Document Preprocessing System – Automatic Selection of Binarization

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

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

***Abstract*—Due to the reason that historical documents present many degradations, the analysis of such documents is considered as a big challenge. In this paper we present a system which allows automatic preprocessing of historical documents. One or many preprocessing methods, as well as sets of input parameters are selected for each book from the used database according to the input image features. Such selection is tested on a subset of every book during the training step, the validation of the carried results is performed on another subset of images. If the validation is not well checked, the training is repeated. The proposed system is applied on a set of books from the Google-Books (23 books, 1000 images) and the Bayerische Staatsbibliothek (10 books, 750 images) collections. The performed results are very promising.**

***Keywords*-Historical document analysis, binarization, automatic parameters selection**

I. Introduction

Preprocessing is considered as one of the important steps of document analysis and recognition system. The better results the preprocessing returns, the higher the recognition rate that will be performed [1]. Preprocessing is the combination of noise removal, binarization or thresholding and foreground/background segmentation algorithms. Binarization is the crucial step of preprocessing because the output of the preprocessing step is a logical image. In order to develop an automatic preprocessing system of historical documents, it is necessary to allow automatic evaluation of binarized images. There are three possibilities of binarization evaluation: visually, according to the recognition and error rates or using ground-truth images. Evaluation algorithms [2], which are based on the visual evaluation from an expert, lack precision and are time consuming. The evaluation based on the recognition rate [3] does not evaluate only the preprocessing step but the whole phases of the system. The evaluation using the ground-truth images shows difficulties [4] when it deals with a large database, because the generation of ground-truth for binarization runs manually [5] or semi-automatically [6]. The proposed works for ground-truth generation have been applied on a limited number of parts of images [7], [8], [9], [10] or on synthetic images [11]. Due to the reason that the generation of ground-truth is a complex function as well as time consuming, it is meaningful to propose an automatic preprocessing system without the need to generate the ground-truth of the totality of images of every historical book.

Digital images belonging to different books of the same database are generally different. This explains that it is not desirable to apply the same preprocessing method with the same parameters on all images of the same document. For that reason the selection of a binarization method as well as its parameters for each book of the database is necessary [12], [13]. According to the first tests carried in [8], the selections of the binarization method and its parameters depend on the characteristics of the input document. In order to allow an automatic preprocessing system, we propose to deal with a system of the selection of the binariaztion method and its parameters on a subset of images of each book and not on the totality of the images. This proposed method has to respond to two questions:

1) How many images have to be used for the most satisfying selection of the preprocessing methods (Training phase)?
2) The choice of the binarization method and the input parameters for such method selected during the training can be generalized for all the books (Validation phase)?

In this paper we try to answer these two questions. For that reason we propose in Section II the description of the proposed system for preprocessing. Section III shows the experimental results. In Section IV we discuss the results reached during our work. Section V summarizes the proposed preprocessing system and some future ideas.

II. Document Preprocessing System Architecture

We propose a system for automatic preprocessing based on two phases, the training and the validation. In this section we present the architecture of the proposed system as well as the different phases.

A. Architecture Overview

Figure 1 shows the process of the preprocessing phase. The parametrization of each binarization method is performed using [8]. The binarization methods $m_l \in M$ are applied on the images $I(x, y) \in S$, where S and M denote

978-0-7695-4661-2/12 $26.00 © 2012 IEEE
DOI 10.1109/DAS.2012.31

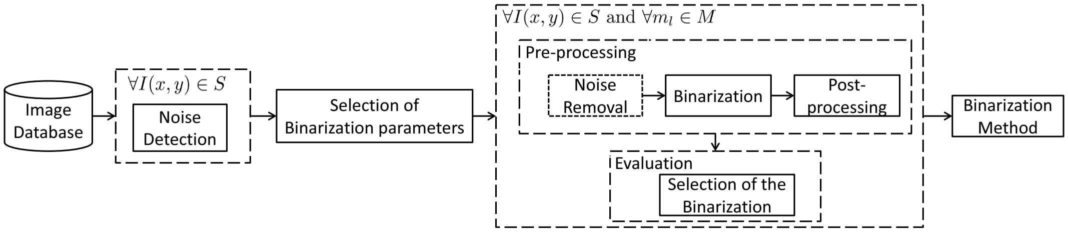


Figure 1. Architecture of the training phase of automatic selection of binarization, where S is a subset of images and M is a set of the tested binarization methods

