2011 International Conference on Document Analysis and Recognition

New Binarization Approach Based on Text Block Extraction

Ines Ben Messaoud, Hamid Amiri
Laboratoire des Systèmes et Traitement de Signal (LSTS)
Ecole Nationale d'Ingénieurs de Tunis (ENIT)
Tunis, Tunisia
ibmnoussa@gmail.com, hamidlamiri@yahoo.com

Haikal El Abed, Volker Märgner
Institute for Communications Technology (IfN)
Technische Universität Braunschweig
Braunschweig, Germany
{*elabed, v.maergner*}*@tu-bs.de*

***Abstract*—Document analysis and recognition systems include, usually, several levels, annotation, preprocessing, segmentation, feature extraction, classification and post-processing. Each level may be dependent on or independent from the other levels. The presence of noise in images can affect the performance of the entire system. This noise can be introduced by the digitization step or from the document itself. In this paper, we present a new binarization approach based on a combination between a preprocessing step and a localization step. The aim of the present approach is the application of binarization algorithms on selected objects-of-interest. The evaluation of the developed approach is performed using two benchmarking datasets from the last two document binarization contests (DIBCO 2009 and H-DIBCO 2010). It shows very promising results.**

***Keywords*-Document image binarization; Binarization evaluation; Document analysis; Preprocessing.**

I. Introduction

A document analysis and recognition system is considered as a complex process. In order to perform the efficiency of such a system, each step has to be efficient beginning from the first steps such as annotation [1] or preprocessing.

The presence of noise in images, especially in historical documents, is unavoidable. This noise is introduced by image scanning, recording or transmission and may cause errors in the processing of these documents. In order to allow better quality of the input image, the application of noise reduction algorithms seems to be necessary. Several techniques were proposed for reducing the noise sensitivity, such as special filters or noise and shadow removal. The better the noise removal methods, the better the binarized image returned. Binarization is the main step in the preprocessing level. Pixels in a binary image are classified either as foreground $\mathcal{F}$ or as background $\mathcal{B}$. The quality of the binarization is critical for the analysis step. If bad binarized images are used, document processing may yield false results.

Binarization is the first step in the preprocessing of a document analysis and recognition system. It is a technique which transforms a gray-scale Ig or a color image Ic to a binary image BW. We can identify three binarization classes [2], global binarization, e.g. the well known Otsu's method [3], local binarization, e.g. of local binarization methods Bernsen [4], Niblack [5], Sauvola [6] and Lu [7] and hybrid binarization, e.g. Kuo [8].

In order to evaluate binarization performance, it is necessary to use an objective evaluation ([9], [10] and [11]) based on evaluation rates and not on visual evaluation. Several binarization methods lose efficiency, if used documents have bad qualities, or if they deals with specific document characteristics (different fonts, different background).

This paper is organized in 5 sections. In Section II we present the proposed approach. Section III describes experimental setup and test results. In Section IV we discuss the obtained results. Section V describes some possible extensions and future works.

II. Proposed Method

In this section we present our approach of binarization. We have integrated a prebinarization step in order to enhance the input image quality. The input of the binarization method is a set of selected image regions. Figure 1 shows an overview of the proposed binarization architecture.

The original image Ic is an RGB color image. Ic is converted to a gray-scale image Ig according to the following equation $Ig = 0.2989 \cdot Ic_1 + 0.5870 \cdot Ic_2 + 0.1140 \cdot Ic_3$. Because most of the test images present different degradations, we have integrated different noise removal methods before binarization, in order to enhance the quality of the gray-scale image Ig. The output of the noise correction methods is the gray-scale image Ig'. We have applied a localization method on Ig', which returns the set of the objects-of-interest $\{O_i\}$. These objects are the inputs of the binarization method. Document pixels, which did not belong to any detected object, are classified as background. We have tested the proposed approach using different noise removal, region localization and binarization methods. We denote in this paper p an image pixel having (x, y) as coordinates, M and N are respectively the width and height of Ig.

A. Noise Removal

Because historical documents present different degradations, e.g. shadows, dirty background and smudges, application of noise removal algorithms seems to be necessary. The

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.243

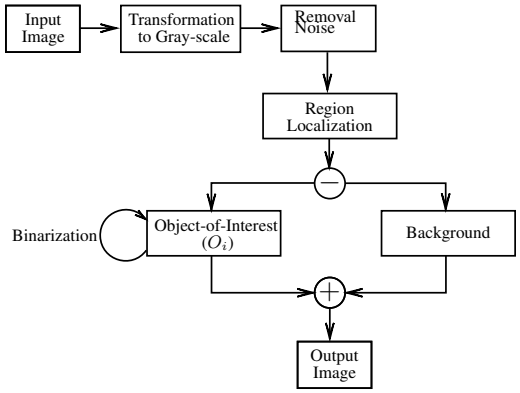

Figure 1. Architecture of the proposed binarization approach including denoising and localization steps

gray-scale images $Ig(x, y)$ and $Ig'(x, y)$ are considered as the input and the output of the noise removal functions.

