptive manner.

$$T(x,y) = \mu(x,y)\,[1 + k\,(\frac{\sigma(x,y)}{R} - 1)] \qquad (2)$$

where the parameters R and k are set to 128 and 0.5 respectively. This method minimizes the effect of background noise and is more suitable for document images. However, Sauvola method fails for images where the assumed hypothesis is not met and accordingly, Wolf and Jolion [7] proposed an improved threshold estimate by taking the local contrast measure into account.

$$T(x,y) = (1-a)\mu(x,y) + a\mathrm{M} + a\frac{\sigma(x,y)}{\mathrm{S}_{max}}(\mu(x,y) - \mathrm{M}) \quad (3)$$

where M is the minimum value of the grey levels of the whole image, S_{max} is the maximum value of the standard deviations of all the windows of the image and 'a' is a parameter fixed at 0.5. This method combines Savoula's robustness with respect to background textures and the segmentation quality of Niblack's method. However it requires two passes since one of the threshold decision parameter S_{max} is the maximum of the standard deviation of all the windows of the images.

With the recent developments on document types, more specialized binarization techniques are required to handle complex documents having both graphics and text. We often encounter text of different colors in a document image as shown in Fig. 1. Conventional methods assume that the polarity of the foreground-background intensity is known a priori. The text is generally assumed to be either bright on a dark background or vice versa. Binarization using a single threshold on such images, without a priori information of the polarity of foreground-background intensities, will lead to loss of textual information as some of the text may be assigned as background. The characters once lost cannot be retrieved back and are not available for further processing. Possible solutions need to be sought to overcome this drawback so that any type of document could be properly binarized without the loss of textual information.

3. SPECIALIZED TEXT BINARIZER

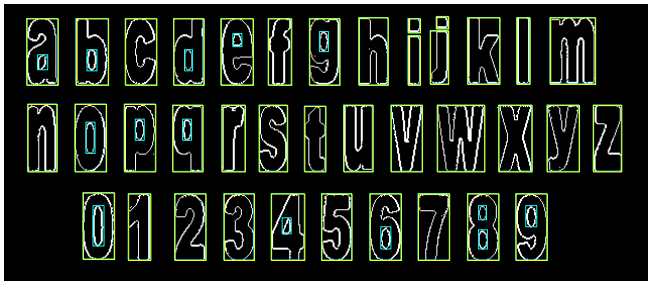
Text is the most important information in a document. We propose a novel method to binarize camera-captured color document images, whereby the foreground text is output as black and the background as white irrespective of the original polarities of foreground-background shades. The proposed method uses an edge-based connected component approach and determines the threshold for each component individually. Canny edge detection [8] is performed individually on each channel of the color image and the edge map $\mathbf{E}$ is obtained by combining the three edge images as follows

$$\mathbf{E} = \mathbf{E}_R \vee \mathbf{E}_G \vee \mathbf{E}_B \qquad (4)$$

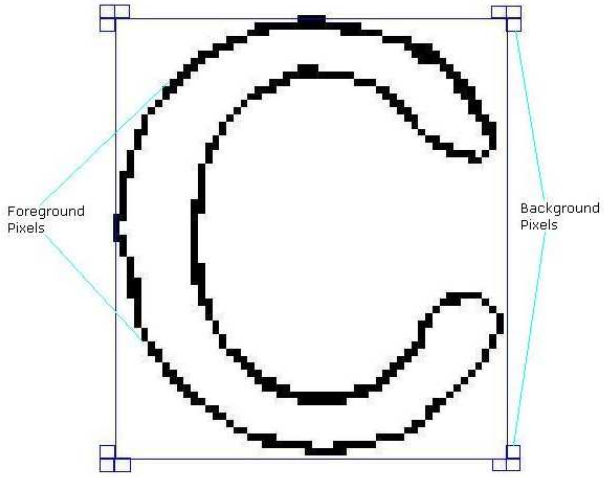
Here, $\mathbf{E}_R$, $\mathbf{E}_G$ and $\mathbf{E}_B$ are the edge images corresponding to the three color channels and $\vee$ denotes the logical OR operation. We have used the thresholds 0.2 and 0.3 for the hysteresis thresholding step of Canny edge detection. An 8-connected component labeling follows the edge detection step and the associated bounding box information is computed. We call each component, thus obtained, an edge-box (EB). We make some sensible assumptions about the document and use the area and the aspect ratios of the EBs to filter out the obvious non-text regions. The aspect ratio is constrained to lie between 0.1 and 10 to eliminate highly elongated regions. The size of the EB should be greater than 15 pixels but smaller than 1/5th of the image dimension to be considered for further processing. Since edge detection captures both the inner and outer boundaries of the characters, it is possible that an EB may completely enclose one or more EBs as illustrated in Fig. 2(a). If a particular EB has exactly one or two EBs that lie completely inside it, the internal EBs can be conveniently ignored as it corresponds to the inner boundaries of the text characters. On the other hand, if it completely encloses three or more EBs, only the internal EBs are retained while the outer EB is removed as such a component does not represent a text character. Thus, the unwanted components are filtered out by subjecting each edge component to the following constraint:

```
if (N_int <3)
   {Reject EB_int,  Accept EB_out}
else
   {Reject EB_out,  Accept EB_int}
```

