International Journal of Computer Applications Technology and Research
Volume 3– Issue 9, 559 - 563, 2014, ISSN: 2319–8656

A Novel Document Image Binarization For Optical Character Recognition

Varada V M Abhinay
S.V. College of Engineering
Tirupati, Andhra Pradesh, India

P.Suresh Babu
S.V. College of Engineering
Tirupati, Andhra Pradesh, India

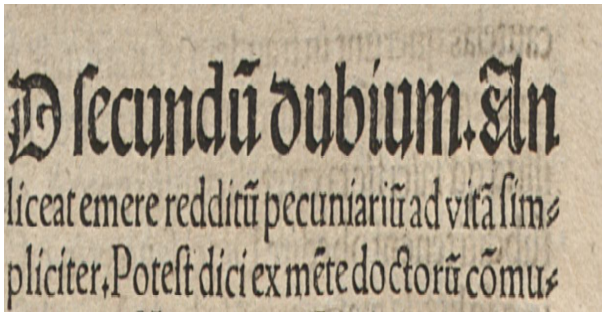
Abstract: This paper presents a technique for document image binarization that segments the foreground text accurately from poorly degraded document images. The proposed technique is based on the Segmentation of text from poorly degraded document images and it is a very demanding job due to the high variation between the background and the foreground of the document. This paper proposes a novel document image binarization technique that segments the texts by using adaptive image contrast. It is a combination of the local image contrast and the local image gradient that is efficient to overcome variations in text and background caused by different types degradation effects. In the proposed technique, first an adaptive contrast map is constructed for a degraded input document image. The contrast map is then binarized by global thresholding and pooled with Canny's edge map detection to identify the text stroke edge pixels. By applying Segmentation the text is further segmented by a local thresholding method that. The proposed method is simple, strong, and requires minimum parameter tuning.
Keywords:Adaptive image contrast, document analysis, pixel intensity, pixel classification.

1. INTRODUCTION

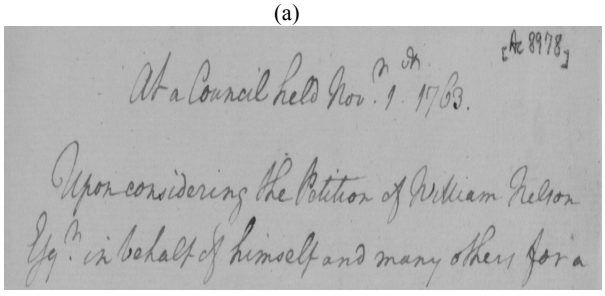
Document image binarization is a preprocessing stage for various document analyses. As more and more number of text document images is scanned, speedy and truthful document image binarization is becoming increasingly important. As document image binarization [1] has been studied for last many years but the thresholding techniques of degraded document images is still an unsettled problem. This can be explained by the difficulty in modeling different types of document degradation such as change in image contrast, uneven illumination, smear and bleeding-through that exist in many document images as illustrated in Fig. 1.

The printed text within the degraded documents often shows a certain amount of variation in terms of the stroke brightness, stroke connection, stroke width and document image background. A large number of document image thresholding techniques have been reported in the literature. For document images of a good quality, global thresholding is efficiently capable to extract the document text.

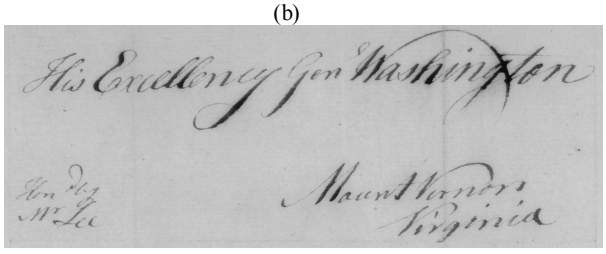
But for document images suffering from different types of document degradation, adaptive thresholding, which estimates a local threshold for each document image pixel, is usually capable of producing much better binarization results. One of the typical adaptive thresholding approach [2] is window based, which estimates the local threshold based on image pixels within a neighborhood window. However, the performance of the window-based methods depends heavily on the window size that cannot be determined properly without prior knowledge of the text strokes.



(a)



(b)



(c)

Figure 1: Degraded Document Images from DIBCO Datasets.

Whereas, some window-based method Nib lack's often introduces a large amount of noise and some method such as Sauvola's[3] is very sensitive to the variation of the image

contrast between the document text and the document background.

