Locating Text in Historical Collection Manuscripts

Basilios Gatos[1], Ioannis Pratikakis[1,2], and Stavros J. Perantonis[1]

[1] Computational Intelligence Laboratory
Institute of Informatics and Telecommunications
National Research Center "Demokritos"
153 10 Athens, Greece
{bgat,ipratika,sper}@iit.demokritos.gr
[2] Department of Information and Communication Systems Engineering
University of the Aegean
83 200 Karlovassi, Samos, Greece
yipratik@aegean.gr

Abstract. It is common that documents belonging to historical collections are poorly preserved and are prone to degradation processes. The aim of this work is to leverage state-of-the-art techniques in digital image binarization and text identification for digitized documents allowing further content exploitation in an efficient way. A novel methodology is proposed that leads to preservation of meaningful textual information in low quality historical documents. The method has been developed in the framework of the Hellenic GSRT-funded R&D project, D-SCRIBE, which aims at developing an integrated system for digitization and processing of old Hellenic manuscripts. After testing of the proposed method on numerous low quality historical manuscripts, it has turned out that our methodology performs better compared to current state-of-the-art adaptive thresholding techniques.

1 Introduction

Historical document collections are a valuable resource for human history. It is common that documents belonging to historical collections are poorly preserved and are prone to degradation processes. This work aims to leverage state-of-the-art techniques in digital image binarization and text identification for digitized documents allowing further content exploitation in an efficient way. The method has been developed in the framework of the Hellenic GSRT-funded R&D project, D-SCRIBE, which aims to develop an integrated system for digitization and processing of old Hellenic manuscripts.

Binarization (threshold selection) is the starting step of the most document image analysis systems and refers to the conversion of the gray-scale image into a binary image. Binarization is a key step in document image processing modules since a good binarization set the base for successful segmentation and recognition of characters. In

G.A. Vouros and T. Panayiotopoulos (Eds.): SETN 2004, LNAI 3025, pp. 476–485, 2004.

old document processing, binarization usually distinguishes text areas from background areas, so it is used as a text locating technique. In the literature, the binarization is usually reported to be performed either globally or locally. The global methods (global thresholding) use one calculated threshold value to classify image pixels into object or background classes [1-5], whereas the local schemes (adaptive thresholding) can use many different adapted values selected according to the local area information [6,7]. Most of the proposed algorithms for optimum image binarization rely on statistical methods, without taking into account the special nature of document images [8-10]. However, recently some document directed binarization techniques have been developed [11-14]. Global thresholding methods are not sufficient for document image binarization since document images usually have poor quality, shadows, nonuniform illumination, low contrast, large signal-dependent noise, smear and strains. The most famous and efficient global thresholding technique is this of Otsu [2].

In this paper, a novel adaptive thresholding scheme is introduced in order to binarize low quality historical documents and locate meaningful textual information. The proposed scheme consists of four basic steps. The first step is dedicated to a denoising procedure using a low-pass Wiener filter. We use an adaptive Wiener method based on statistics estimated from a local neighborhood of each pixel. In the second step, we use Niblack's approach for a first rough estimation of foreground regions. Usually, the foreground pixels are a subset of Niblack's result since Niblack's method usually introduces extra noise. In the third step, we compute the background surface of the image by interpolating neighboring background intensities into the foreground areas that result from Niblack's method. A similar approach has been proposed for binarizing camera images [15]. In the last step, we proceed to final thresholding by combining the calculated background surface with the original image. Text areas are located if the distance of the original image with the calculated background is over a threshold. This threshold adapts to the gray-scale value of the background surface in order to preserve textual information even in very dark background areas. The proposed method has been tested with a variety of low quality historical manuscripts and has been reported to work better than the most famous adaptive thresholding techniques.

2 Previous Work

Among the most known approaches for adaptive thresholding is Niblack's method [8] and Sauvola's method [11].

Niblack's algorithm [8] calculates a pixel-wise threshold by shifting a rectangular window across the image. The threshold T for the center pixel of the window is computed using the mean m and the variance s of the gray values in the window:

$$T = m + k\,s \tag{1}$$

where k is a constant set to -0.2. The value of k is used to determine how much of the total print object boundary is taken as a part of the given object. This method can distinguish the object from the background effectively in the areas close to the objects.