where EB_{int} denotes the EBs that lie completely inside the current EB under consideration and N_{int} is the number of



(a)



(b)

Fig. 2. (a) Edge-boxes for the English alphabet and numerals. Note that there is no character that completely encloses more than two edge components. (b) The foreground and the background pixels of each edge component

EB_{int}. These constraints on the edge components effectively remove the obvious non-text elements, while retaining all the text-like elements. Only these filtered set of EBs are considered for binarization.

For each EB, we estimate the foreground and background intensities and the threshold is computed individually. Fig. 2(b) shows the foreground and the background pixels which are used for obtaining the threshold and inversion of the binary output. The foreground intensity is computed as the mean gray-level intensity of the pixels that correspond to the edge pixels.

$$\mathbf{F}_{EB} = \frac{1}{\mathrm{N}_E} \sum_{(x,y)\in\mathbf{E}} \mathbf{I}(x,y) \tag{5}$$

where $\mathbf{E}$ represents the edge pixels, $\mathbf{I}(x,y)$ represents the intensity value at the pixel (x,y) and N_E is the number of edge pixels in an edge component. For obtaining the background intensity, we consider three pixels each at the periphery of the corners of the bounding box as follows

$$\begin{aligned} B = \{&\mathbf{I}(x-1,y-1), \mathbf{I}(x-1,y), \mathbf{I}(x,y-1), \\ &\mathbf{I}(x+w+1,y-1), \mathbf{I}(x+w,y-1), \mathbf{I}(x+w+1,y), \\ &\mathbf{I}(x-1,y+h+1), \mathbf{I}(x-1,y+h), \mathbf{I}(x,y+h+1), \\ &\mathbf{I}(x+w+1,y+h+1), \mathbf{I}(x+w,y+h+1), \\ &\mathbf{I}(x+w+1,y+h)\} \end{aligned}$$

where (x, y) represent the coordinates of the top-left corner of the bounding-box of each edge component and w and h are its width and height, respectively. The local background intensity is then computed as the median intensity of these 12 background pixels.

$$\mathbf{B}_{EB} = \text{median}(B) \tag{6}$$

Assuming that each character is of uniform color, we binarize each edge component using the estimated foreground intensity as the threshold ($\mathbf{T}_{EB}$). Depending on whether the foreground intensity is higher or lower than that of the background, each binarized output is suitably inverted so that the foreground text is always black and the background, always white.

$$\mathbf{T}_{EB} = \begin{cases} \mathbf{F}_{EB}, & if\, \mathbf{F}_{EB} < \mathbf{B}_{EB} \\ (255 - \mathbf{F}_{EB}), & if\, \mathbf{F}_{EB} > \mathbf{B}_{EB} \end{cases} \tag{7}$$

4. EXPERIMENTS

The test images used in this work are acquired using a Sony digital still camera at a resolution of 1280 $\times$ 960. The images are taken from both physical documents such as book covers and newspapers and non-paper document images like text on 3-D real world objects. Fig. 3 compares the results of our method with some popular local binarization techniques, viz, Niblack's method, Sauvola's method and Wolf's method on a document image having large variation in text sizes with the smallest and the largest components being 5 $\times$ 16 and 414 $\times$ 550 respectively. Clearly, these local binarization methods fail when the size of the window is smaller than the stroke width. A large character is broken up into several components and undesirable voids occur within thick characters. It requires a priori knowledge of the polarity of foreground-background intensities as well. On the other hand, our method can deal with characters of any size and color as it only uses edge connectedness. The generality of the algorithm is tested on 50 complex color document images and is found to have a high adaptivity and performance. Some results of binarization using our method are shown in Fig. 4. The algorithm deals only with the textual information and it does not threshold the edge components that were already filtered out. In the resulting binary images, as desired, all the text regions are output as black while the background as white, irrespective of their colors in the input images.

