Enhancement of Historical Document Images by Combining Global and Local Binarization Technique

E. Zemouri, Y. Chibani, and Y. Brik

***Abstract*—In this paper we present a combined binarization technique for historical document images. Usually, many binarization techniques are implemented in the literature for different types of binarization problems. The few simple available thresholding methods cannot be applied to many binarization problems. In order to improve the quality of historical document images, we propose a combined approach based on global and local thresholding methods. The method was evaluated on the benchmarking dataset used in the Handwritten Document Image Binarization Contest (H-DIBCO 2012) and an Arabic historical document from National Library of Algeria. The evaluation based on the word spotting system showed the efficiently of our approach.**

***Index Terms*—Historical document, binarization, global and local threshold, word spotting.**

I. Introduction

Binarization is an important step in historical document image preprocessing to eliminate background noise and improve the document quality. This process consists of converting the gray-level image in binary image which can be used for further processing (Optical character recognition 'OCR', Intelligent character recognition 'ICR', Word spotting…).

Many thresholding algorithms have been previously proposed. However, the quality of these algorithms still shows quality shortcoming in document image analysis systems. An early histogram-based global binarization algorithm, Otsu's method [1], is widely used. Isodata's method [2] also used as a global method to calculate the explicit thresholds. Niblack [3], Sauvola [4] and NICK [5] use a local thresholding. It has been proved that all previously reported methods are effective for certain types of document images. However, none has been proved to be effective for all examples of degraded document images. Historical document images are particularly challenging for the thresholding or information separation problem (Document Image Binarization Contest: DIBCO 2009 [6], H-DIBCO 2010 [7], DIBCO 2011 [8] and H-DIBCO 2012 [9]). Many historical documents have become degraded and are difficult for a human to decipher due to long ineffective storage conditions and inevitable differences in paper quality and ink.

In order to improve the quality of binarized image, we propose to enhance them before the binarization. The proposed method makes use of the global thresholding to enhance the document image, and then we apply a local thresholding method.

This paper is structured as follow; Section II reviews the state of the art of binarization techniques. Our proposed method is presented in Section III. Then, experimental results are reported in Section IV. Finally, conclusion and future work are presented in Section V.

II. State of the Art

Thresholding historical document image converts the gray-level image to binary format by separating the useful font and information from the background. There are two main approaches of binarization namely global and local thresholding. In the global method, only one threshold is used in the whole image, if the pixel value of an input image is more than T, the pixel is set to background. Otherwise, it is foreground.

Otsu's method [1] assumes the presence of two distributions (one for the text and another one for the background). It calculates a threshold value in such a way that it maximizes the variance between the two distributions.

Isodata's method [2] calculates a threshold by separating iteratively the gray-level histogram into two classes.

The main drawback of global methods is that they can't adapt well to uneven illumination and noise. Hence, they do not perform well on low quality document images.

Unlike global thresholding, local threshold is calculated for each pixel in the image according to the properties of its neighborhood. This method generally performs better for low quality images.

Niblack's method [3] calculates the thresholding values of each window over the image separately by the following formula:

$$T = m + k.s \quad (1)$$

where m is the mean value and s is the standard deviation value of the pixels inside the window. The value of k is generally fixed to -0.2.

Sauvola's method [4] was developed from Niblack's method. It aims to solve the problem of black noise depending on the impact on the standard deviation value by using a range of gray-level values in the images. The thresholding formula is:

Manuscript received September 20, 2013; revised November 21, 2013.

ET-Tahir Zemouri is with the Speech Communication and Signal Processing Laboratory, University of Sciences and Technology Houari Boumediene, Algiers, Algeria (e-mail: tzemouri@ usthb.dz).

Youcef Chibani and Youcef Brik are with the Speech Communication and Signal Processing Laboratory, University of Sciences and Technology Houari Boumediene, Algiers, Algeria (e-mail: ychibani@ usthb.dz, ybrik@ usthb.dz).

DOI: 10.7763/IJIEE.2014.V4.397

$$T = m.\left[1 + k.(s / R - 1)\right] \quad (2)$$

where R is the dynamic range of standard deviation and the parameter k gets positive values (k = 0.5, R = 128).

The NICK method [5] is an improvement of Niblack's method developed by Khurshid. It aims to solve the problem of black noise in Niblack's method and the low contrast problem in Sauvola's method by shifting the thresholding value downward. The thresholding formula is:

$$T = m + k.\sqrt{\frac{\sum P_i^2 - m}{N}} \quad (3)$$

where k is a factor in the range [-0.2, -0.1], P_i is the gray-level value of the pixel, and N is the total number of pixels.