The proposed method is simple, straightforward and able to handle different types of degraded document images with minimum parameter tuning. It use of the adaptive image contrast that mixes the local image contrast and the local image gradient adaptively and therefore is liberal to the text and background variation caused by different types of degradations of document images. In particular, the proposed technique addresses the over-normalization problem of the local maximum minimum algorithm. At the same time, the parameters used in the algorithm can be adaptive estimated.

2. RELATED WORK

Many degraded documents do not have a clear bimodal pattern; global thresholding is usually not a suitable approach for the degraded document binarization. Adaptive thresholding [2], which estimates a local threshold for each document image pixel, is again a better approach to deal with different types variations in degraded document images. The early window-based adaptive thresholding [2] techniques estimate the local threshold by using the mean and the standard variation of image pixels within a local neighborhood window.

The weakness of these window-based thresholding techniques is that the thresholding performance depends deeply on the window size and hence the character stroke width. The other different approaches have also been reported, including background subtraction, texture analysis[4], recursive method [5], decomposition method, contour completion, Markov Random Field [3], cross section sequence graph analysis. These methods combine different types of image information and domain knowledge and are often complex. These methods are very useful features for segmentation of text from the document image background because the document text usually has certain image contrast to the neighboring document background. They are very effective and have been used in many document image binarization techniques.

3. PROPOSED METHOD

This section describes the proposed document image binarization techniques

A. Contrast Image Construction.

B. Canny Edge Detector.

C. Local Threshold Estimator.

D. Post Processing Procedure.

In the proposed technique, first an adaptive contrast map is constructed for an input image degraded badly. Then the binarized contrast map is combined with edge map obtained from canny edge detector to identify the pixels in edges of text stroke. By using local threshold the foreground text is further segmented which is based on the intensities of detected text stroke edge pixels within a local window. The block diagram of proposed method is as shown in figure 2.

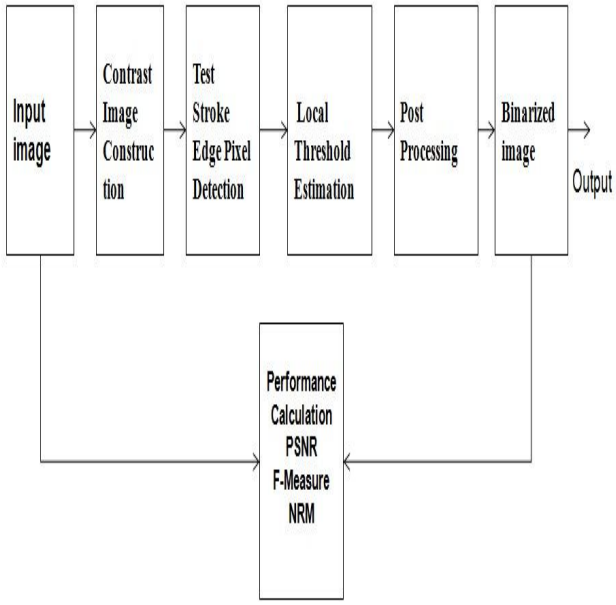


Figure 2: Block diagram of the proposed method

3.1 Contrast Image Construction

The image gradient has been extensively used for edge detection from uniform background image. Degraded document may have certain variation in input image because of patchy lighting, noise, or old age documents, bleed-through, etc. In Bernsen's paper, the local contrast is defined as follows:

$$C(i,j) = Imax(i,j) - Imin(i,j) \quad (1)$$

where C(i, j) denotes the contrast of an image pixel (i, j), *I*max*(i, j)* and *I*min*(i,)* denote the maximum and minimum intensities within a local neighborhood windows of (i, j), respectively.

If the local contrast C(i, j) is smaller than a threshold, the pixel is set as background directly. Otherwise it will be classified into text or background by comparing with the mean of *I*max*(i, j)* and *I*min*(i, j)* in Bernsen's method. The earlier proposed a novel document image binarization method [1] by using the local image contrast that is evaluated as follows

$$C(i,j) = \frac{Imax(i,j) - Imin(i,j)}{Imax(i,j) + Imin(i,j) + \epsilon} \quad (2)$$

Where ϵ is a positive but infinitely small number that is added in case the local maximum is equal to 0. By comparing with Bernsen's contrast in Equation ,and the local image contrast in Equation 2 introduces a normalization factor by extracting the stroke edges properly; the image gradient can be normalized to recompense the image variation within the document background. To restrain the background variation the local image contrast is evaluated as described in Equation 2.