(a) Input image (b) Niblack (c) Sauvola

(d) Wolf (e) Proposed

Fig. 3. Comparison of the proposed method with some popular local binarization methods for a document image with large variation in text size. While our method is able to handle characters of any size and color, all other methods fail to binarize properly the components larger than the size of the window (35×35 used here).

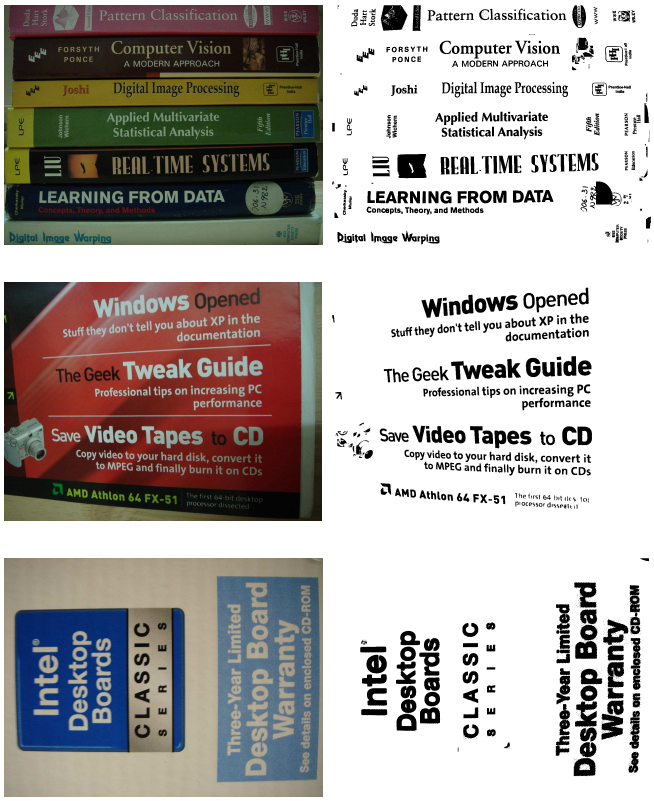


Fig. 4. Some examples of binarization results obtained using the proposed method. All the text regions are output as black and the background as white, irrespective of their original colors in the input images.

5. CONCLUSIONS AND FUTURE WORK

We have developed a specialized binarization technique well-suited for camera-based document images. It has good adaptability without the need for manual tuning and can be applied to a broad domain of target document types. It simultaneously handles the ambiguity of the polarity of the foreground-background intensities and the algorithm's dependency on the parameters. The edge-box analysis captures all the characters, irrespective of size and color, thereby enabling us to perform local binarization without the need to specify any window. The proposed method retains the useful textual information more accurately and thus, has a wider range of target document types compared to conventional methods.

The edge detection method is good in finding the character boundaries irrespective of the foreground-background polarity. However, if the background is textured, the edge components may not be detected correctly due to edges from the background. This can affect the performance of our edge-box filtering strategy. Overcoming these challenges is considered as a future extension to this work.

6. REFERENCES

[1] D. Doermann, J. Liang, and H. Li, "Progress in camera-based document image analysis," *ICDAR*, vol. 1, pp. 606–615, 2003.

[2] J. N. Kapur, P. K. Sahoo, and A.K.C. Wong, "A new method for gray-level picture thresholding using the entropy of the histogram," *Computer Vision Graphics Image Process.*, vol. 29, pp. 273–285, 1985.

[3] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Systems Man Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[4] W. Niblack, "An introduction to digital image processing," *Prentice Hall*, pp. 115–116, 1986.

[5] O. D. Trier and A.K. Jain, "Goal-directed evaluation of binarization methods," *IEEE Trans. PAMI*, vol. 17, no. 12, pp. 1191–1201, 1995.

[6] J. Sauvola and M. Pietikainen, "Adaptive document image binarization," *Pattern Recognition*, vol. 33, pp. 225–236, 2000.

[7] C. Wolf and J.M. Jolion, "Extraction and recognition of artificial text in multimedia documents," *Pattern Analysis and Applications*, vol. 6, no. 4, pp. 309–326, 2003.

[8] J. Canny, "A computational approach to edge detection," *IEEE Trans. PAMI*, vol. 8, no. 6, pp. 679–698, 1986.