

Comparison of Niblack inspired Binarization methods for ancient documents

Khurram Khurshid¹, Imran Siddiqi¹, Claudie Faure², Nicole Vincent¹

¹Laboratoire CRIP5 – SIP, Université Paris Descartes

45, rue des Saints-Pères, 75006 Paris France

{khurshid ; siddiqi ; nicole.vincent} @ math-info.univ-paris5.fr

²UMR CNRS 5141 - GET ENST, 46 rue Barrault, 75634 Paris

cfaure@enst.fr

ABSTRACT

In this paper, we present a new sliding window based local thresholding technique ‘NICK’ and give a detailed comparison of some existing sliding-window based thresholding algorithms with our method. The proposed method aims at achieving better binarization results, specifically, for ancient document images. NICK has been inspired from the Niblack’s binarization method and exhibits its robustness and effectiveness when evaluated on low quality ancient document images.

Keywords: Ancient documents, local binarization, Niblack

1. INTRODUCTION

The importance of digital libraries for information retrieval cannot be denied. The ancient historical books contain invaluable knowledge but it is very time consuming to search the required information in these paper books. Different methods have been devised to facilitate this information search. These include word spotting, optical character recognition etc. Although a lot of work has already been done in this field, it still remains an inviting and challenging field of research, mainly because the results achieved so far are not satisfactory for huge volumes of data; especially if the document base consists of a set of ancient printed documents of relatively degraded quality [9, 10, 11].

In an OCR, one of the main processing stage is binarization of document images, i.e. separation of foreground from background [12,16]. Binarization of a text image should give us, in an ideal case, the foreground text in black and noisy background in white. Though different thresholding methods already exist in literature, they don’t give perfect results for all types of documents. Some algorithms might work better for one type of documents where there are marks of strain while they might give poor results for other types where there are extremely low intensity variations.

Different binarization methods have been evaluated in [1, 2, 8] for different types of documents. [2] presents an evaluation of eleven locally adaptive binarization methods for gray scale images with low contrast, variable background intensity and noise. In that evaluation, Niblack’s method [4] was found to be the better of them all. Different improvements have since been made to the original Niblack method to improve the results. These include Sauvola’s algorithm [5], Wolf’s work [6] and Feng’s method [7]. Our method also introduces an improvement in the Niblack algorithm with an advantage that it improves binarization for "white" and light page images by shifting down the binarization threshold. We have named our algorithm ‘NICK’ after the first letters of the names of this paper’s authors. There are other more complicated multi-step thresholding processes like [8,16], but we will limit our discussion to the simple Niblack inspired sliding window thresholding methods.

Paper is divided into multiple sections beginning with the description of different sliding-window based thresholding methods. This is followed by the detailed description of our proposed method. At the end we show the comparison of these different methods and the results achieved. Resulting binarization images by different algorithms are also shown in the experimental results section.

2. RELATED WORK

The problem of document binarization is as old as the document image analysis itself. A large number of binarization techniques have been proposed over the last two decades. These techniques can generally be classified into two categories i.e. global thresholding and local thresholding. Global thresholding methods employ a single intensity threshold value. This value is calculated based on some heuristics or statistics of some global image attributes to classify image pixels into foreground (text) or background (non-text) pixels [3]. The main issue with global methods is that they can not adapt well to uneven illumination and noise, hence do not perform well on low quality document images.

Local thresholding methods, on the other hand, compute a threshold for each pixel in the image on the basis of the content in its neighborhood. As opposed to global thresholding, local methods generally perform better for low quality images, which are the ones that we will focus on. In the following section, we give an account of some of the well known sliding window binarization methods. Our objective, however, is not to provide a survey of the existing methods as found elsewhere [1, 2]. We will limit our discussion only to the simple sliding-window thresholding methods, highlighting their pros and cons and finally suggesting our improved method.

