Region Based Local Binarization Approach for Handwritten Ancient Documents

Ines Ben Messaoud, Hamid Amiri
Laboratoire de Recherche
Signale Image et Traitement de l'Information (LR-SITI)
ENIT
Tunis, Tunisia
ibmnoussa@gmail.com, hamidamiri@yahoo.com

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 fact that historical handwritten documents present many degradations, pre-processing of such documents is considered as a big challenge. Most pre-processing methods and specifically binarization return better results when they are applied on printed documents. We present in this paper a binarization approach adaptive for handwritten historical documents based on extraction of regions-of-interest. During our tests several images datasets are used, the benchmarking datasets for binarization DIBCO 2009 and H-DIBCO 2010 (15 images) as well as complete handwritten documents from the IAM historical database (about 60 images). The evaluation of the proposed binarization method is based on several evaluation metrics for binarization. The results show that the proposed method fit with handwritten historical documents ($FM \approx 88\%$) for images of the binarization competitions.**

***Keywords*-Binarization; regions-of-interest; evaluation metrics;**

I. Introduction

Following the scan of several historical handwritten books in different libraries as the Library of Congress, the Göttingen State and University library, and the Bayerische Staatsbibliothek, many works have been proposed in the field of handwritten documents processing. Most document analysis and recognition systems have been tested on printed documents first due to the fact that the characteristics of such documents are easier to extract. Such systems have been adapted to be used with handwritten historical documents subsequently due to the fact that these documents present much degradations.

In order to extract information from handwritten documents, many steps are performed. Pre-processing is applied of noise removal and binarization. Page layout analysis describes the physical structure of the document. Segmentation is the step which extracts text-lines, words or subwords. The determination of objects' attributes is known as feature extraction, while the classification of such features into different classes is the classification step.

Due to the presence of many degradations in historical handwritten documents, pre-processing of such documents is considered as a big challenge. Binarization is considered as the crucial step of pre-processing due to the reason that its output is the input of the succeeded steps. Several binarization methods have been first proposed for printed documents and afterwards adapted for handwritten documents. While some binarization methods are applied on color images [1], the majority take as input a gray-scale image [2]. Binarization methods can be classified according to the technique used for thresholding. Edge-detection-based methods use the edge pixels for the calculation of the local threshold [3]. Features-extraction-based methods use several characteristics of the image as stroke-width to classify pixels as background or foreground [4]. Statistical-based binarization includes methods where the adaptive threshold is estimated according to the specific pixel neighborhoods using statistical models as Hidden Markov Model (HMM) [5].

Due to the reason that many binarization methods have been proposed, the challenge is how to compare between such methods and how to select the most appropriate method for a specific input image. Methods for binarization evaluation have been improved from visually, to semi-automatically and automatically. Visual comparison was performed by an expert who judges the quality of the binary image, such way of evaluation is time consuming and lacks precision [6]. Semi-automatic evaluation combines both principles manually and automatically and it is applied on parts of images only [7]. Automatic evaluation is based on the former and it is applied on complete images of historical documents [8].

The paper is organized as follows. Section II presents a description of the proposed approach for binarization in detail. In Section III the evaluation method for binarization is explained. Section IV describes the realized tests and the achieved results. In Section V the discussion is drawn. Conclusions and future works are presented in Section VI.

II. Overview of the Proposed Method for Binarization

The proposed approach for binarization is an adaption of our method [9] to handwritten historical documents. Our method was classified 4th during the binarization competition DIBCO 2011 [10]. The description of the proposed approach for binarization of handwritten documents is shown in Figure