Generally, global thresholding methods are not able to remove noise that is not distributed in the image. Furthermore, local thresholding approaches are significantly more time-consuming and computationally expensive.

III. Proposed Method

The proposed hybrid approach includes both global and local thresholding techniques to deal with noisy historical documents, where firstly a global thresholding T is applied to whole document.

The algorithm is based on the fact that a document image includes very rare pixels of useful information (foreground) compared to the size of the image (foreground and background). The fact is that rarely, the amount of black pixels exceeds the 10% of the total pixels in the document. Taking advantage of this fact, we accept that the average of the pixel values of a document image is determined mainly by the background even if the document is quite clear. This claim is demonstrated in Fig. 1. In the same figure two thresholds (one of Otsu's method and the second of Isodata's method) and the average value in each case are given. It is obvious that the average value is always on the background side, considering either threshold.

The proposed algorithm of document binarization consists of the following steps:

Algorithm

i) Apply global thresholding (T) to the whole document.

where T = *average*(I),

if $I(x,y) > T$: $I_{int}(x,y) = 255$

else $I_{int}(x,y) = I(x,y)$

end

ii) Apply local thresholding method on the intermediate image I_{int}.

We choose Sauvola's method [4] which ensures a good compromise between the extraction of useful information and the elimination of the background.

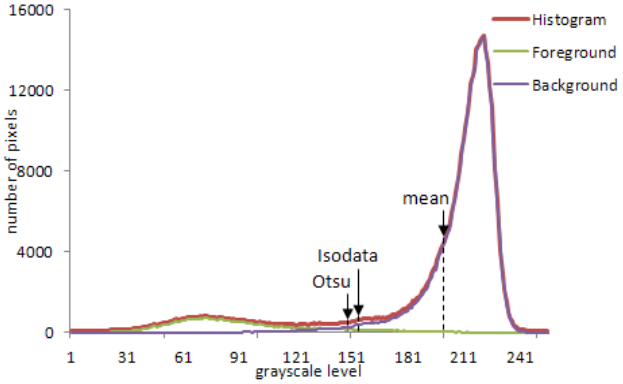


Fig. 1. Histogram of the document image. Thresholds extracted using Otsu's method, Isodata's method and the average value of the pixels.

IV. Test and Results

The proposed method has been evaluated on two different datasets. First, an Arabic historical document dataset supplied by the National Library of Algeria is used based on word spotting system. Second, a Handwritten Document Image Binarization Contest dataset is employed based on H-DIBCO 2012 evaluation measures [9].

A. Arabic Historical Document Dataset

The proposed method will be applied to degraded samples from the National Library of Algeria. The National Library recently created a dataset containing the scanned images of ancient books. 116 printed Arabic pages were selected from the book "الاقوال المرضية في علم الكرة الارضية -أحمد فايد- 1842". A representative example of the dataset is shown in Fig. 2.

Fig. 2. Example of an Arabic historical document image.

To evaluate the performance on this dataset we apply word spotting system, which is composed by the following modules:

1) Preprocessing: Several preprocessing steps have taken place, such as:
   - Binarization: All the previews methods are used to the binary image.
   - Pages separation: The document page is separated into two pages; the vertical projection profile is used in this process.
   - Skew angle correction: The horizontal projection profile is used for estimating the skew angle [10], which can be performed for different angles and the largest magnitude variations correspond to the skew angle.
   - Border removal: It aims at enhancing the document images by detecting and cutting out noisy black borders

as well as noisy text regions appearing from neighboring pages. It is based on projection profiles.

- Segmentation: It aims to extract the words from the document. Segmentation is performed in two consecutive steps: line segmentation and word segmentation. Both steps make use of the projection profiles [11].

2) Feature generation: Features are generated from every word capable of capturing the word similarities and discarding the small differences due to remaining noise or different style of writing. They are carefully selected in order to describe the contour and region shape of the word. Four features are then selected, which are:

- Projection profile: Captures the distribution of ink along one of the two dimensions in a binary word image. A vertical projection is computed by summing the intensity values in each image column separately.
- Upper word profile: The word image is scanned from top to bottom. When a black pixel is found, all the following pixels of the same column are converted to black.
- Lower word profile: The word image is scanned from bottom to top and all the pixels are converted to black until a black pixel is found.
- Number of vertical pixel transition White/Black: when a pixel changes from foreground to background in each column of the image.