In particular, the numerator (i.e. the difference between the local maximum and the local minimum) captures the local image difference that is similar to the traditional image gradient. The denominator is a normalization factor that suppresses the image variation within the document background. For pixels within bright regions of a image, it will produce a large normalization factor to neutralize the numerator and accordingly result in a relatively low image contrast. For the pixels within dark regions of an image, it will produce a small denominator and accordingly result in a relatively high image contrast.

3.2 Canny's Edge Detection

Through the contrast image construction the stroke edge pixels are detected of the document text. The edges can be detected through canny edge detection algorithm, firstly by smoothing the noise from the image and then algorithm finds for the higher magnitude of image accordingly the edges of image gradient will be marked. While marking only local edges of image should be marked.

As these methods are evaluated by the difference between the maximum and minimum intensity in a local window, the pixels at both sides of the text stroke will be selected as the high contrast pixels. The binary map can be improved further through the combination with the edges by Canny's edge detector, through the canny edge detection the text will be identified from input image.

3.3 Local Threshold Estimation

Once the text stroke edges are detected, then the document text can be extracted based on the observation that the document text is surrounded by text stroke edges and also has a lower intensity level compared with the detected stroke edge pixels[2]. The document text is extracted based on the detected text Stroke edges as follows:

$$R(x,y) = \begin{cases} 1 & I(x,y) \le E_{\text{mean}} + \frac{E_{\text{std}}}{2} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where *E*mean and *E*std are the mean and standard deviation of the intensity of the detected text stroke edge pixels within a neighborhood window *W*.

3.4 Post-Processing Procedure

Document image thresholding often introduces a certain amount of error that can be corrected through a series of post-processing operations. Document thresholding error can be corrected by three post-processing operations based on the estimated document background surface and some document domain knowledge. In particular, first remove text components (labeled through connected component analysis) of a very small size that often result from image noise such as salt and pepper noise. The real text components are usually composed based on the observation that of much more than 3 pixels, the text components that contain no more than 3 pixels in our system is simply removed.

Next, remove the falsely detected text components that have a relatively large size. The falsely detected text components of a relatively large size are identified based on the observation that they are usually much brighter than the surrounding real text strokes. Then observations are then captured by the image difference between the labeled text component and the corresponding patch within the estimated document background surface.

4. APPLICATION

Foreign language data acquired via Arabic OCR is of vital interest to military and border control applications. Various hardcopy paper types and machine- and environment-based treatments introduce artifacts in scanned images. Artifacts such as speckles, lines, faded glyphs, dark areas, shading, etc. complicate OCR and can significantly reduce the accuracy of language acquisition. For example, Sakhr Automatic Reader, a leader in Arabic OCR, performed poorly in initial tests with noisy document images. We hypothesized that performing image enhancement of bi-tonal images prior to Arabic OCR would increase the accuracy of OCR output. We also believed that increased accuracy in the OCR would directly correlate to the success of downstream machine translation.

We applied a wide variety of paper types and manual treatments to hardcopy Arabic documents. The intent was to artificially model how documents degrade in the real world. Four hardcopies of each document were created by systematically applying four levels of treatments. Subsequent scanning resulted in images that reflect the progressive damage in the life-cycle of each document – the Manually Degraded Arabic Document (MDAD) corpus. Applying the assigned image enhancement settings, three types of images were captured for each document:

- Without image enhancement,
- With Fujitsu TWAIN32 image enhancement, and
- With both Fujitsu TWAIN32 and ScanFix image enhancement.

The MDAD corpus default scans already established the images without image enhancement. The dynamic threshold capability (i.e., SDTC) was disabled in order to gain full control of the scan brightness. Discovering the ideal brightness setting involved re-scanning and reducing the brightness setting repeatedly until white pixels appeared inside glyphs. The last scan with solid black glyphs was selected as the optimal scan. The three types of images for each document were then processed through the OCR tool. CP1256 files were output and compared against the ground truth using the UMD accuracy tool.

We discovered that the evaluation metrics may not be reflecting the OCR output well. We have already mentioned that the OCR tool expects clean documents and on noisy documents it attempts to recognize speckles as characters. For noisy documents, the OCR tool produced several failure characters in the output file or caused Automatic Reader to abnormally end. Since accuracy was calculated as the number of correct characters minus error characters, divided by the number of correct characters, the tool produced negative and zero values.