The results are not very sensitive to the window size as long as the window covers at least 1-2 characters. However, noise that is present in the background remains dominant in the final binary image. Consequently, if the objects are sparse in an image, a lot of background noise will be left. Sauvola's method [11] solves this problem by adding a hypothesis on the gray values of text and background pixels (text pixels have gray values near 0 and background pixels have gray values near 255), which results in the following formula for the threshold:

$$T = m + (1 - k (1 - s/R)) \quad (2)$$

where R is the dynamics of the standard deviation fixed to 128 and k is fixed to 0.5. This method gives better results for document images.

3 Methodology

The proposed methodology for low quality historical document binarization and text preservation is illustrated in Fig. 1 and it is fully described in this section.

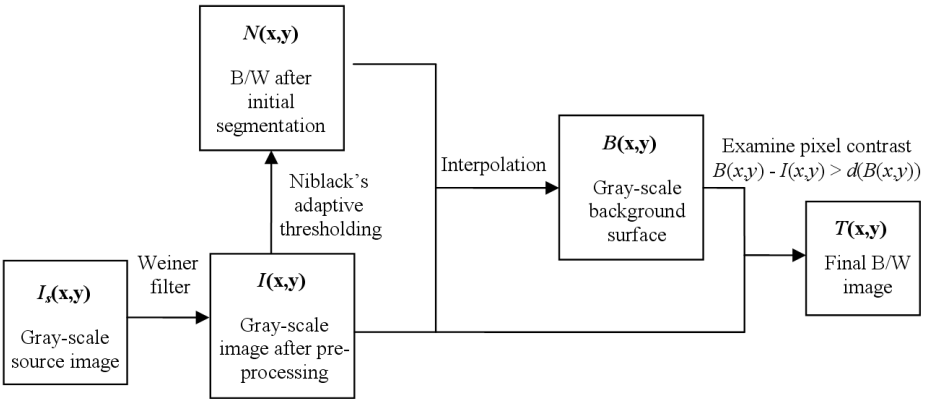


Fig. 1. Block diagram of the proposed methodology for low quality historical document text preservation.

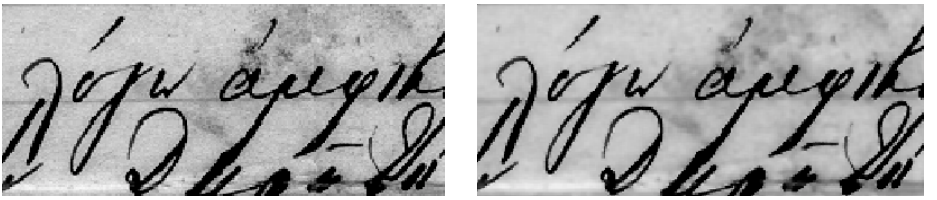
3.1 Stage1: Pre-processing

Since historical document collections are usually of very low quality, a pre-processing stage of the grayscale source image is essential in order to eliminate noise areas, to smooth the background texture and to highlight the contrast between background and text areas. The use of a low-pass Wiener filter [16] has proved efficient for the above goals. Wiener filter is commonly used in filtering theory for image restoration. Our pre-processing module implements an adaptive Wiener method based on statistics

estimated from a local neighborhood around each pixel. The grayscale source image I_s is transformed to grayscale image I according to the following formula:

$$I(x,y) = \mu + (\sigma^2 - v^2)(I_s (x,y) - \mu) / \sigma^2 \quad (3)$$

where μ is the local mean and σ^2 the variance at a NxM neighborhood around each pixel. We used a 5x5 Wiener filter for documents with thick characters or an alternative 3x3 for documents with thin characters. Fig. 2 shows the results of applying a 3x3 Wiener filter to a document image.