2.1 Niblack Algorithm

Niblack's algorithm [4] calculates a pixel-wise threshold by sliding a rectangular window over the gray level image. The computation of threshold is based on the local mean m and the standard deviation s of all the pixels in the window and is given by the equation 1 below:

$$\begin{aligned} T_{Niblack} &= m + k * s \\ T_{Niblack} &= m + k \sqrt{\frac{1}{NP} \sum (p_i - m)^2} \\ &= m + k \sqrt{\frac{\sum p_i^2}{NP} - m^2} = m + k \sqrt{B} \end{aligned} \quad (1)$$

where NP is the number of pixels in the gray image, m is the average value of the pixels p_i , and k is fixed to -0.2 by the authors. Advantage of Niblack is that it always identifies the text regions correctly as foreground but on the other hand tends to produce a large amount of binarization noise in non-text regions also.

2.2 Sauvola's Algorithm

Sauvola's algorithm [5] claims to improve Niblack's method by computing the threshold using the dynamic range of image gray-value standard deviation, R :

$$T_{Sauvola} = m * (1 - k * (1 - \frac{S}{R}))$$

where k is set to 0.5 and R to 128. This method outperforms Niblack's algorithm in images where the text pixels have near 0 gray-value and the background pixels have near 255 gray-values. However, in images where the gray values of text and non-text pixels are close to each other, the results degrade significantly.

2.3 Wolf's Algorithm

To address the issues in Sauvola's algorithm, Wolf et al. [WOF 03] propose to normalize the contrast and the mean gray value of the image and compute the threshold as:

$$T_{Wolf} = (1 - k) * m + k * M + k * \frac{S}{R} (m - M)$$

where k is fixed to 0.5, M is the minimum gray value of the image and R is set to the maximum gray-value standard deviation obtained over all the local neighborhoods (windows).

This method in most cases outperforms its predecessors. However, degradation is observed in its performance if there is a sharp change in background gray values across the image. This is due to the fact that the values of M and R are calculated from the whole image. So even a small noisy patch could significantly influence M and R values, thus eventually calculating misleading binarization thresholds.

2.4 Feng's Algorithm

Instead of calculating dynamic range of gray-value standard deviation from the whole image like [6], Feng et al. [7] propose calculating it locally introducing the notion of two local windows, one contained within the other. The values of local mean m , the minimum gray-level M , and standard deviation s are calculated in the primary local window while the dynamic range standard deviation R_s is calculated in the larger window termed as 'secondary local window'. Binarization threshold is then computed as:

$$T_{Feng} = (1 - \alpha_1) * m + \alpha_2 * \left(\frac{s}{R_s} \right) * (m - M) + \alpha_3 * M$$

where $\alpha_2 = k_1 (s/R_s)^\gamma$ and $\alpha_3 = k_2 (s/R_s)^\gamma$. Based on the experimental experiences of authors, γ is set to 2 while the values of other parameters, α_1 , k_1 and k_2 are proposed to be in the ranges 0.1-0.2, 0.15-0.25 and 0.01-0.05 respectively. This method addresses well the R-problem in the Wolf's algorithm. However the introduction of three parameters, determined empirically, leaves the robustness of this method questionable.

3. OUR APPROACH

We now put forward our proposition of calculating the binarization threshold which is likely to work better for many (if not all) types of degraded and noisy ancient documents. Instead of following the chain of one algorithm proposing modifications in its predecessors, we derive our thresholding formula from the basic Niblack algorithm, the parent of all the methods discussed earlier. The major advantage our method achieves over Niblack is that it considerably improves binarization for "white" and light page images by shifting down the binarization threshold.

The thresholding that we propose is:

$$\begin{aligned} T &= m + k \sqrt{(\sum p_i^2 - m^2) / NP} \\ &= m + k \sqrt{A} \end{aligned} \quad (2)$$

where ,

k is the Niblack factor
 m = mean grey value
 p_i = pixel value of grey scale image
 NP = number of pixels

Now using equation (1) and (2) , we find a relation between A and B:

$$\begin{aligned} (A) - (B) &= \frac{\sum p_i^2}{NP} - \frac{m^2}{NP} - \frac{\sum p_i^2}{NP} + m^2 = \\ m^2 - \frac{m^2}{NP} &= m^2 \left(1 - \frac{1}{NP}\right) = m^2 \left(\frac{NP-1}{NP}\right) \end{aligned}$$