1) Shading Correction: Shading correction filters have been used in order to minimize the signal inhomogeneity within an image [12]. The shadings can be described as multiplicative or additive components for the document. Equations 1 and 2 describe shading correction filters using division and subtraction, respectively.

$$I'_g(x, y) = \frac{Ig(x, y)}{b(x, y)} \cdot K_d \tag{1}$$

$$Ig'(x, y) = Ig(x, y) - b(x, y) + K_s \tag{2}$$

$b(x, y)$ is estimated using the low-pass or the median filter, in our approach we have used the median filter. K_d and K_s are the amplification and the correction parameters.

2) Wiener Filter: The Wiener filter is applied locally to $N_x \times N_y$ neighborhoods of the pixel p. σ^2 is the variance at $N_x \times N_y$ neighborhood, μ is the local mean and v^2 is considered as the average of all estimated variances for each p and its neighborhoods. Equation 3 describes the output of the Wiener filter.

$$Ig'(x, y) = \mu + \frac{\left(\sigma^2 - \upsilon^2\right)\left(Ig(x, y) - \mu\right)}{\sigma^2} \tag{3}$$

B. Localization

We have integrated a localization for the object-of-interest $\{O_i\}$ at the prebianrization step. L is the total number of objects in an image. We used for this task either edge detection or connected component. Selected objects-of-interest $O_i \in \{O_1, \ldots, O_L\}$ are considered as the inputs of the method of binarization. The rest of the image is considered as background. The input image is denoted as $Ig'(x, y)$.

1) Edge Detection: We have used as first method of object-of-interest localization an edge detection method. We have estimated edges using the Canny method [13]. The output image $Bw'(x, y)$ of the Canny edge detection is described by Equation 4.

$$Bw'(p) = \begin{cases} 1 & \text{, if } p \in Edge(Ig'(p)) \\ 0 & \text{, otherwise} \end{cases} \tag{4}$$

An object region O_1 is the minimum box including the result of the Canny edge detection ($Bw'(p) = 1$).

2) Connected Component: The second method of the localization of object-of-interest uses connected component technique. This method has as input the binary image $Bw'(x, y)$ obtained by Canny's method. The labeled matrix C describes labels of the connected components in $Bw'(x, y)$, where J denotes the total number of connected components. The object O_1 is the box containing the connected component having the label 1 ($C = 1$). For each connected component, having the label $C = j$ where $1 < j \leq J$, the intersection between the box B containing the current connected component and the objects $\{O_i\}$ was calculated. For the first intersection between B and one object O_i, the comparison was stopped. O_i was updated as a union of O_i and B. If the intersection was empty, we added a new object $O_{i+1} = B$. L is the total number of the objects-of-interest.

C. Binarization

We have used different binarization methods in order to choose the most efficient on the test dataset and noise correction filters.

1) Otsu: We consider t^* as a threshold returned by Otsu's method. t^* is determined using Equation 5,

$$t^* = \arg \max_{0 \leq t < G} \left\{\omega_1(t) \cdot \mu_1^2(t) + \omega_2(t) \cdot \mu_2^2(t)\right\} \tag{5}$$

where gray-level pixel intensities in $Ig'(x, y)$ are ranged from 0 to $G-1$. $\omega_1(t)$ and $\omega_2(t)$ are the probabilities of the foreground class ($p \leq t$) and background class ($t < p < G$). $\mu_1(t)$ and $\mu_2(t)$ are considered as the mean gray level values of the foreground and background classes, respectively.

2) Sauvola: $T(x, y)$ threshold is calculated using Equation 6,

$$T(x, y) = \mu(x, y) + \left[1 + c \cdot \left(1 - \frac{\sigma(x, y)}{A}\right)\right] \tag{6}$$

where $\mu(x, y)$ and $\sigma(x, y)$ are the average and standard deviation, respectively. c, A and sliding window w size are input parameters. A is the dynamic range of the standard deviation and c is a fixed parameter.