a subset of images and a subset of binarization methods respectively. One or many binarization methods are selected as the best methods according to a set of evaluation metrics.

B. Data Preparation

The proposed system allows the selection of the binarization and its input parameters. For that reason each step of the proposed system is evaluated. In order to select one of the appropriate binarization methods, we have evaluated the binary images, results of the application of $m_l \in M$, using the ground-truth images for binarization corresponding to the input image $I(x, y)$. The ground-truth images for binarization are generated using the method proposed in [14]. This method is an adaption of the semi-automatic method for ground-truth generation used in the last competitions of binarization [6].

Black pixels both in ground-truth $GT(x, y)$ and in binary images are classified as foreground and white pixels as background. Every pixel is clustered within one of four classes, true positive (TP), false positive (FP), false negative (FN) and true negative (TN). The evaluation metrics used to evaluate the results of the proposed framework were adopted from the binarization competitions DIBCO2009 [7] and H-DIBCO2010 [9], and which are Fmeasure (FM), pseudo Fmeasure ($p - FM$), peak signal-to-noise ratio ($PSNR$), negative rate metric (NRM), misclassification penalty metric (MPM), geometric-mean pixel accuracy (GA) [11] and normalized cross correlation (ρ) [15]. The totality of those metrics were used for the evaluation of binarization images in [14].

An accumulation rank R_{m_l} is in accordance with each binarization method m_l. R_{m_l} denotes the sum of $r(m_l, e)$, which is the rank of the binarization method m_l using the e^{th} evaluation metric, $e \in \{1, \ldots, 7\}$. R_{m_l} is calculated using

$$R_{m_l} = \sum_{p=1}^{7} r(m_l, e) \tag{1}$$

$rank(m_l)$ denotes the function which returns the rank of the binarization method m_l giving B_k. The first ranking binarization is the one which has the minimum R_{m_l} giving B_k.

The selection of the input parameters of binarization methods is performed according to the input document features. Due to the reason that historical documents have the common characteristic of the presence of noise, the type of noise was used as the unique feature to select the input parameters. We have defined four classes of noise denoted C_j, where $1 \leq j \leq 4$. The class with images presenting show-through is C_1, images presenting high similarity between foreground and background belong to C_2, images with variable background belong to C_3, otherwise the images are classified into C_4. The ground-truth of the noise detection method is performed subjectively by an expert. Each image $I(x, y)$ is classified by an expert into one of the classes C_j, $\forall j, 1 \leq j \leq 4$.

C. Training

The training phase is applied on a subset θ_k of each book B_k of the used collection. For each document $I(x, y) \in \theta_k$ the class of noise C_j is selected using the method proposed in [8]. The type of noise is defined based on the gray-scale image according to the image histogram and to the Otsu's method of binarization due to the reason that this method does not need any input parameters. The detection of the type of noise is performed in order to select the input parameters of the binarization methods as shown by

$$\forall m_l \in M \text{ and } \forall j,\ 1 \leq j \leq 4$$
$$param(m_l|C_j) = (a_1, a_2, \cdots, a_{n_l}), \tag{2}$$

where n_l is the number of the method m_l input parameters. The used parameters during the training for each method m_l are those selected for the class C_j, which is the class containing the majority of images $I(x, y) \in \theta_k$.

A set of binarization methods $m_l \in M$ is applied using the input parameters $param(m_l|C_j)$. For each binarization m_l a set of evaluation metrics is performed and the best binarization methods (1^{st} and 2^{nd}) are those having the minimum accumulated ranking R_{m_l}.