If the value of NP is high, we can approximate $(A - B)$ by m^2 .

$$A - B \cong m^2$$

$$\text{thus} \quad A \cong B + m^2$$

It shows that if the image is very dark, the value of m is low meaning that the difference between A and B is very small. But if the image is light, the value of m is high and thus the difference between A and B is greater which lowers the binarization threshold for NICK as shown in figure 1.

$$T \cong m + k\sqrt{B + m^2} \dots\dots\dots (3)$$

The value of Niblack factor k can vary from -0.1 to -0.2 depending upon the application requirement. k close to -0.2 make sure that noise is all but eliminated but characters can break a little bit, while with values close to -0.1, some noise pixels can be left but the text will be extracted crisply and unbroken. So, for an OCR application, the value of k must be set at -0.1 and in applications where we don't desire any noise, k should be -0.2.

Equation (3) shows that the difference between values of T and $T_{Niblack}$ increases with m . To highlight this difference, we applied NICK and Niblack's method, globally (computing the binarization threshold using mean and standard deviation of complete image), to a set of 20 document images having different mean gray values. Fig. 1 shows the relation between T and $T_{Niblack}$ for these images. For whiter images (m near 255) this difference becomes more significant. Now coefficient k being negative, T becomes smaller than $T_{Niblack}$, thus implying that less pixels are coded as black pixels. Fig. 2 shows the result obtained with *with Niblack and NICK* at $k = -0.2$ on the first three pages of a book.

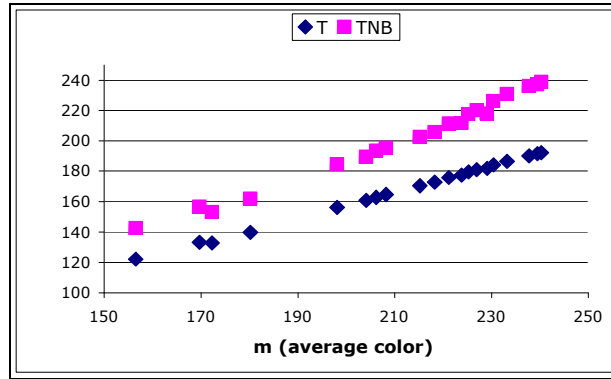
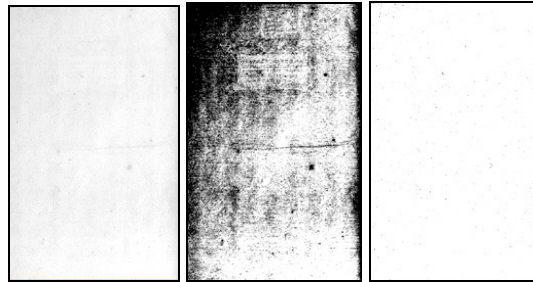


Fig. 1: T and TNB thresholds for different average gray levels m .



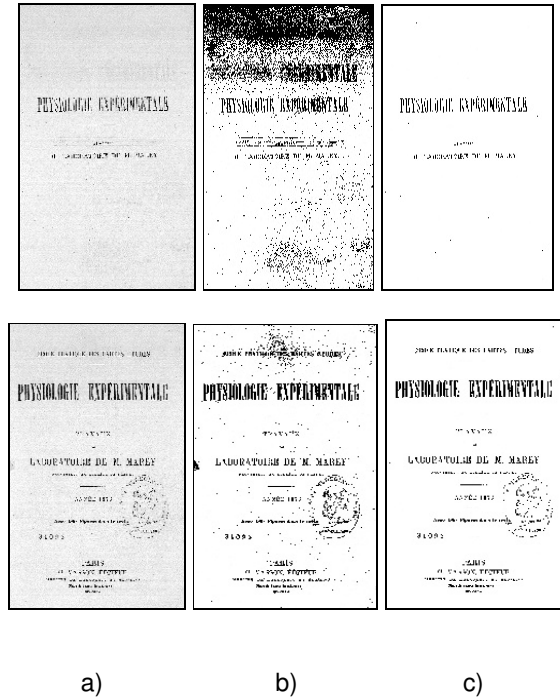


Fig. 2: a) Original page. b) Binarization with Niblack. c) Binarization with NICK

To evaluate our method and compare the results with the local methods described above, we binarize the document images locally, computing the binarization threshold for each pixel of the input image using a sliding window. The window size (optimized empirically) has been set to 19x19 in our case. The results are given in the following section.