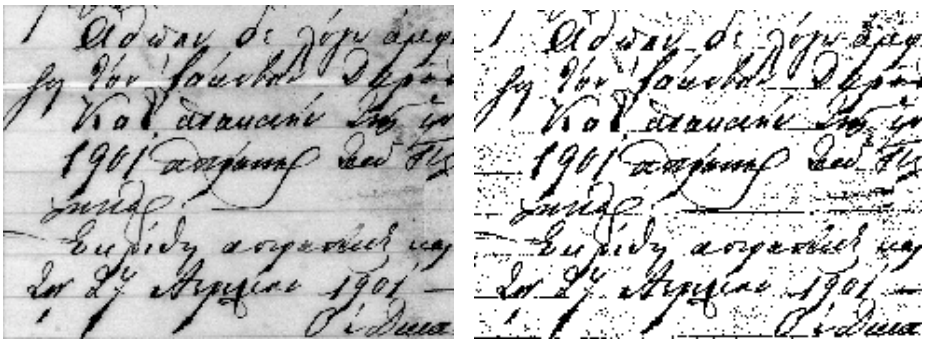


(a) (b)

Fig. 2. Image pre-processing. (a) Original image; (b) 3x3 Wiener filter.

3.2 Stage2: Rough Estimation of Foreground Regions

At this step of our approach, we want to obtain a rough estimation of foreground regions. Our intention is to proceed to an initial segmentation to foreground and background regions and find a set of foreground pixels that is a superset of the correct set of foreground pixels. In other words, we intend to obtain a set of pixels that contains the foreground pixels plus some noise. Niblack's approach for adaptive thresholding [8] is suitable for this case since Niblack's method usually detects text regions but introduces extra noise (see Fig. 3). At this step, we process image $I(x,y)$ in order to extract the binary image $N(x,y)$ that has 1's for the rough estimated foreground regions.



(a) (b)

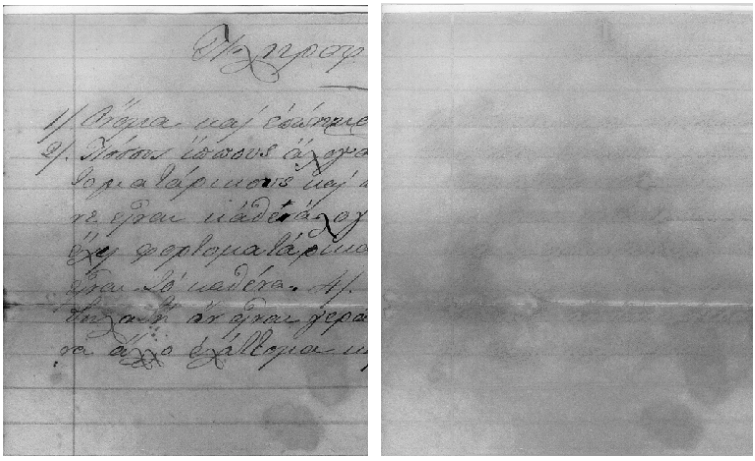
Fig. 3. Adaptive thresholding using Niblack's approach (a) Original image; (b) Estimation of foreground regions.

3.3 Stage3: Background Surface Estimation

At this stage, we compute an approximate background surface $B(x,y)$ of the image. The pixels of the pre-processed source image $I(x,y)$ belong to the background surface $B(x,y)$ only if the corresponding pixels of the resulting rough estimated foreground image $N(x,y)$ have zero values. The remaining values of surface $B(x,y)$ are interpolated from neighboring pixels. The formula for $B(x,y)$ calculation is as follows:

$$B(x,y) = \begin{cases} I(x,y) & \text{if } N(x,y) = 0 \\ \dfrac{\sum_{ix=x-dx}^{x+dx} \sum_{iy=y-dy}^{y+dy} (I(ix,iy)(1 - N(ix,iy)))}{\sum_{ix=x-dx}^{x+dx} \sum_{iy=y-dy}^{y+dy} (1 - N(ix,iy))} & \text{if } N(x,y) = 1 \end{cases} \tag{4}$$

The interpolation window of size dx x dy is defined to cover at least two image characters. An example of the background surface estimation is demonstrated in Fig. 4.



(a) (b)

Fig. 4. Background surface estimation (a) Original image I; (b) Background surface B.