D. Validation

During the validation phase, the selection performed during the training is evaluated. The validation is applied on a different subset of documents β_k for each book B_k, where $\theta_k \cap \beta_k = \emptyset$ and $\#\beta_k = \frac{\#\theta_k}{2}$, ($\#$ denotes the number of images). During the validation we judge if the number of images used during the training is sufficient for the selection of the binarization method and its parameters. A binarization method is chosen as the best m_l for the images $I(x,y) \in \beta_k$ using the same input parameters during the training. A comparison between the results determined during training and validation is performed. If the method carried out during the validation figures between the two best binarization methods selected during the training, the training is validated. Otherwise new images are added to the last subset, where the new subset θ_k is composed from the images used in the last iteration and the new images. At every iteration we add 5% of the image book B_k and the training is repeated at maximum three times. That means that 15%, 20% and 25% of the totality of images are used during the first, second and third iterations, respectively. If after the third iteration the training is not validated the binarization method selected in the training and which returns the best value of $PSNR$ is chosen, because it has been proved in [15] that $PSNR$, FM and ρ are considered as good metrics for binarization evaluation.

III. Experimental Setup

The proposed system is evaluated on sets of printed historical books from the Google-Books collection (Version 1.0 Aug, 17, 2007)[1], and handwritten books from the Bayerische Staatsbibliothek (BSB)[2] collection. 23 printed Latin books are used from the Google-Books collection and 10 Arabic handwritten books from the BSB collection. The tests are achieved on first 75 page images from each book of the both collections.

A. Tests and Results

The method for the selection of the binarization method input parameters according to the noise features is tested on the benchmarking dataset of binarization DIBCO 2009 [7]. During our tests six binarization methods are used, namely Otsu (m_1) [16], Bernsen (m_2) [17], Niblack (m_3) [18], Sauvola (m_4) [19], Gatos (m_5) [20], and Ben Messaoud (m_6) [10]. Otsu's method (m_1) has no input parameters. Every binarization method m_l, $l \in \{2, \cdots, 6\}$ has n_l input parameters, Bernsen (t: threshold, w: window size), Niblack (k: weight, w: window size), Sauvola (r: adaptive range, w: window size, k: weight), Gatos (m: binarization method, q, q_1, q_2: thresholding parameters, w: window size) and Ben Messaoud (v, min: thresholding parameters and can: Canny's threshold). Gatos' method is combined with a Wiener's filter and use the local theresholding method m which refers to either Niblack or Sauvola's method. As post-processing the Gatos' binarization is combined with shrink and swell filtering [21]. In Ben Messaoud's method a Wiener's filter is applied on the gray-scale image and after the binarization a post-processing method is performed, which allows the elimination of small connected components considered as false alarms. The input parameters for each pair (m_l, C_j) are found according to our tests performed in [8] and resumed in Table I.

Table I
Selection of the input parameters of the Binarization methods used according to the class of noise

		C_1	C_2	C_3	C_4
m_2	t	100	140	100	100
	w	35	25	35	35
m_3	k	-0.2	-0.2	-0.2	-0.2
	w	80	125	180	180
m_4	r	80	128	71	71
	k	0.2	0.2	0.2	0.2
	w	24	30	30	20
m_5	m	m_3	m_2	m_2	m_2
	q	0.5	0.5	0.5	0.5
	q_1	0.6	0.6	0.6	0.6
	q_2	0.8	0.8	0.8	0.8
	w	20	31	70	80
m_6	v	60	15	50	50
	min	110	20	30	35
	can	0.7	0.2	0.3	0.5