3) Lu: We used a modified version of Lu's method. For the estimation of the document background we have used a digital smoothing polynomial filter [14]. The binary image $BW(x, y)$ output of Lu's method before the application of post processing techniques is given in Equation 7.

$$BW(x,y) = \begin{cases} 0 & \text{, if } N_e \geq N_{min} \text{ and} \\ & f(x,y) \leq E_{mean}(x,y) \\ 1 & \text{, otherwise} \end{cases} \quad (7)$$

$f(x,y)$ is the output of the application of preprocessing methods on $Ig'(x,y)$. $E_{mean}(x,y)$ is a threshold, which is calculated locally. N_e is calculated locally within a window w centered in p, denoting the number of the detected stroke edge pixels and N_min is an input parameter.

III. Tests and Results

A. Datasets

We have evaluated the proposed approach on two document datasets proposed at the last binarization competitions DIBCO 2009 [9] and H-DIBCO 2010 [10]. DIBCO 2009 dataset[1] originates from the collections of different libraries. It includes selected parts of images extracted of 10 different historical documents. 5 are printed text images and 5 handwritten text images. The selected images contain representative degradations. H-DIBCO2010 dataset[2] consists of 10 handwritten document images. It originates from the library of congress and contains also representative degradations. DIBCO is a set of color images and their corresponding ground-truth GT, H-DIBCO dataset is a set of color images and their corresponding GT and skeleton ground-truth SG.

B. Evaluation Metrics

The evaluation metrics were adopted for the evaluation of participating methods during the last two competitions (DIBCO 2009 and H-DIBCO2010) including $Fmeasure$ and pFM (pseudo $Fmeasure$), peak signal-to-noise ratio ($PSNR$), negative rate metric (NRM) and misclassification penalty metric (MPM). These metrics are calculated using the ground-truth GT, skeleton ground-truth SG and the binary image BW resulting from the binarization method. Binarization quality increases when $Fmeasure$, pFM and $PSNR$ increase and NRM and MPM decrease.

- $Fmeasure$ and pFM are defined in Equation 8 using $Precision$ and $Recall$ (respectively $pRecall$). $Precision$ and $Recall$ are calculated pixel-wise using ground-truth GT image and binarized image BW. For $pRecall$, used to calculate pFM, the skeleton image SG is used as reference ground-truth.

$$Fmeasure = \frac{2Recall \cdot Precision}{Recall + Precision} \quad (8)$$

- $PSNR$ measures (Equation 9) how close the binary image is to ground-truth using the mean square error (MSE) and a constant C as measure for the difference between $\mathcal{F}$ and $\mathcal{B}$ pixel intensities (C is set to 1 as maximum distance).

$$PSNR = 10 \cdot log_{10}\left(\frac{C^2}{MSE}\right) \quad (9)$$

- NRM represents the relationship (defined in Equation 10) between the ground-truth pixels and the binarized image pixels (N_{TP}, N_{FP}, N_{FN} and N_{TN} are number of true positives, false positives, false negatives and true negatives, respectively).

$$NRM = \frac{\frac{N_{FN}}{N_{FN}+N_{TP}} + \frac{N_{FP}}{N_{FP}+N_{TN}}}{2} \quad (10)$$

- MPM, as defined in Equation 11, measures the distance between the contours of the ground-truth and the binarized image. d^i_{FN} and d^j_{FP} represent the distance of the i^{th} false negative and the j^{th} false positive from the contour of the ground-truth, respectively. The normalization factor D is the sum over all the pixel-to-contour distances of the ground-truth object.

$$MPM = \frac{\sum_{i=1}^{N_{FN}} d^i_{FN} + \sum_{j=1}^{N_{FP}} d^j_{FP}}{2D} \quad (11)$$

The best binarization method is the one that has the best accumulated rank R_m [10], where m denotes the tested binarization method. R_m is the accumulation of $r(m,e)$, where $r(m,e)$ is the rank of the method m using the e^{th} evaluation measure.

C. Experimental Setup

In the first experiment we have studied the performance of each binarization method with different denoising techniques using DIBCO 2009 dataset. The object-of-interest localization was applied using Canny's edge detection method. We have applied 3 different binarization methods (local and global algorithms), i.e Otsu, Sauvola and Lu. Each method was applied first without using any noise removal technique, then using a median shading filter and finally using Wiener filtering.

In the second experiment, we have fixed a binarization method according to the results from the first test. In order to perform better quality of binarization images, we have tested different combinations of noise removal with different proposed methods for object-of-interest localization, as described in Sections II-A and II-B respectively. We used 3 different methods in our test experiments, we have fixed the binarization method and we have varied the denoising and the object-of-interest localization techniques. In method 1 we have applied Wiener filter and canny edge, in method 2 Wiener filter and connected component and in method 3 we have applied only connected component without any denoising technique.