3.4 Stage4: Final Thresholding

In the last step, we proceed to final thresholding by combining the calculated background surface B with the original image I. Text areas are located if the distance of the original image with the calculated background is over a threshold d. We suggest the threshold d must change according to the gray-scale value of the background surface B in order to preserve textual information even in very dark background areas. For this

reason, we propose a threshold d that has smaller values for darker regions. The final binary image T is given by the following formula:

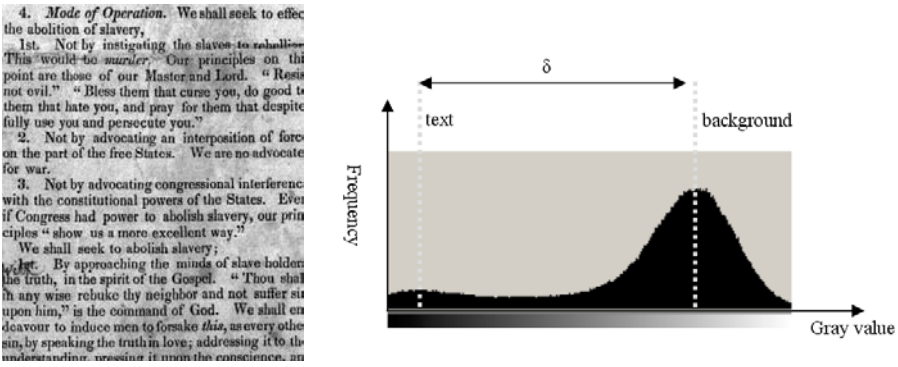
$$T(x,y) = \begin{cases} 1, & \text{if } B(x,y) - I(x,y) > d(B(x,y)) \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

A typical histogram of a document image (see Fig. 5) has two peaks, one for text regions and one for background regions. The average distance δ between the foreground and background can be calculated by the following formula:

$$\delta = \frac{\sum_x \sum_y (B(x,y) - I(x,y))}{\sum_x \sum_y N(x,y)} \tag{6}$$

For the usual case of document images with uniform illumination, the minimum threshold d between text pixels and background pixels can be defined with success as $q \cdot \delta$ where q is a variable near 0.8 that helps preserving the total character body in order to have successful OCR results [15]. In the case of very old documents with low quality and non-uniform illumination we want to have a smaller value for the threshold d for the case of darker regions since it is a usual case to have text in dark regions with small foreground-background distance. To achieve this, we first compute the average background values b of the background surface B that correspond to the text areas of image N:

$$b = \frac{\sum_x \sum_y (B(x,y)(1 - N(x,y)))}{\sum_x \sum_y (1 - N(x,y))} \tag{7}$$



(a) (b)

Fig. 5. Document image histogram (a) Original image; (b) Gray level histogram.

We wish the threshold to be approximately equal to the value $q \cdot \delta$ when the background is large (roughly greater than the average background value b) and approximately equal to $p_2 \cdot q \cdot \delta$ when the background is small (roughly less than $p_1 \cdot b$) with $p_1, p_2 \in [0,1]$. To simulate this desired behaviour, we use the following logistic sigmoid function that exhibits the desired saturation behaviour for large and small values of the background as shown in Fig. 6:

$$d(B(x,y)) = q\,\delta \left(\frac{(1-p_2)}{1+\exp\left(\frac{-4\,B(x,y)}{b\,(1-p_1)} + \frac{2(1+p_1)}{(1-p_1)}\right)} + p_2 \right) \tag{8}$$

After experimental work, for the case of old manuscripts, we suggest the following parameter values: $q = 0.6$, $p_1 = 0.5$, $p_2 = 0.8$.

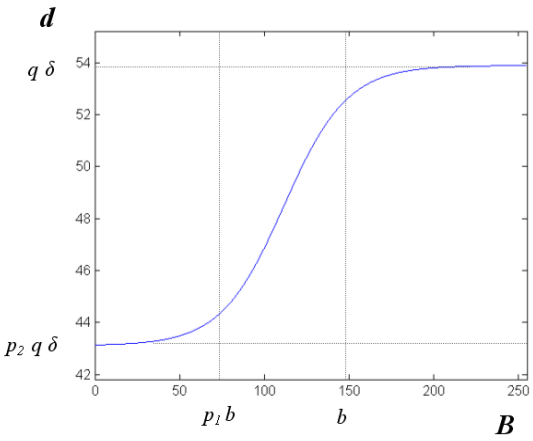


Fig. 6. Function $d(B(x,y))$.