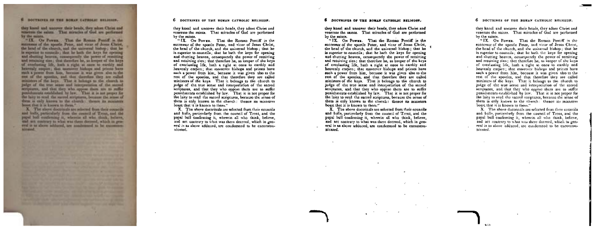
Only images used during the training are classified into one class C_j, the input parameters for the methods m_l are those defined for the class C_j (see Table I) which contains the majority of images used during the training. The results of the training and the validation phases applied on the BSB and the Google-Books collection are shown in Tables II and III, respectively. In every iteration the books which are not validated are marked in bold. Based on Table II, the validation of the training of 8 books of the BSB collection is achieved during the first iteration (only 10% of the book images), the training of the B_5 is validated using 15% of the totality of images and the training of B_4 is performed in iteration 3. The training of 21 books of the Google-Books collection is validated after the third iteration. As detailed in Section II-D, if after three iterations the training is not yet validated, the binarization method is the one of both methods performed during the training and having the best $PSNR$. In this case the Gatos' method (m_5) returns the best $PSNR$ for both books B_4 and B_6. Based on Figure 2 it is notable that for a sample image from the training subset of B_4, the binary images returned during the training (2(b) and 2(c)) are better than the binary image selected during the validation 2(d).

In order to evaluate the proposed system of automatic preprocessing, the whole concept is tested on the rest of the images (which are not used during the training and the validation steps) so called γ_k, where $\gamma_k \cap \theta_k = \emptyset$. For this evaluation step, we have developed three different methods:

- 1^{st} method: The binarization methods classified as the first ones ($rank(m_l) = 1$) during the training for each

[1] http://books.google.com

[2] http://www.bsb-muenchen.de



(a) (b) (c) (d)

Figure 2. Sample image from the book B_4 from the Google-Books used during the training (a) binarized using m_6 in (b), m_5 in (c), and m_2 in (d)

Table II
TRAINING AND VALIDATION PHASES APPLIED ON HANDWRITTEN BOOKS FROM THE BSB COLLECTION

	Training		Validation
	1^{st}	2^{nd}	1^{st}
Iteration 1			
B_1	m_2	m_4	m_2
B_2	m_6	m_2	m_6
B_3	m_6	m_4	m_6
$\mathbf{B_4}$	$\mathbf{m_6}$	$\mathbf{m_2}$	$\mathbf{m_5}$
$\mathbf{B_5}$	$\mathbf{m_4}$	$\mathbf{m_5}$	$\mathbf{m_6}$
B_6	m_2	m_1	m_6
B_7	m_6	m_2	m_2
B_8	m_6	m_2	m_6
B_9	m_2	m_5	m_2
B_{10}	m_2	m_1	m_2
Iteration 2			
$\mathbf{B_4}$	$\mathbf{m_6}$	$\mathbf{m_2}$	$\mathbf{m_5}$
B_5	m_4	m_6	m_4
Iteration 3			
B_4	m_2	m_6	m_2

book B_k are applied on γ_k.

- 2^{nd} method: The binarization methods classified as the second ($rank(m_l) = 2$) during the training for each B_k are applied on γ_k.
- 3^{rd} method: All the binarization methods $m_l \in M$ are applied on γ_k on each book, the classification of the binarization methods is performed, one binarization method is selected as the most appropriate method for each book B_k. Those methods are applied on the set γ_k.

The averages of the evaluation metrics after the application of the selected binarization for each book in first to third methods are calculated. The best average values of the evaluation metrics are those carried out in third method, because the most appropriate binarization methods for the new subset γ_k are selected. Based on the results shown in Table IV, it is notable that if we apply the methods selected during the training either the first or the second ones (first method and third method), the results are very close to the best values in third method. It is shown that the choice of the binarization as well as the input parameters are well evaluated.