978-0-7695-4774-9/12 $26.00 © 2012 IEEE
DOI 10.1109/ICFHR.2012.261

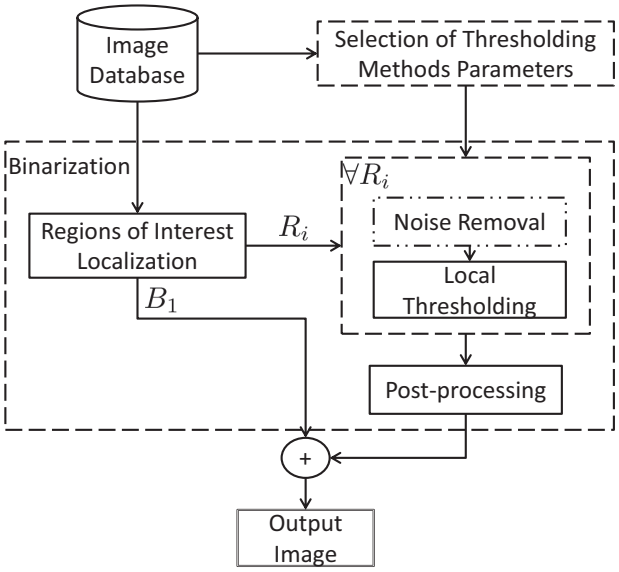


Figure 1. Overview of the proposed adaptive binarization approach for handwritten historical documents

1. The input image is transformed to gray-scale in the case if it is a color image. We denote with Ig the gray-scale input image for the proposed approach of binarization. We consider p an image pixel and (x,y) its coordinates in the image, where $x \in \{1,\ldots,Nx\}$ and $y \in \{1,\ldots,Ny\}$. In our work $\mathscr{B}$ and $\mathscr{F}$ refer to the background and the foreground respectively. First a pre-classification of each pixel is performed, a pixel p is first classified as $\{R_i\}$ or B_1, where R_i denotes a region-of-interest and $B_1 = Ig \setminus \{R_i\}$. $\{R_i\}$ is a set of gray-scale regions containing information ($\mathscr{F}$ pixels). A local thresholding method is applied only on each R_i. The final background is calculated using Equation 1.

$$\mathscr{B} = B_1 \sqcup B_2 \tag{1}$$

Where B_2 is the set of pixels classified as background after the binarization of all R_i, and $i \in \{1,\ldots,N\}$. The method for the detection of regions-of-interest is described in the following section. In order to adapt our binarization approach for handwritten historical documents, a new method for the detection of R_i as well as new thresholding methods are performed for the classification of R_i pixels.

A. Regions-of-interest Detection

We have used in [9] the Canny's edge detection method in order to detect R_i, in this work we adapt a new method for the selection of R_i based on the detection of connected components. We consider S_i the set of connected components grouped together according to y axis, where $i \in \{1,\ldots,N\}$. Each S_i is characterized with its position n_i according to the $y-$axis. First the connected components are labeled and Nb is the number of connected components. A Gaussian low pass filter is applied as a smoothing algorithm (Equation 2).

$$G(y) = \frac{1}{\sigma\sqrt{2\pi}} exp^{-\frac{\left(y^2-\mu\right)^2}{2\sigma^2}} \tag{2}$$

The size of the filter depends on the estimated height of S_i. σ and μ denote the standard deviation and the mean value of the Gaussian filter. The position of the gravity center of each connected component according to the $y-$axis is denoted with g_j, where $j \in \{1,\ldots,Nb\}$. The positions n_i are equal to the local maximums of the positions of the gravity-centers g_j. Each connected component is associated to a specific $S_{i'}$, $i' \in \{1,\ldots,N\}$, if the condition shown in Equation 3 is satisfied.

$$\left\|g_j, n_{i'}\right\| = min(\left\|g_j, n_i\right\|),\ i,i' \in \{1,\ldots,N\} \tag{3}$$

All Nb connected components are associated to a specific set S_i. A region-of-interest R_i is equal to the polygon formed by the set of connected components S_i.