4 Experimental Results

In this section, we compare the performance of our algorithm with those of Otsu [2], of Niblack [8] and of Sauvola *et al.* [11]. We also give the results of the application of a global thresholding value. The testing set includes images from handwritten or typed historical manuscripts. All images are of poor quality and have shadows, non-uniform illumination, smear and strain. Fig. 7 demonstrates an example of a typed manuscript, while Fig. 8 demonstrates an example of handwritten characters. As shown in all cases, our algorithm out-performs all the rest of the algorithms in preservation of meaningful textual information.

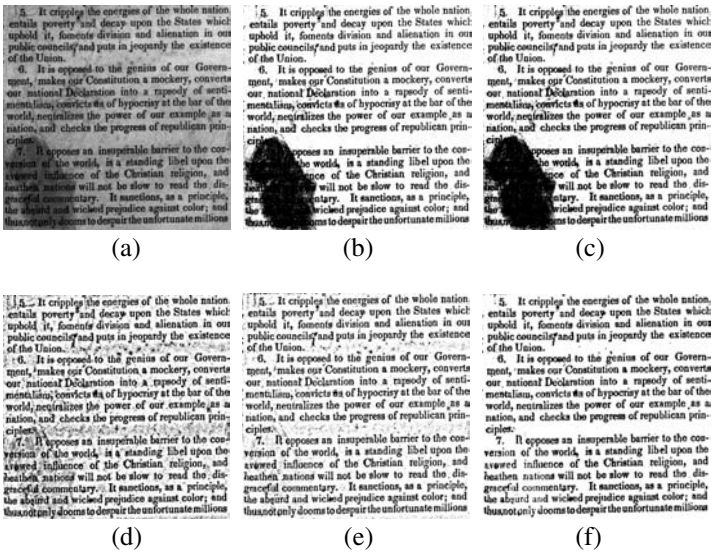


Fig. 7. Experimental results - binarization of a typed manuscript (a) Original image; (b) Global thresholding; (c) Otsu's method; (d) Niblack's method; (e) Sauvola's method; (f) The proposed method.

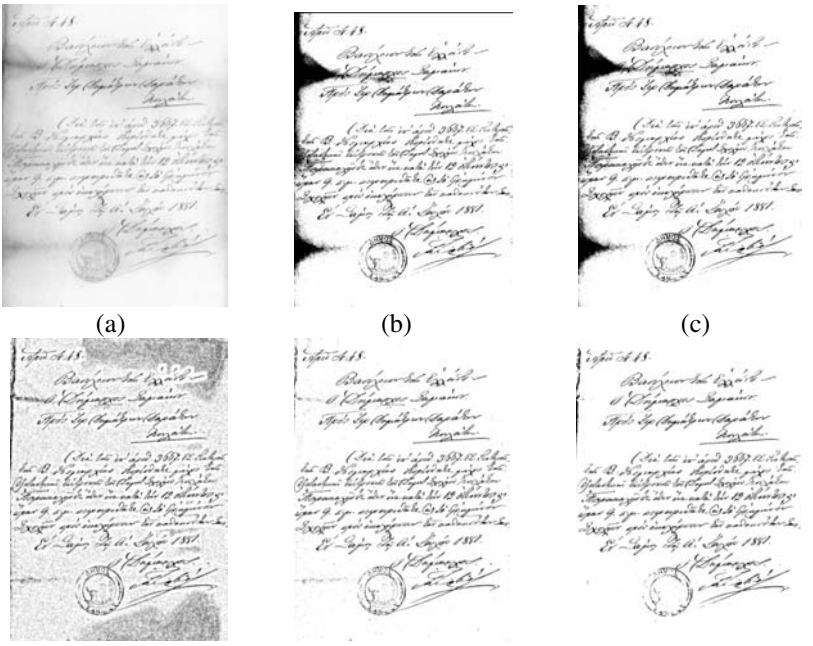


Fig. 8. Experimental results - binarization of a handwritten manuscript (a) Original image; (b) Global thresholding; (c) Otsu's method; (d) Niblack's method; (e) Sauvola's method; (f) The proposed method.