3) Dynamic Time Warping (DTW): The DTW is an algorithm that efficiently computes the similarity between two sequences which can have different length. It is applied to the problem of word matching in historical documents [12].

After the binarization stage by several methods, we select 7 query words from this document to evaluate the system by word spotting criteria using DTW (see Table I).

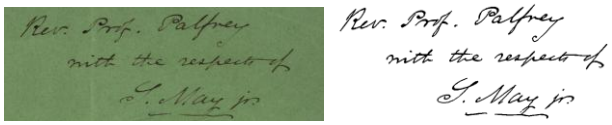
The evaluation will be based on the recall (RC), precision (PR) and F-measure (FM) which are defined as follows:

$$FM1 = \frac{2.RC1.PR1}{RC1 + PR1} \tag{4}$$

where $RC1 = \frac{C}{N}$, $PR1 = \frac{C}{M}$, C is the correctly detected word instances, N is the total number of word instances for every query (ground truth) and M is the total number of detected word instances.

B. H-DIBCO 2012 Dataset

We have also evaluated our approach on document datasets proposed in the H-DIBCO 2012 [9]. We mention that in this testing dataset, 14 handwritten images with their associated ground truth were built for the evaluation. Fig. 3 illustrates a representative example of the dataset.



(a) (b)

Fig. 3. Representative samples of H-DIBCO 2012. (a) Original image, (b) Ground truth image.

The evaluation we did, comprises an ensemble of measures that are suitable for assessment purposes in the context of document analysis and recognition [9].

- F-Measure ($FM2$)

$$FM2 = \frac{2.RC2.PR2}{RC2 + PR2} \tag{5}$$

$$RC2 = \frac{TP}{TP + FN},\ PR2 = \frac{TP}{TP + FP}$$

where *TP, FP, FN* denote the true positive, False positive and False negative values, respectively.

- Peak Signal to Noise Ratio (PSNR)

$$PSNR = 10\log(\frac{D^2}{MSE}) \tag{6}$$

where D is the difference between foreground and background, and MSE is the mean square error.

- Distance Reciprocal Distortion Metric (DRD)

It properly correlates with the human visual perception and it measures the distortion for all the flipped pixels [9].

TABLE I: EVALUATION MEASURES OF THE OTSU'S, SAUVOLA'S, NICK AND THE PROPOSED METHOD USING WORD SPOTTING CRITERIA

	Otsu (%)			Sauvola (%)			Nick (%)			Proposed method (%)		
	RC1	PR1	FM1	RC1	PR1	FM1	RC1	PR1	FM1	RC1	PR1	FM1
الكرة	100	31.77	48.22	41.09	100	58.25	78.68	100	88.07	97.22	86.00	91.48
المياه	93.93	70.45	80.51	68.75	100	81.48	96.66	85.29	90.62	75.00	100	85.71
الجبال	80.76	100	89.36	100	100	100	96.29	59.09	73.23	92.00	100	95.83
المواد	53.19	100	69.44	92.30	100	94.73	100	22.22	36.36	92.85	100	95.41
سطح	74.25	100	85.24	90.47	100	95	94.44	89.47	91.89	90.90	100	95.23
الحرارة	91.89	91.89	91.89	94.28	82.50	100	78.94	69.76	74.07	74.19	100	85.18
الارض	98.11	100	99.04	94.61	100	97.23	99.21	36.84	53.73	89.24	100	94.31
Total	84.59	84.91	80.58	83.07	97.50	87.99	92.03	66.35	72.86	87.41	98.00	**91.87**

C. Results

This experiment was conducted to show the performance of our method in several challenges such as disparity in the size of text; large, non-uniform illumination, low-quality images, thin pen stroke lines and low-contrast between the text and background.

To evaluate the performance of the proposed method, we

compare its results with those of Otsu's, Niblack's, Sauvola's and NICK methods. The parameters of the methods are fixed according to the values proposed by the authors: k=-0.2 and ws=15×15 window size for Niblak's method, k=0.5, R=128 and ws=25×25 window size for Sauvola's method and k=-0.1 and ws=19×19 window size for the NICK method. (See Fig. 4- Fig. 5).

TABLE II: EVALUATION RESULTS OF THE DATASET H-DIBCO 2012

Methods	FM2 (%)	PSNR	DRD
Otsu	75,07	15,03	26,46
Niblack	54,63	9,67	45,58
Sauvola	53,93	14,74	12,15
Nick	82,68	16,78	7,29
Winning 2012 [9]	89.47	21.80	3.44
Proposed	76,40	15,90	8,86

The presented results in Table I-II shows the efficiently of our method (ws=21×21), we have FM1=91.87 (See Table I) and FM2=76.40 (See Table II).

