# Font and Background Color Independent Text Binarization

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, jayantkmishra@gmail.com, ramikag@ee.iisc.ernet.in

## Abstract

*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-background shades. The method employs an edge-based connected component approach and automatically determines a threshold for each component. It has several advantages over existing binarization methods. Firstly, it can handle documents with multi-colored texts with different background shades. Secondly, the method is applicable to documents having text of widely varying sizes, usually not handled by local binarization methods. Thirdly, the method automatically computes the threshold for binarization and the logic for inverting the output from the image data and* ***does not require any input parameter****. The proposed method has been applied to a broad domain of target document types and environment and is found to have a good adaptability.*

## 1 Introduction

There has been an increased use of cameras in acquiring document images as an alternative to traditional flatbed scanners and research towards camera based document analysis is growing [3]. 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. Its ability to capture non-paper document images like scene text has several potential applications like licence plate recognition, road sign recognition, digital note taking, document archiving and wearable computing. But 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 effortlessly acquire and manage information in documents.

In most document processing systems, a binarization process precedes the analysis and recognition procedures. The use of two-level information greatly reduces the computational load and the complexity of the analysis algorithms. 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 classes. Global thresholding techniques are generally based on histogram analysis [4, 6]. It works 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 unlike the case of the scanner. The histogram of such images are generally not bi-modal and a single threshold can never yield an accurate binary document image. 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 [5] 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 $T_{(x,y)}$ as follows:

$$T_{(x,y)} = \mu_{(x,y)} - k\,\sigma_{(x,y)}, \quad k = 0.2 \tag{1}$$

Yanowitz and Bruckstein [10] introduced a threshold that varies over different image regions so as to fit the spatially changing background and lighting conditions. Based on the observation that the location and gray level values at the

edge points of the image are good choices for local thresholds, a threshold surface is created by relaxation initialized on the edge points. The method is however computationally very intensive. Trier and Jain [8] 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 size of the window, the method produces a noisy output since the expected sample variance becomes the background noise variance. Sauvola and Pietikainen [7] 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.

$$T_{(x,y)} = \mu_{(x,y)} \left[1 + k \left(\frac{\sigma_{(x,y)}}{R} - 1\right)\right] \quad (2)$$

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. As pointed out by Wolf *et al* in [9], the Sauvola method fails for images where the assumed hypothesis is not met and accordingly, they 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, $\mathrm{S}_{max}$ is the maximum value of the standard deviations of all windows of the image and 'a' is a parameter fixed at 0.5. The Wolf's method requires two passes since one of the threshold decision parameter $\mathrm{S}_{max}$ is the maximum of all standard deviation of all windows of the images. The computational complexity is therefore slightly higher in this case. This method combines Savoula's robustness with respect to background textures and the segmentation quality of Niblack's method.

However, recent developments on document types, for example, documents with both graphics and text, where the text varies in color and size, call for more specialized binarization techniques. It is relatively difficult to obtain satisfactory binarization with various kinds of document images. The choice of window size in local methods can severely affect the result of binarization and may give rise to broken characters and voids, if the characters are thicker than the size of the window considered. Moreover, we often encounter text of different colors in a document image. Conventional methods assume that the polarity of the foreground-background intensity is known a priori. The text


**Figure 1. Some example images with multi-colored textual content and varying background shades. A conventional binarization technique, using a fixed foreground-background polarity, will treat some characters as background, leading to the loss of some textual information**

is generally assumed to be either bright on a dark background or vice versa. If the polarity of the foreground-background intensity is not known, the binary decision logic could treat some text as background and no further processing can be done on those text. Clark and Mirmhedi [2] use a simple decision logic to invert the result of binarization based on the assumption that the background pixels far outnumber the text pixels. Within each window, the number of pixels having intensity values higher or lower than the threshold are counted and the one which is less in number is treated as the foreground text. This simple inversion logic cannot handle the case where the characters are thick and occupy a significant area of the window under consideration. Moreover, a document image can have two or more different shades of text with different background colors as shown in Fig. 1. 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.

## 2 System Description

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 polarity of foreground-background shades. The proposed method uses an edge-based connected component approach to automatically obtain a threshold for each component. Canny edge detection [1] 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

$$\mathbf{E} = \mathbf{E}_R \vee \mathbf{E}_G \vee \mathbf{E}_B \tag{4}$$

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. 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.