Table III
TRAINING AND VALIDATION PHASES APPLIED ON A SELECTION OF PRINTED BOOKS FROM THE GOOGLE-BOOKS COLLECTION

	Training		Validation
	1^{st}	2^{nd}	1^{st}
Iteration 1			
B_0	m_4	m_6	m_6
B_1	m_6	m_4	m_6
$\mathbf{B_2}$	$\mathbf{m_4}$	$\mathbf{m_5}$	$\mathbf{m_6}$
B_3	m_5	m_4	m_4
$\mathbf{B_4}$	$\mathbf{m_5}$	$\mathbf{m_4}$	$\mathbf{m_6}$
$\mathbf{B_6}$	$\mathbf{m_6}$	$\mathbf{m_5}$	$\mathbf{m_2}$
$\mathbf{B_{10}}$	$\mathbf{m_4}$	$\mathbf{m_6}$	$\mathbf{m_5}$
$\mathbf{B_{13}}$	$\mathbf{m_5}$	$\mathbf{m_4}$	$\mathbf{m_2}$
$\mathbf{B_{17}}$	$\mathbf{m_2}$	$\mathbf{m_4}$	$\mathbf{m_5}$
$\mathbf{B_{19}}$	$\mathbf{m_4}$	$\mathbf{m_6}$	$\mathbf{m_2}$
Iteration 2			
B_2	m_5	m_6	m_6
$\mathbf{B_4}$	$\mathbf{m_5}$	$\mathbf{m_4}$	$\mathbf{m_6}$
$\mathbf{B_6}$	$\mathbf{m_5}$	$\mathbf{m_6}$	$\mathbf{m_2}$
$\mathbf{B_{10}}$	$\mathbf{m_4}$	$\mathbf{m_6}$	m_5
$\mathbf{B_{13}}$	$\mathbf{m_5}$	$\mathbf{m_4}$	$\mathbf{m_2}$
$\mathbf{B_{17}}$	$\mathbf{m_2}$	$\mathbf{m_4}$	$\mathbf{m_5}$
B_{20}	m_4	m_2	m_2
Iteration 3			
$\mathbf{B_4}$	$\mathbf{m_6}$	$\mathbf{m_5}$	$\mathbf{m_2}$
$\mathbf{B_6}$	$\mathbf{m_5}$	$\mathbf{m_6}$	$\mathbf{m_2}$
B_{10}	m_5	m_6	m_6
B_{13}	m_5	m_2	m_2
B_{17}	m_5	m_2	m_2

IV. DISCUSSION

The objective of the proposed work is to answer both questions mentioned at in Section I.

1) As a response to the first question and based on the results performed in the training phase, 25% of the images is sufficient for the selection of the binarization method, with 0% error for the BSB collection and about 9% error for the Google-Books collection. It is notable that such result is very promising because we didn't achieve the threshold number of documents used during the training phase of recognition systems (limited to 60%).
2) Based on the tests of the whole concept, it can be concluded that the choice of the method and its parameters for each book is efficient. 90% of the selected methods as the best binarization applied on the rest of images γ_k of each book from the BSB collection are present between the methods selected during the training. 83% of the selected methods as the best binarization applied on the rest of images γ_k of each book from the Google-Books collection are present between the method selected during the training.

V. CONCLUSION

In this work an automatic preprocessing system for historical documents is proposed. Two phases are adopted, the

Table IV
EVALUATION OF THE SELECTED METHODS ON A NEW SUBSET OF DOCUMENTS FROM BSB AND GOOGLE-BOOKS COLLECTIONS

		FM	$P-FM$	$PSNR$	NRM	MPM	GA	ρ
		(%)	(%)		$\cdot 10^{-2}$	$\cdot 10^{-3}$	$\cdot 10^{2}$	$\cdot 10^{2}$
BSB	1^{st} method	86.48	86.58	16.43	5.48	3.32	92.09	85.49
	2^{nd} method	86.58	86.54	16.18	5.51	4.79	92.95	85.36
	3^{rd} method	**88.65**	**89.05**	**16.94**	**5.30**	**2.74**	**93.97**	**87.49**
Google-Books	1^{st} method	86.72	87.56	16.57	3.14	**3.88**	95.59	86.13
	2^{nd} method	85.61	86.12	16.1	**2.5**	6.97	96.07	85.1
	3^{rd} method	**89.55**	**90.91**	**17.81**	3.42	6.32	**95.63**	**88.83**