4. EXPERIMENTAL RESULTS

The binarization results formulated have been based on tests performed on images provided by the Bibliothèque Interuniversitaire de Médecine, Paris [13] and the Institute de recherche et d'histoire des textes [14]. A total of 120 images, of size 1536 x 2549 pixels for BIUM and miscellaneous sizes for IRHT images, were selected from the base and were binarized using Niblack's, Sauvola's, Wolf's and Feng's methods. The results of these four methods were compared with the results achieved by our method. All these methods have been evaluated for a window size of 19x19, with bigger window in Feng's method being kept at 33x33. Some of the results obtained by these methods are shown in table 1. Based on visual criteria, the proposed algorithm seems to outperform the other methods with respect to image quality and preservation of meaningful textual information. After a thorough visual examination of the experimental results, important observations are summarized in the following:

- With Niblack's approach, the resulting binary image generally suffers from a large amount of background noise, especially in areas without text.
- With Sauvola's method, the background noise problem that appears in Niblack's approach is solved but in many cases where there are less intensity variations, characters become extremely thinned and broken. In some cases, the characters disappear totally giving a white output image.

- In most of the cases Wolf's algorithm outperforms the two predecessors, however there are occasions when the characters disappear or break if intensity variations are very small or there is some noisy patch with a very sharp intensity variation from the rest of the image.
- Feng's method generally works very well but the main drawback remains its susceptibility to the empirically determined parameter values as discussed earlier. A slight change in parameter values could drastically affect the binarization results, as was observed in our experiments. One set of parameter values could give excellent results for one image but the same set would not work for another image with different intensity and illumination variations. Secondly, the introduction of a larger secondary window (around the primary window) also makes this method computationally inefficient as compared to the rest.
- NICK shows improved performance when compared to the other methods tested, and it performs better specially when the images have extremely low intensity variations and for whiter images. Also computationally, NICK is much more efficient as compared to Wolf's and Feng's method. Table 1 shows portions of some output images.

As in [7], we did a threshold analysis to see the difference in local threshold values between these methods. Figure 3a shows the thresholds determined by the five methods along a scanline (central line) of a manually created synthetic image having width 140 pixels and height 19 pixels (height was set to 19 because of the window size used). This image was generated in a way that background columns (lighter gray values) are randomly put around the foreground columns (darker gray values) to simulate a degraded text image. It enables us to see how each of the methods respond to the changing background conditions in an image.

This analysis shows the effectiveness of our method in correctly discriminating the foreground and background regions. In fig 3a, our method successfully classified the right portion of the scanline (from pixel 80 onwards) as background as opposed to Niblack's and Sauvola's methods while Wolf's and Feng's algorithm couldn't identify the foreground that crisply in the left portion of scanline.

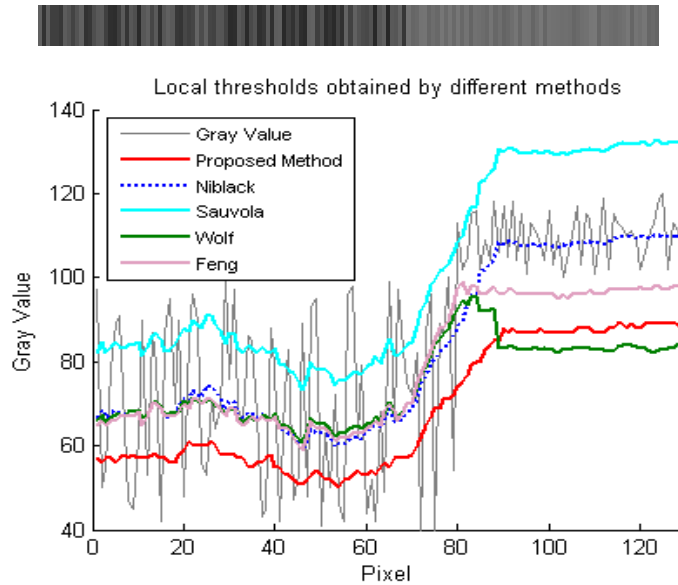
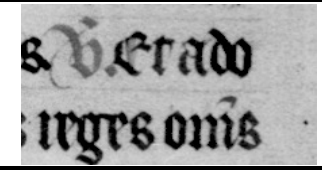