(a)

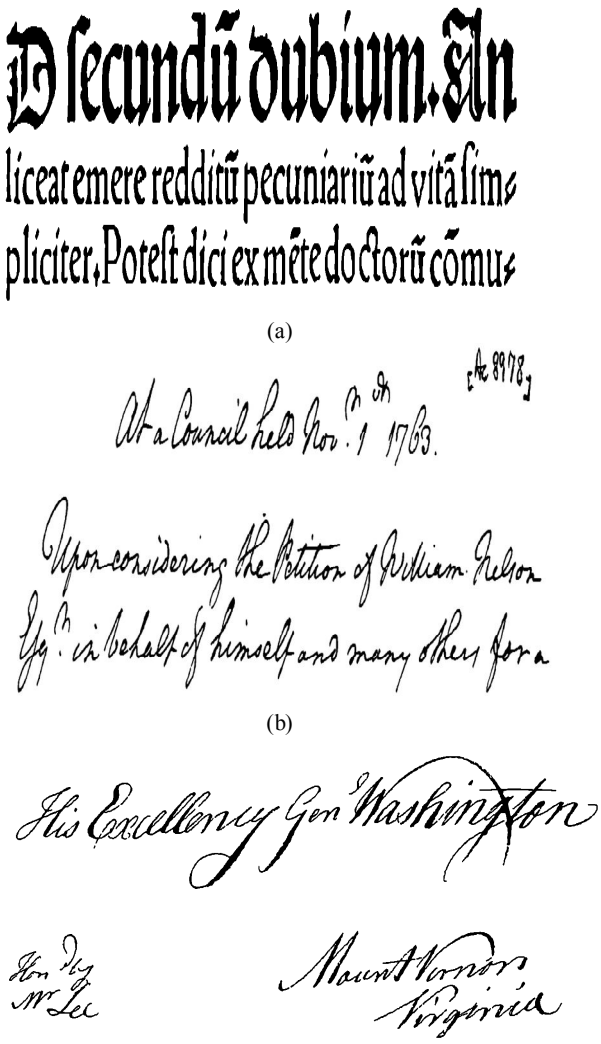
(b)

Figure 3. Text localization and recognition results of proposed binarization method.

5. DISCUSSION

As described in previous sections, the proposed method involves several parameters, most of which can be automatically estimated based on the statistics of the input document image. This makes our proposed technique more stable and easy-to-use for document images with different kinds of degradation. Binarization results of the sample document images are as shown in figure 4.

The superior performance of our proposed method can be explained by several factors. First, the proposed method combines the local image contrast and the local image gradient that help to suppress the background variation and avoid the over-normalization of document images with less variation. Second, the combination with edge map helps to produce a precise text stroke edge map. Third, the proposed method makes use of the edges of the text stroke that help to extract the foreground text from the document background accurately.



(a)

(b)

(c)

Figure 4: Binarization results of the sample document images as shown in figure 1.

6. CONCLUSION

The proposed method follows numerous different steps, Firstly pre-processing procedure collect the document image information, then proposed technique makes use of the local image contrast that is valuated based on the local maximum and minimum. Through canny edge detection the stroke edges are detected based on the local image variation, then local threshold is estimated based on the detected stroke edge pixels within a local neighborhood window and then through post processing procedure the quality of binarized result is improved.

7. REFERENCES

[1] Bolan Su, Shijian Lu, and Chew Lim Tan, ―Robust Document Image Binarization Technique for Degraded Document Images‖ IEEE TRANS ON IMAGE PROCESSING, VOL. 22, NO. 4, APRIL 2013.

[2] B. Gatos, I. Pratikakis, and S. Perantonis, "Adaptive degraded document image binarization," Pattern Recognit., vol. 39, no. 3, pp. 317–327, 2006.

[3] T. Lelore and F. Bouchara, "Document image binarisation using Markov field model," in Proc. Int. Conf. Doc. Anal. Recognit., pp. 551–555, Jul. 2009.

[4] Y. Liu and S. Srihari, "Document image binarization based on texture features," IEEE Trans. Pattern Anal. Mach. In tell., vol. 19, no. May 1997.

[5] M. Cheriet, J. N. Said, and C. Y. Suen, "A recursive thresholding technique for image segmentation," in Proc. IEEE Trans. Image Process., June 1998.