**Figure 2. Edge-boxes for the English alphabet and numerals. Note that there is no character that completely encloses more than two edge components**

Since the 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. For example, the letter 'O' gives rise to two components; one due to the inner boundary $\text{EB}_{int}$ and the other due to the outer boundary $\text{EB}_{out}$. 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

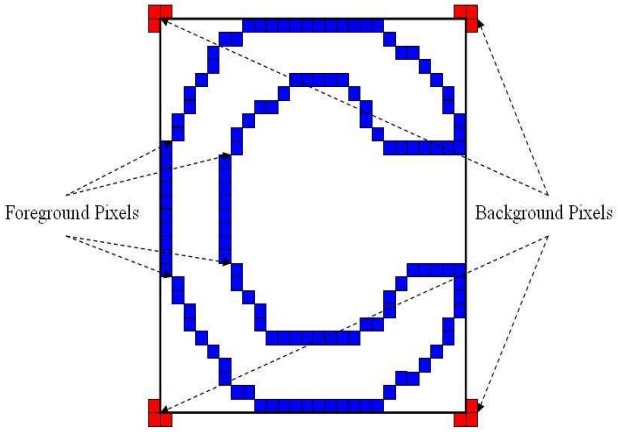



**Figure 3. The foreground and the background pixels of each edge component**

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 constraints:

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

where $\text{EB}_{int}$ denotes the EBs that lie completely inside the current EB under consideration and $\text{N}_{int}$ is the number of $\text{EB}_{int}$. These constraints on the edge components effectively remove the obvious non-text elements while retaining all the text-like elements. Only the filtered set of EBs are considered for binarization.

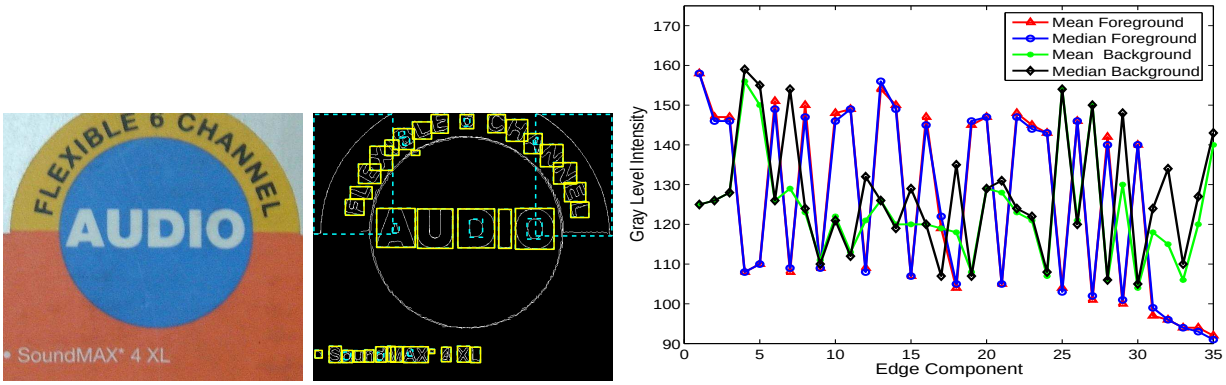
## 3 Estimation of Threshold

For each EB, we estimate the foreground and background intensities and the threshold is computed individually. Fig. 3 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.

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

where $\mathbf{E}$ represent the edge pixels, $\mathbf{I}(x,y)$ represent the in-




(a) (b) (c)

**Figure 4. (a) Input Image (b) Output of Edge-Box filtering. The dotted boxes in cyan are filtered out and only the yellow solid boxes (35 in number) are considered for binarization (c) The threshold parameters for the valid edge components. Observe that the mean and median intensities of the foreground pixels are almost the same for all characters. The same holds true for the background estimate for horizontally (or vertically) aligned text. However, when the text is aligned diagonally, the mean intensity of the background pixels is affected due to overlapping of the adjacent bounding boxes. Hence, the median intensity gives a more reliable logic for inverting the binary output**

tensity 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

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

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. Fig. 4 shows the output of the edge-box filtering and the threshold parameters for each of the valid edge components. As it is observed in Fig. 4(c), the mean or median intensity are almost the same for the foreground pixels irrespective of the text orientation. However, for a diagonally aligned text, the bounding boxes can have some overlap with the adjacent components and can interfere in the background intensity estimate. This is the case for the text 'FLEXIBLE 6 CHANNEL' printed in black in a semi-circular manner which are represented in Fig. 4(c) by the edge components whose estimated foreground intensity is lower than that of the background. The mean background intensity for these components are affected by the adjacent components while the median is not. Thus, the local background intensity can be estimated more reliably by considering the median intensity of the 12 background pixels instead of the mean intensity.