training and the validation. According to the results found during our experiments it is notable that the training is well evaluated. The selected binarization and its input parameters as the best method for a limited subset of images is classified from the best binarization methods for the rest of images from the same book with a minimum error. Such work can be generalized on larger collection of historical books.

ACKNOWLEDGMENT

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

REFERENCES

[1] H. El Abed and V. Märgner, "ICDAR 2009-arabic handwriting recognition competition," *IJDAR*, vol. 14, no. 1, pp. 3–13, March 2011.

[2] A. Trier and T. Taxt, "Evaluation of binarization methods for document images," *T-PAMI*, vol. 17, no. 3, pp. 312–315, March 1995.

[3] M. A. Ramirez-Ortegon, E. A. Duenez-Guzman, R. Rojas, and E. Cuevas, "Unsupervised measures for parameter selection of binarization algorithms," *Pattern Recognition*, vol. 44, no. 3, pp. 492–501, 2011.

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

[5] E. Saund, J. Lind, and P. S. and, "Pixlabeler: User interface for pixel-level labeling of elements in document images," in *ICDAR*, Barchelona, Spain, July 2009, pp. 646–650.

[6] K. Ntirogiannis, B. Gatos, and I. Pratikakis, "An objective evaluation methodology for document image binarization techniques," in *DAS*, Nara, Japan, September 2008, pp. 217–224.

[7] B. Gatos, K. Ntirogiannis, and I. Pratikakis, "ICDAR 2009 document image binarization contest (DIBCO 2009)," in *ICDAR*, Barchelona, Spain, July 2009, pp. 1375–1382.

[8] I. Ben Messaoud, H. El Abed, H. Amiri, and V. Märgner, "New method for the selection of binarization parameters based on noise features of historical document," in *Joint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data (J-MOCR-AND)*, Beijin, China, September 2011, pp. 3–10.

[9] I. Pratikakis, B. Gatos, and K. Ntirogiannis, "H-DIBCO 2010-Handwritten Document Image Binarization Competition," in *ICFHR*, Kalkutta, India, November 2010, pp. 727–732.

[10] I. Ben Messaoud, H. El Abed, H. Amiri, and V. Märgner, "New binarization approach based on text block extraction," in *ICDAR*, Beijin, China, September 2011, pp. 1205–1209.

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

[12] M. Stommel and G. Frieder, "Automatic estimation of the legibility of binarised historic documents for unsupervised parameter tuning," in *ICDAR*, Beijin, China, September 2011, pp. 104 –108.

[13] R. D. Lins, S. Banergee, and M. Thielo, "Automatically detecting and classifying noises in document images," in *ACM Symposium on Applied Computing*, March 2010, pp. 33–39.

[14] I. Ben Messaoud, H. El Abed, H. Amiri, and V. Märgner, "A design of a preprocessing framework for large database of historical documents," in *Historical Document Imaging and Processing*, Beijin, China, September 2011, pp. 177–183.

[15] E. Barney Smith, "An anlysis of binarization ground truth," in *DAS*, Boston, Massachusetts, USA, June 2010, pp. 27–34.

[16] N. Otsu, "A threshold selection method from gray level histograms," *SMC*, vol. 9, pp. 62–66, 1979.

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

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

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

[20] B. Gatos, I. Pratikakis, and S. Perantonis, "Adaptive degraded document image binarization," *Pattern Recognition*, vol. 39, pp. 317–327, September 2006.

[21] R. Schilling, *Fundamentals of Robotics Analysis and Control*, E. Cliffs, Ed. Prentice-Hall, 1990.