B. Thresholding Methods

In the first step the gray-scale regions-of-interest, which are the input of the thresholding methods, are detected. In this section we present the methods used in order to classify the pixels of R_i either as $\mathscr{F}$ or B_2. A Wiener filter is applied on all R_i in order to remove noise pixels. In our previous work [9], we have tested two noise removal filters, shading correction and Wiener filter. According to the binarization results reported in [9], the Wiener filter has given the best results. For that reason the latter is applied as noise removal filter. Due to the reason that the proposed binarization approach is adaptive for handwritten historical documents, we have used two local thresholding methods well evaluated for this type of documents. First we have proved in previous tests presented in [8] that a modified version of Bernsen [11] gives good results when it is applied on handwritten historical documents. The second thresholding method is based on the one proposed by Su et al. [12]. The latter was ranked 1st during the last handwritten document image binarization competition held at ICFHR 2010 [13]. Both used thresholding methods are applied locally on each pixel in a specific neighborhood limited within a window centered in p. We have used square windows specified with their width w, which are chosen as odd values. We denote with Rb_i the output of the local thresholding method for the input gray-scale region R_i.

1) Bernsen's Method: The local threshold proposed by Bernsen is applied for high contrast pixels and calculated by Equation 4.

$$t(p) = \frac{max(R_i(X,Y)) + min(R_i(X,Y))}{2} \tag{4}$$

Where $X = x+k$, $Y = y+k$ and $k \in \left\{-\frac{w-1}{2}, \ldots, \frac{w-1}{2}\right\}$. High contrast pixels must satisfy the condition shown in Equation 5.

$$t(p) = max(R_i(X,Y)) - min(R_i(X,Y)) > v \quad (5)$$

Where v is a threshold given as input parameter. For the pixels which are not classified as high contrast pixels, $t(p)$ is equal to the global threshold calculated by Otsu's method [14].

2) Su's Method: The binary region Rb_i, which is the output of Su's method, is extracted from the high contrast image Ih calculated as it is shown in Equation 6

$$Ih = 1 - (Ih_1 + Ih_2) \quad (6)$$

Where Ih_1 is the high contrast image, which is the result of the application of Otsu's method on the contrast image Is (Equation 7). Ih_2 is the high contrast image obtained after the application of Canny's edge detector to R_i.

$$Is(x,y) = \frac{max(R_i(X,Y)) - min(R_i(X,Y))}{max(R_i(X,Y)) + min(R_i(X,Y)) + \varepsilon} \quad (7)$$

Where ε is an infinity small number. X, Y, and k are defined as in Equation 4. The binary region Rb_i is calculated by Equation 8.

$$Rb_i(x,y) = \begin{cases} 0 & \text{, if } N_e \geq min \text{ and} \\ & R_i(x,y) \leq K_{mean}(x,y) + K_{std}(x,y) \\ 1 & \text{, otherwise} \end{cases} \quad (8)$$

The binary region K_{mean} and K_{std} are calculated using Equations 9 and 10. Ne is the number of high contrast pixels in the neighborhoods of p calculated in the image Ih, and min is an input parameter.

$$K_{mean}(x,y) = \frac{\sum_{k=-\frac{w-1}{2}}^{\frac{w-1}{2}} R_i(X,Y) \cdot Ih(X,Y)}{Ne} \quad (9)$$

$$K_{std}(x,y) = \sqrt{\frac{\sum_{k=-\frac{w-1}{2}}^{\frac{w-1}{2}} (R_i(X,Y) - K_{mean}(X,Y)) \cdot Ih(X,Y)^2}{Ne}} \quad (10)$$

Where $X = x+k$, $Y = y+k$ and $k \in \left\{-\frac{w-1}{2}, \ldots, \frac{w-1}{2}\right\}$ in both Equations 9 and 10.

3) Post-processing: After the classification of R_i pixels as $\mathscr{F}$ or B_2, a post-processing method is applied. The connected components, having an area less than 10% of the average area of the Nb connected components, are removed.