[1] http://users.iit.demokritos.gr/~bgat/DIBCO2009/benchmark/
[2] http://www.iit.demokritos.gr/~bgat/H-DIBCO2010/benchmark

Table I
AVERAGE OF THE EVALUATION METRICS OF OUR APPROACH USING DIFFERENT BIANRIZATION METHODS WITH DIFFERENT NOISE REMOVAL FILTERS

	$Fmeasure$	$PSNR$	NRM $(\cdot 10^{-2})$	MPM $(\cdot 10^{-3})$
	Without noise removal			
Lu	**90.23**	**18.16**	6.73	**0.78**
Sauvola	84.26	16.64	11.98	1.23
Otsu	80.73	15.88	**5.67**	3.66
	Median shadow			
Lu	**89.57**	**17.85**	**7.28**	0.92
Sauvola	63.35	13.17	23.27	6.17
Otsu	88.24	17.23	7.79	**0.78**
	Wiener Filter			
Lu	**90.39**	**18.30**	6.65	**0.63**
Sauvola	82.91	16.40	13.08	1.21
Otsu	80.59	15.88	**5.55**	3.67

We have compared proposed methods 1, 2 and 3 with methods participating at the last binarization competitions and using document datasets of DIBCO 2009 and H-DIBCO 2010 respectively.

D. Results

According to the results shown in Table I, the evaluation of binarization approach performance was calculated according to 4 evaluation metrics. It is notable that the application of denoising techniques can increase the performance of the binarization method or it can cause loss of information. The application of median shadow filter affects Sauvola binarization performance, the average of 4 evaluation metrics became worse (e.g. $Fmeasure$ has decrease from 84.26% to 63.35%). The application of the Wiener filter before Otsu binarization algorithm have increased the performance of Otsu's method, the average of 3 evaluation measures was ameliorated. The best results were achieved at the application of the combination of Wiener filter as the denoising method and Lu as the binarization method. According to the first experiment results, we have fixed in the next tests Lu as binarization method because it has the best accumulated rank compared with the other binarization methods.

In order to evaluate the performance of the proposed approach of binarization using localisation of object-of-interest, we have compared three methods 1, 2 and 3, with the participated methods at the last binarization competitions. These methods are defined as follows, method 1(noise removal: Wiener filter, localization: Canny edge, binarization: Lu), method 2 (noise removal: Wiener filter, localization: connected components, binarization: Lu) and method 3 (noise removal: without noise correction, localization: connected components, binarization: Lu). Samples from the test dataset binarized using proposed methods are shown in Figure 2. Figure 2(b) and 2(d), samples of handwritten and printed document, respectively, originated from DIBCO dataset, were binarized using method 1 and 2,

Figure 2. Samples of the test dataset resulted from binarization using localization of object-of-interest 2(b): image binarization of 2(a), 2(d): image binarization of 2(c), 2(f): image binarization of 2(e)

Table II
COMPARISON OF OUR BINARIZATION METHODS WITH THOSE PARTICIPATING AT DIBCO 2009 COMPETITION

	$Fmeasure$	$PSNR$	NRM $(\cdot 10^{-2})$	MPM $(\cdot 10^{-3})$
1st	**91.24**	**18.66**	4.31	0.55
2nd	90.06	18.23	4.75	0.89
3rd	89.34	17.79	5.32	1.9
Proposed methods				
m_1	90.39	18.30	6.65	0.63
m_2	90.60	18,37	6,64	**0,31**

respectively. Figure 2(f), sample of H-DIBCO dataset was binarized using method 3.

Methods 1 and 2 were tested on DIBCO dataset. Results are shown in Table II, both proposed methods 1 and 2 are classified as the 4^{th} and the 2^{nd}, respectively, among 45 , according to the accumulated rank R_m using the 4 evaluation metrics $Fmeasure$, $PSNR$, NRM and MPM.

Methods 1 and 3 were tested using H-DIBCO dataset. Results are shown in Table III, both proposed methods 1 and 2 are classified as the 8^{th} and 6^{th}, respectively, among 17 participating methods, using the accumulated rank R_m and 5 evaluation measures $Fmeasure$, pFM, $PSNR$, NRM and MPM.