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

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

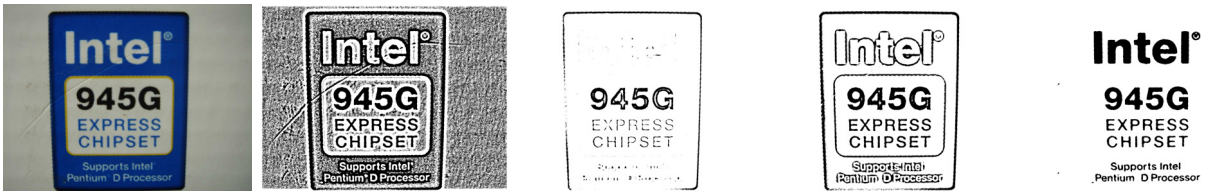
$$\text{If } F_{EB} < B_{EB},\ \mathbf{BW}_{EB}(x,y) = \begin{cases} 1, \mathbf{I}(x,y) \geq F_{EB} \\ 0, \mathbf{I}(x,y) < F_{EB} \end{cases} \tag{7}$$

$$\text{If } F_{EB} > B_{EB},\ \mathbf{BW}_{EB}(x,y) = \begin{cases} 0, \mathbf{I}(x,y) \geq F_{EB} \\ 1, \mathbf{I}(x,y) < F_{EB} \end{cases} \tag{8}$$

All the threshold parameters explained in this section are derived from the image data and the method is thus completely free from user-defined parameters.

## 4 Experiments

The test images used in this work are acquired from a Sony digital still camera at a resolution of 1280 $\times$ 960. The images are taken from both physical documents such




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

**Figure 5. Comparison of some popular local binarization methods for (a) a document image with multiple text colors and sizes (b) Niblack's Method (c) Sauvola's Method and (d) Wolf's Method with the (e) Proposed method. While the proposed 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 (25 $\times$ 25 used here) and require a priori knowledge of the polarity of foreground-background intensities as well**

.

as book covers, newspapers etc and non-paper document images like text on 3-D real world objects. The connected component analysis performed on the edge map captures all the text characters irrespective of the polarity of their foreground and background intensities. We have used the thresholds 0.2 and 0.3 for the hysteresis thresholding step of Canny edge detection. The variance of the associated Gaussian function is taken to be 1. The constraints on the edge components effectively removes the obvious non-text elements while retaining all the text-like components. From each valid edge component, the foreground and background intensities are automatically computed and each of them is binarized individually.

Fig. 5 compares the results of our method with some popular local binarization techniques, namely, Niblack's method, Sauvola's method and Wolf's method on a document image having multi-colored text and large variations in sizes with the smallest and the largest components being $7\times5$ to $291\times174$ respectively. Clearly, these local binarization methods fail when the size of the window is smaller than 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 binarization logic developed here is tested on documents having foreground text with different background shades. Though these kinds of images are quite commonly encountered, none of the existing binarization techniques can deal with such images. The generality of the algorithm is tested on more than 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. 6 adjacent to its respective input images. 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.

## 5 Conclusions and Future Work

We have developed a novel technique for binarization of text from digital camera images. It has a good adaptability without the need for manual tuning and can be applied to a broad domain of target document types and environment. It simultaneously handles the ambiguity of polarity of the foreground-background shades and the algorithm's dependency on the parameters. The edge-box analysis captures all the characters, irrespective of their sizes thereby enabling us to perform local binarization without the need to specify any window. The use of edge-box has enabled us to automatically compute the foreground and the background intensities reliably and hence the required threshold for binarization. The proposed method retains the useful textual information more accurately and thus, has a wider range of applications compared to other 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 and our edge-box filtering strategy fails. This has been observed for the image shown in Fig. 7. Overcoming these challenges is considered as a future extension to this work. The method is able to capture all the text while at the same time, filter out most of the components due to the background. The method can be extended to incorporate text localization as well.