To quantify the efficiency of the proposed binarization method, we compared the results obtained by the well-known OCR engine FineReader 6 [17], using the binarization results of Niblack [8], Sauvola *et al.* [11] and the proposed method, as the inputs of this engine. To measure the quality of the OCR results we calculated the Levenshtein distance [18] between the correct text (ground truth) and the resulting text. As shown at Table 1, the application of the proposed binarization technique has shown the best performance regarding the final OCR results.

Table 1. Representative OCR results after applying several binarization schemes.

	Levenshtein Distance from the Ground truth			
	Document 1	Document 2	Document 3	Document 4
Niblack's method	228	619	513	447
Sauvola's method	60	394	276	694
The proposed method	**56**	**207**	**177**	**153**

5 Conclusions

In this paper, we present a novel methodology that leads to preservation of meaningful textual information in low quality historical documents. The proposed scheme consists of four (4) distinct steps: a pre-processing procedure using a low-pass Wiener filter, a rough estimation of foreground regions using Niblack's approach, a background surface calculation by interpolating neighboring background intensities and finally a thresholding by combining the calculated background surface with the original image. Text areas are located if the distance of the original image with the calculated background is over a threshold. This threshold adapts to the gray-scale value of the background surface in order to preserve textual information even in very dark background areas. The proposed methodology works with great success even in cases of historical manuscripts with poor quality, shadows, nonuniform illumination, low contrast, large signal-dependent noise, smear and strain. Experimental results show that our algorithm out-performs the most known thresholding approaches.

Acknowledgment

The authors would like to thank Dr. Ergina Kavallieratou for providing us samples from her private historical document collection.

References

1. Rosenfeld, A., Kak, A. C.: Digital Picture Processing, 2nd edition, Academic Press, New York (1982).
2. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. Systems Man Cybernet. 9 (1) (1979) 62-66.

3. Kittler, J., Illingworth, J.: On threshold selection using clustering criteria. IEEE Trans. Systems Man Cybernet. 15 (1985) 652-55.
4. Brink, A. D.: Thresholding of digital images using two-dimensional entropies. Pattern Recognition 25 (8) (1992) 803-808.
5. Yan, H.: Unified formulation of a class of image thresholding techniques. Pattern Recognition 29(12) (1996) 2025-2032.
6. Sahoo, P. K., Soltani, S., Wong, A. K. C.: A survey of Thresholding Techniques. Computer Vision, Graphics and Image Processing, 41(2) (1988) 233-260.
7. Kim, I. K., Park, R. H.: Local adaptive thresholding based on a water flow model. Second Japan-Korea Joint Workshop on Computer Vision, Japan (1996) 21-27.
8. Niblack, W.: An Introduction to Digital Image Processing. Englewood Cliffs, N. J., Prentice Hall (1986) 115-116.
9. Yang, J., Chen, Y., Hsu, W.: Adaptive thresholding algorithm and its hardware implementation. Pattern Recognition Lett. 15(2) (1994) 141-150.
10. Parker, J. R., Jennings, C., Salkauskas, A. G.: Thresholding using an illumination model. ICDAR'93 (1993) 270-273.
11. Sauvola, J., Seppanen, T., Haapakoski, S., Pietikainen, M.: Adaptive Document Binarization. International Conference on Document Analysis and Recognition (1997) 147-152.
12. Chang, M., Kang, S.,Rho, W., Kim, H., Kim, D.: Improved binarization algorithm for document image by histogram and edge detection. ICDAR'95 (1993) 636-643.
13. Trier, O. D., Jain, A. K.: Goal-Directed Evaluation of Binarization Methods. IEEE Trans. on Patt. Anal. and Mach. Intell., 17(12) (1995) 1191-1201.
14. Eikvil, L., Taxt, T., Moen, K.: A fast adaptive method for binarization of document images. Int. Conf. Document Analysis and Recognition, France (1991) 435-443.
15. Seeger, M., Dance, C.: Binarising Camera Images for OCR. Sixth International Conference on Document Analysis and Recognition (ICDAR'01), Seattle, Washington (2001) 54-58.
16. Jain, A.: Fundamentals of Digital Image Processing. Prantice Hall, Englewood Cliffs, NJ (1989).
17. www.finereader.com
18. Levenshtein, V. I.: Binary codes capable of correcting deletions, insertions and reversals. Sov. Phys. Dokl., 6 (1966) 707-710.