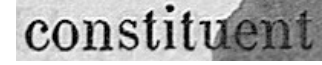



Fig. 3: Local thresholds obtained by the five methods.

Tab. 1: Portion of output images obtained by using a) Niblack b) Sauvola c) Wolf d) Feng e) NICK at windows size 19

Portion of original Image	a	b
c	d	e
		
		
		
		
		
		
		
		
		
		

To quantify the efficiency of our proposed binarization method, an OCR experiment was also performed on text samples having different intensity variations and quality. Four of the samples are shown in figure 4. These samples were selected from 25 representative degraded BIUM images. The character recognition was performed by the well-known OCR

engine ABBYY Fine Reader 9 [15] on the binarization results of Niblack, Sauvola, Wolf, Feng and NICK on each of the 25 images. OCR results are quantified by character recognition rate, calculated as:

$$\text{recognition rate} = \frac{\text{Number of Correctly Detected Characters}}{\text{Total Characters in Ground Truth}}$$

The recall percentages for each of the images are given in table 2.

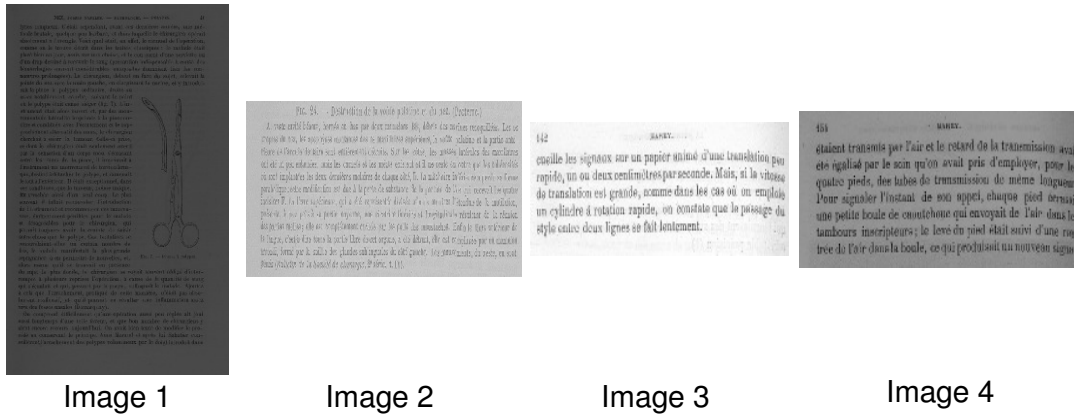


Fig. 4: Four sample document images of the 25 used in the OCR test

Tab. 2: Recall rates achieved by the ABBYY OCR

Image	Total Characters	Number of characters correctly recognized				
		Niblack	Sauvola	Wolf	Feng	NICK
Image 1	2012	0	1907	2003	1932	2002
Image 2	944	887	0	821	938	940
Image 3	241	239	233	233	230	238
Image 4	364	0	356	354	344	363
Other 21 images	5446	2687	3783	4828	5364	5411
Total	9007	3813	6279	8239	8808	8954
RECG RATE		42.33	69.71	91.47	97.79	99.41