C. Selection of Thresholding Methods' Parameters

The proposed approach for binarization is parameterless. For that reason in order to estimate the input values of the thresholding methods, a method for the selection of parameters proposed in [15] is used. This method classifies the input image Ig into one class C_l, $l \in \{1, \ldots, 4\}$. C_l depends on the detected noise type in Ig, C_1: images with show-through, C_2: images presenting variable background, C_3: images where the similarity between background and foreground pixels is very high, C_4: images without noise. For each class a set of parameters is selected for each thresholding method used during our work.

III. Evaluation of Binarization Methods

The evaluation of binarization methods' performance is based on the comparison between the binary image Ib and its corresponding ground-truth noted with GT. While some databases also contain the corresponding ground-truth, the most datasets don't include them. It has been proved in [16] that there are differences between ground-truth images either calculated manually or semi-automatically, for that reason a method for the generation of ground-truth images proposed in [8] is used. The comparison between both binary images is based on the number of pixels classified as true positive ($GT = 1$ and $Ib = 1$), false positive ($GT = 0$ and $Ib = 1$), false negative ($GT = 1$ and $Ib = 0$) and true negative ($GT = 0$ and $Ib = 0$). According to these values several evaluation metrics are calculated. F-measure (FM), Peak Signal-to-Noise Ratio ($PSNR$), Negative Rate Metric (NRM), and Miss-classification Penalty Metric (MPM) proposed first in [17]. Geometric mean Accuracy (GA) used as the only evaluation metric during [18], and Normalized Cross Correlation (ρ) proposed in [16].

IV. Experiments and Results

In this section we describe the used datasets during our tests as well as the experimental setup and the achieved results. During our tests three different images' datasets are used. The benchmarking datasets DIBCO 2009 and H-DIBCO 2010 containing 5 and 10 handwritten images respectively. These images belong to the library of congress[1] and the Göttingen State and University Library and present much degradation as show-through and variable background. The third images' dataset is composed of images from the IAM historical database (IAM-HistDB) described in [19]. The IAM-HistDB contains about 60 images and transcriptions of handwritten Latin documents from the 9th century written in Carolingian script.

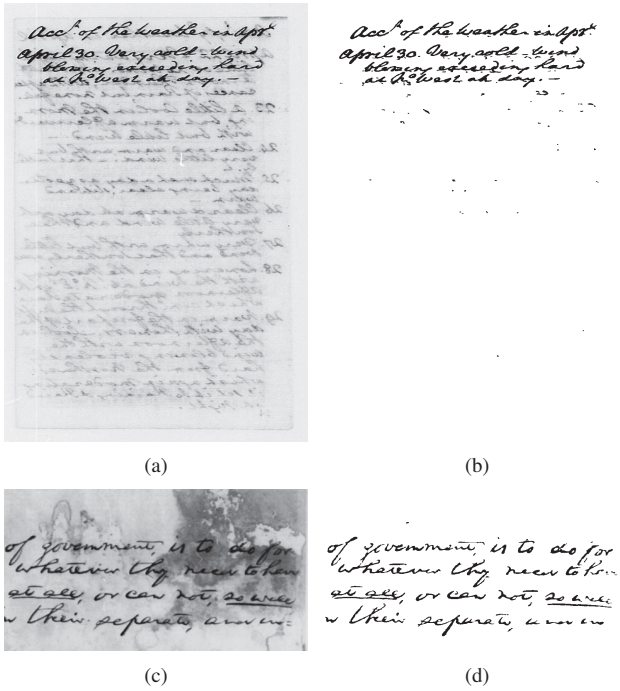
A. Results

In the first test, the proposed approach for binarization is applied on images from the DIBCO 2009 and the H-DIBCO 2010 datasets. The comparison between our approach and the state-of-the-art methods is shown in Tables I and II respectively. This comparison is based on 4 evaluation metrics FM, $PSNR$, NRM, and MPM. The results of the proposed approach are compared to those given by the methods participated to the binarization competitions. The proposed approach is ranked 1st when it is applied on the DIBCO 2009