Based on the obtained results (See Fig. 4-Fig. 5), Niblack's method detects the text body. However, its drawback is that it produces a large amount of black noise in the empty windows. Sauvola's method solves the problem of black noise. However, it fails if the contrast between the foreground and background is small or if the text is in thin pen stroke text. The NICK method solves the low contrast problems. However, it still fails when the contrast is too small or the text is in thin pen stroke text. Our proposed approach (ws=21×21) constitutes a good compromise between the extraction of useful information and the elimination of the background.

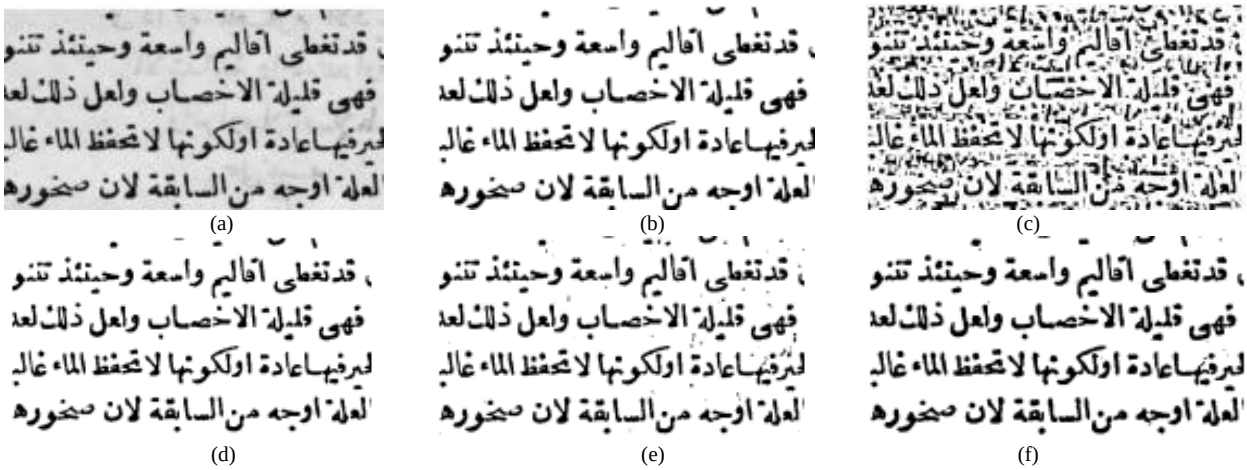


Fig. 4. Binarization results of and Arabic historical document image: (a) Original image, (b) Otsu's, (c) Niblack's, (d) Sauvola's, (e) NICK and (f) Proposed method.

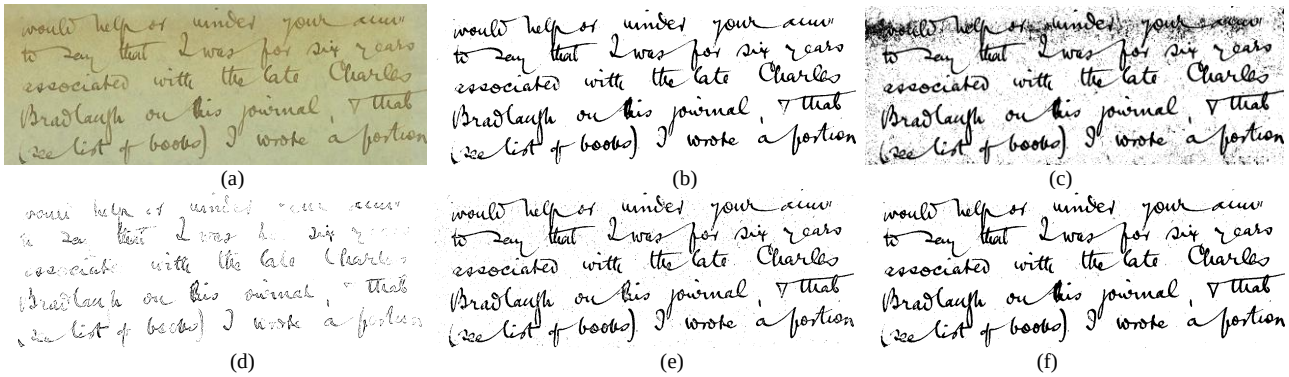


Fig. 5. Binarization results of H-DIBCO 2012 document image: (a) Original image, (b) Otsu's, (c) Niblack's, (d) Sauvola's, (e) NICK and (f) Proposed method.