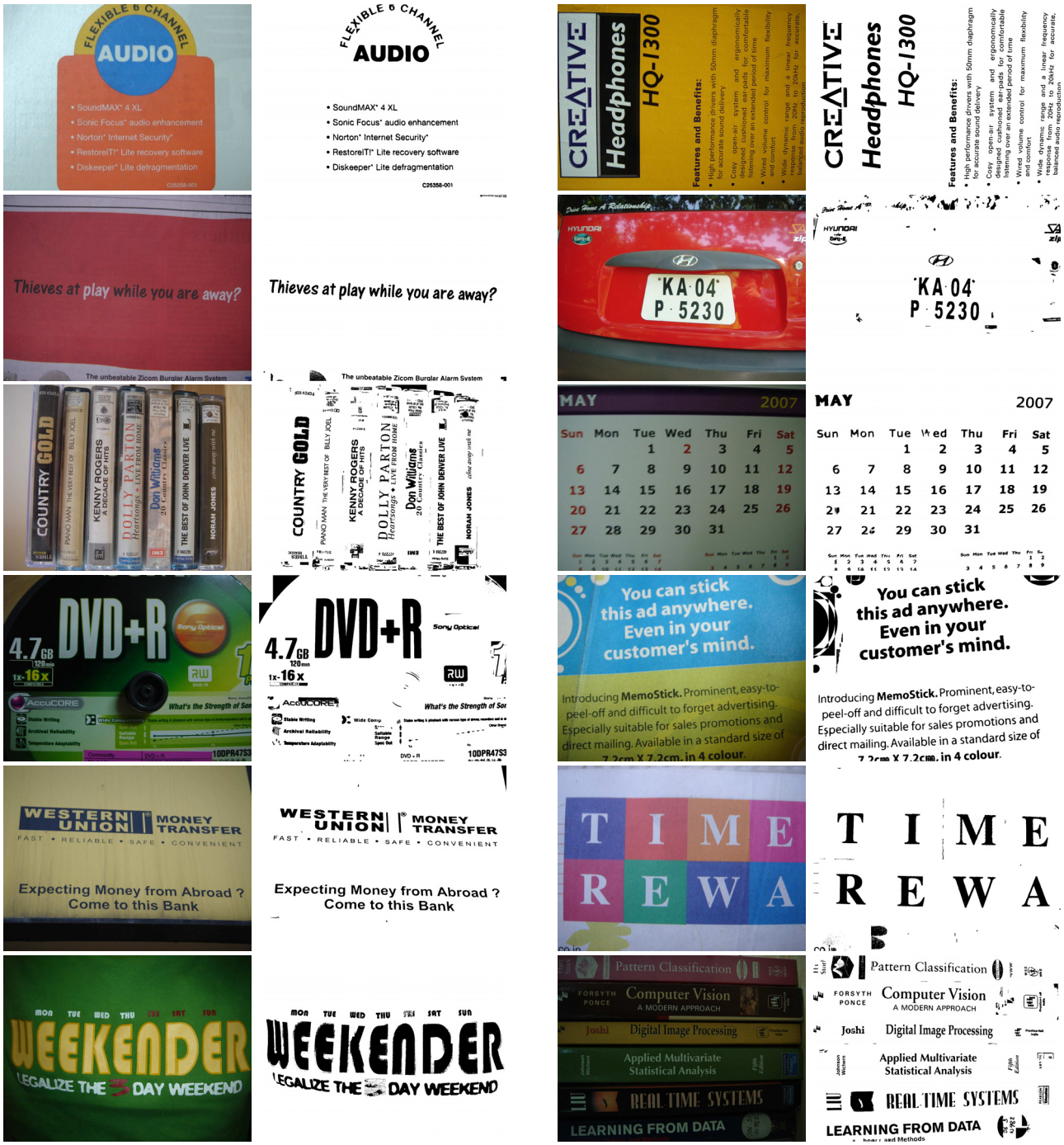



**Figure 6. Some examples of binarization results obtained using the proposed method. Based on the estimated foreground and background intensities, each binarized component is suitably inverted so that all the text are represented in black and the background in white**


**Figure 7. An example image for which the proposed algorithm fail to binarize properly**

## 6 Acknowledgements

The authors wish to thank the anonymous reviewers for their useful comments and suggestions. The author, T. Kasar, would also like to thank his colleague A. Bhavna for the useful discussions on several aspects of this work.

## References

[1] J. Canny. A computational approach to edge detection. *IEEE trans. PAMI*, 8(6):679–698, 1986.

[2] P. Clark and M. Mirmhedi. Rectifying perspective views of text in 3-d scenes using vanishing points. *Pattern Recognition*, 36:2673–2686, 2003.

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

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

[5] W. Niblack. An introduction to digital image processing. *Prentice Hall*, pages 115–116, 1986.

[6] N. Otsu. A threshold selection method from gray-level histograms. *IEEE Trans. Systems Man Cybernetics*, 9(1):62–66, 1979.

[7] J. Sauvola and M. Pietikainen. Adaptive document image binarization. *Pattern Recognition*, 33:225–236, 2000.

[8] O. D. Trier and A. Jain. Goal-directed evaluation of binarization methods. *IEEE Trans. PAMI*, 17(12):1191–1201, 1995.

[9] C. Wolf, J. Jolion, and F. Chassaing. Text localization, enhancement and binarization in multimedia documents. *ICPR*, 4:1037–1040, 2002.

[10] S. Yanowitz and A. Bruckstein. A new method for image segmentation. *Computer Vision, Graphics and Image Processing*, 46(1):82–95, 1989.