IV. DISCUSSION

According to our results, the proposed approach of binarization using localization of object-of-interest gives good results for printed documents (e.g. $Fmeasure = 91.17\%$). There is some loss of information for handwritten documents having fine stroke width. In this case the foreground pixels

Table III
COMPARISON OF OUR BINARIZATION METHODS WITH THOSE PARTICIPATING AT H-DIBCO 2010 COMPETITION

	$Fmeasure$	$p-FM$	$PSNR$	NRM ($\cdot 10^{-2}$)	MPM ($\cdot 10^{-3}$)
1st	**91.50**	93.58	**19.78**	5.981	0.492
1st	89.70	95.15	19.15	8.180	**0.288**
2nd	91.78	94.43	19.67	**4.771**	1.334
Proposed methods					
m_1	86.74	93.36	18.16	10.48	0.541
m_3	86.33	**93.79**	18.03	10.76	0.379

may be considered as background pixels. This explains the fact that our method has better quality for DIBCO dataset (printed and handwitten documents) than for H-DIBCO dataset (only handwritten documents).

Table I shows that our approach of binarization using object-of-interest localization returns promising results, when we apply Otsu binarization and median shadow filtering. Compared to the methods participating at DIBCO 2009, this method ($Fmesaure = 88.24\%$, $PSNR =$17.23, $NRM =$7.79 $\cdot 10^{-2}$ and $MPM =$0.78 $\cdot 10^{-2}$) can be classified as the 5^{th} method. These results are very promising because Otsu's method checks automatically the best threshold without any input parameters in opposition to the other methods (e.g. Sauvola and Lu), which need adjusted input parameters. According to these results the application of our approach using Otsu can be extended to large databases.

It is obvious that our approach limits the input of the binarization method, only a set of objects-of-interest was binarized, the other regions were considered as background. This approach minimizes misclassification errors (It returns the 1^{st} and the 5^{th} best MPM values using DIBCO and H-DIBCO datasets, respectively).

One important improvement step for binarization approaches is the decrease of the dependency of the manual settings for different filters. This step is required for the developed binarization approach on a large set of historical books, including different effects. In other words, it is important to invest effort in research works to increase the adaptivity of binarization methods and to develop ground-truth independent evaluation approaches.

V. CONCLUSIONS

In this paper we propose a new approach of binarization based on object-of-interest localization. The tests of this approach were realized on document datasets from the last binarization competitions DIBCO 2009 and H-DIBCO 2010, which propose original images and their corresponding ground-truth. Because test dataset images present different degradations we apply different filters for correction of noise before the application of binarization. The evaluation of the proposed approach is based on the calculation of evaluation metrics. The proposed approach gives promising results and it returns better results using DIBCO dataset than H-DIBCO. This work will be extended using automatic methods for the choice of best parameters according to the input image. Another kind of documents will also be included in our tests.

ACKNOWLEDGMENT

A part of this work was supported by the DAAD (German Academic Exchange Service).

REFERENCES

[1] I. Ben Messaoud and H. El Abed, "Automatic annotation for handwritten historical documents using markov models," in *ICFHR*, 2010, pp. 381–386.

[2] M. Sezgin and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," *Journal of Electronic Imaging*, vol. 13, pp. 146–165, 2004.

[3] N. Otsu, "A threshold selection method from gray level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. 19, pp. 62–66, 1979.

[4] J. Bernsen, "Dynamic thresholding of grey-level images," in *ICPR*, 1986, pp. 1251–1255.

[5] W. Niblack, "An introduction to digital image processing," in *Prentice Hall, Englewood Cliffs*, 1986, pp. 115–116.

[6] J. Sauvola and M. Pietikinen, "Adaptive document image binarization," *Pattern Recognition*, vol. 33, no. 2, pp. 225–236, 2000.

[7] S. Lu, B. Su, and C. L. Ta, "Document image binarization using background estimation and stroke edge," *IJDAR*, vol. 13, no. 4, pp. 303–314, 2010.

[8] T. Kuo, Y. Lai, and Y. Lo, "A novel image binarization method using hybrid thresholding," in *IEEE International Conference on Multimedia & Expo (ICME)*, 2010, pp. 608–612.

[9] B. Gatos, K. Ntirogiannis, and I. Pratikakis, "ICDAR2009 document image binarization contest (DIBCO2009)," in *ICDAR*, 2009, pp. 1375–1382.

[10] ——, "H-DIBCO 2010-handwritten document image binarization competition," in *ICFHR*, 2010, pp. 727–732.

[11] R. Paredes and E. Kavallieratou, "ICFHR2010 contest : Quantitative evaluation of binarization algorithms," in *ICFHR*, 2010, pp. 733–736.

[12] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. Pearson Prentice Hall, 2008.

[13] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8, no. 6, pp. 679–698, 1986.

[14] S. Lu and C. L. Tan, "Binarization of badly illuminated document images through shading estimation and compensation," in *ICDAR*, 2007.