V. CONCLUSION

This paper presents a document image binarization combination framework. The proposed method uses the fact that the pixels which compose the text in the document, usually, do not exceed 10% of its size.

The proposed framework divides the image pixels into two categories based on the global thresholding. Then, the intermediate image is binarized by using local thresholding method. The experimental results issued from the Arabic historical document images based on word spotting criteria using DTW and H-DIBCO 2012 dataset show the performance of our method.

In the future, we plan to implement other techniques to enhance the complex background in historical document images.

ACKNOWLEDGMENT

The authors wish to thank the National Library of Algeria for supplying historical document to carry out our experiment.

REFERENCES

[1] N. Ostu, "A Threshold Selection Method From Gray-Level Histogram," *IEEE Trans on Systems, Man and Cybernetics*, vol. 9, pp. 62-66, 1979.

[2] T. Ridler and S. Calvard, "Picture Thresholding using an Iterative Selection Method," *IEEE Trans on Systems, Man and Cybernetics*, 1978.

[3] W. Niblack, "An Introduction to Digital Image Processing. 1986," presented at International Conference on Document Analysis and Recognition, 2003.
[4] J. Sauvola and M. Pietikäinen, "Adaptive Document Image Binarization," *Pattern Recognition*, vol. 33, pp. 225-236, 2000.
[5] K. Khurshid, I. Siddiqi, C. Faure, and N. Vincent, "Comparison of Niblack Inspired Binarization Methods for Ancient Documents," in *Proc. IS&T/SPIE Electronic Imaging*, pp. 72470U-72470U-9, 2009.
[6] B. Gatos, K. Ntirogiannis, and I. Pratikakis, "ICDAR 2009 Document Image Binarization Contest (DIBCO 2009)," in *Proc. International Conference on Document Analysis and Recognition*, pp. 1375-1382, 2009.
[7] I. Pratikakis, B. Gatos, and K. Ntirogiannis, "H-DIBCO 2010-Handwritten Document Image Binarization Competition," in *Proc. International Conference on Frontiers in Handwriting Recognition*, pp. 727-732, 2010.
[8] I. Pratikakis, B. Gatos, and K. Ntirogiannis, "ICDAR 2011 Document image Binarization Contest (DIBCO 2011)," in *Proc. International Conference on Document Analysis and Recognition*, pp. 1506-1510, 2011.
[9] I. Pratikakis, B. Gatos, and K. Ntirogiannis, "ICFHR 2012 Competition on Handwritten Document Image Binarization (H-DIBCO 2012)," in *Proc. International Conference on Frontiers in Handwriting Recognition*, pp. 817-822, 2012.
[10] E. Zemouri and Y. Chibani, "Machine Printed Handwritten Text Discrimination using Radon Transform and SVM Classifier," in *Proc. International Conference on Intelligent Systems Design and Applications*, pp. 1306-1310, 2011.
[11] E. Ataer and P. Duygulu, "Retrieval of Ottoman documents," in *Proc. ACM international workshop on Multimedia Information Retrieval*, pp. 155-162, 2006.
[12] T. M. Rath and R. Manmatha, "Word Image Matching using Dynamic Time Warping," in *Proc. Computer Vision and Pattern Recognition Proceedings*, pp. II-521-II-527, vol. 2, 2003.

ET-Tahir Zemouri was born in M'sila, Algeria. He received the M.Sc. degree from the Faculty of Electronic and Computer Science, University of Sciences and Technology Houari Boumediene, Algiers, Algeria, in 2011. Actually, he is a Ph.D. student in the same faculty. His research interests include Pattern Recognition, Machine Learning and Historical Document Image Analysis and Recognition.

Youcef Chibani was born in Algiers, Algeria. He received the master's and State Doctoral degrees in electrical engineering from the University of Science and Technology Houari Boumediene, Algiers, Algeria. He has been teaching and researching as an Assistant Professor since 2002. His research interests include the use of the wavelet decomposition, neural networks, and support vector machines in many applications as multi-sensor image fusion, change detection, pattern recognition, and multimedia signal processing, as well as document image analysis. He coauthored many papers published in international peer-reviewed journals and conferences.

Youcef Brik received the M.Sc. degree from the Faculty of Electronic and Computer Science, University of Sciences and Technology Houari Boumediene, Algiers, Algeria, in 2010. Actually, he is a Ph.D. student in the same faculty. His research interests include Machine Learning, Pattern Recognition, Artificial Intelligence and Soft Computing, as well as Document Image Processing.