It can be observed from the analysis of first four images that ABBYY was not able to read characters from Image1 & Image4 binarized with Niblack's method, and Image 2 binarized with Sauvola's method as it classified them as figures due to bad binarization. With Niblack, it was the case of extreme noise in the binary image which prevented the OCR from detecting the text, while with Sauvola, it was due to very thin and broken characters, which as a matter of fact were impossible to read from naked eye as well. With Wolf's method, binarization was better in most cases. In image 2 though, there were a lot of broken characters which the OCR could not read. With Feng's formula, OCR results were better than Wolf's method, though in Feng's images, there were bit more noisy patches in addition which affected the OCR performance in some images. With NICK, ABBYY was always able to distinguish and read text indicating the effectiveness and robustness of our method for different illumination and intensity variations in document images. It can be noted that though our method does not always correspond to the best OCR performance for each image, it has no

strong weakness. Thus in each case it is able to provide a binary image that the OCR reads as text even if the original image quality was very poor. Based on this test set, NICK achieves a recall rate of 99.41% compared to 97.79% by Feng, and that too at a lower computational cost, thus showing that NICK outperforms the other sliding-window thresholding methods in efficiency and robustness for low quality ancient document images.

5. CONCLUSION

We have proposed a new binarization method for low quality ancient document. It performs better than the contemporary sliding window methods specially when the input image has very little or no text (white image) and also when the intensity variations between the text and background are extremely low. The results obtained serve to strengthen the arguments put forward in this paper. An exhaustive evaluation of the OCR results on a much varied document set, however, rests to be done which will be the subject of the work to follow.

REFERENCES

- [1] Graham Leedham, Chen Yan, Kalyan Takru, Joie Hadi Nata Tan, Li Mian, "Comparison of Some Thresholding Algorithms for Text/Background Segmentation in Difficult Document Images" *7th International Conference on Document Analysis and Recognition ICDAR*, (2003).
- [2] Øivind Due Trier, Torfinn Taxt, "Evaluation of Binarization Methods for Document Images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17: 312 – 315, (1995).
- [3] N. Otsu, "A threshold selection method from grey level histogram", *IEEE Trans. Syst. Man Cybern.*, vol. 9 no. 1, pp. 62-66 (1979).
- [4] W.Niblack, *An Introduction to Digital Image Processing*. Prentice Hall, Englewood Cliffs, (1986).
- [5] J.Sauvola, T.Seppanen, S.Haapakoski, M.Pietikainen, "Adaptive Document Binarization", 4th Int. Conf. On Document Analysis and Recognition, Ulm, Germany, pp.147-152 (1997).
- [6] C. Wolf, J-M. Jolion, "Extraction and Recognition of Artificial Text in Multimedia Documents", *Pattern Analysis and Applications*, 6(4):309-326, (2003).
- [7] Meng-Ling Feng and Yap-Peng Tan, "Contrast adaptive binarization of low quality document images", *IEICE Electron. Express*, Vol. 1, No. 16, pp.501-506, (2004).
- [8] B. Gatos, I. Pratikakis, S.J. Perantonis, "Adaptive degraded document image binarization", *Pattern Recognition* 39 : 317 – 327, (2006).
- [9] P. J. Burt, C. Yen, X. Xu, "Recovery of distorted document images from bound volumes", *Sixth International Conference on Document Analysis and Recognition*, pp. 429-233 (2001).
- [10] A.Antonacopoulos, Karatzas D., Krawczyk H. and Wiszniewski B., "The Lifecycle of a Digital Historical Document: Structure and Content", *ACM Symposium on Document Engineering*, 147 –154 (2004).
- [11] Baird H. S., "Difficult and urgent open problems in document image analysis for libraries", *1st International workshop on Document Image Analysis for Libraries*, (2004).
- [12] K. Khurshid, C. Faure, N. Vincent, "Feature based word spotting for ancient printed documents", *Eight international workshop on pattern recognition in information systems*, Spain (2008).
- [13] Digital Library of BIUM (Bibliothèque Interuniversitaire de Médecine, Paris), <http://www.bium.univ-paris5.fr/histmed/medica.htm>
- [14] Base de l'institute de recherche et d'histoire des textes, <http://www.irht.cnrs.fr>
- [15] ABBYY fine reader version 9.0, www.finereader.com
- [16] Mauritius Seeger, Christopher Dance, "Binarising Camera Images for OCR", *ICDAR*, 2001