[1] http://www.loc.gov/library/libarch-digital.html

Table I
COMPARISON BETWEEN THE PROPOSED APPROACH FOR BINARIZATION AND THE BEST RANKED METHODS IN DIBCO 2009 COMPETITION APPLIED ON HANDWRITTEN IMAGES FROM THE DIBCO 2009

	FM (%)	*PSNR*	*NRM* ($\cdot 10^{-2}$)	*MPM* ($\cdot 10^{-3}$)
1 st	88.65	19.42	**5.11**	0.34
2 nd	86.02	18.57	6.39	0.95
Ben Messaoud 2011	88.12	19.33	8.39	**0.31**
Proposed Method	**88.70**	**19.43**	5.87	1.02



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

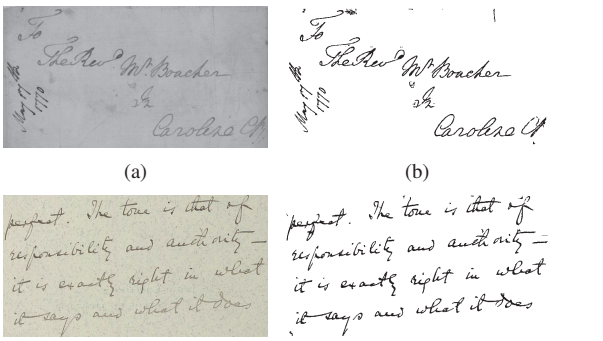
Figure 2. (a) and (c): Samples of handwritten images from the DIBCO 2009 dataset, (b) and (d): the corresponding binary images given with the proposed binarization approach

handwritten images, and 5th when it is applied on H-DIBCO 2010 dataset [13]. Figures 2 and 3 show samples of the DIBCO 2009 and the H-DIBCO 2010 datasets, respectively, binarized using our approach for binarization.

In the second experiment the proposed approach is applied on complete handwritten images, which belong to the IAM-HistDB database. The ground-truth images of such images

Table II
COMPARISON BETWEEN THE PROPOSED APPROACH FOR BINARIZATION AND THE BEST RANKED METHODS IN H-DIBCO 2010 COMPETITION APPLIED ON HANDWRITTEN IMAGES FROM THE H-DIBCO 2010

	FM (%)	*PSNR*	*NRM* ($\cdot 10^{-2}$)	*MPM* ($\cdot 10^{-3}$)
1st	**91,50**	**19.78**	**5.98**	0.49
1st	89.70	19.15	8.18	**0.29**
Ben Messaoud 2011	86.33	18.03	10.76	0.38
Proposed Method	87.88	18.28	8.39	0.79



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

Figure 3. (a) and (c): Samples of handwritten images from the H-DIBCO 2010 dataset, (b) and (d): the corresponding binary images given with the proposed binarization approach

were realized using our method for ground-truth generation [8]. The comparison of the proposed method for binarization with the state-of-the-art methods is described in Table III. The comparison is based on the evaluation metrics *FM*, *PSNR*, *NRM* , *MPM*, *GA*, and ρ. Based on the results shown in Table III it can be concluded that the proposed method for binarization gives on average the best values of evaluation metrics *FM*, *PSNR*. Gatos's method gives the best values of *MPM* and *GA* and Sauvola's method returns the best value of *NRM*. Figure 4 shows sample of the IAM-HistDB dataset binarized using our approach for binarization.

V. DISCUSSION

Different ground-truth images were realized and different state-of-the-art binarization methods have been compared according to evaluation metrics in [16]. While some evaluation metrics as *FM*, *PSNR*, and ρ gave a lower value of standard deviation, the *NRM* returned a high standard deviation. Therefore *FM*, *PSNR* and ρ are considered as better evaluation metrics for binarization than *NRM*. If we consider this fact and we compare the state-of-the-art methods based on *FM* and *PSNR* during DIBCO 2009 and H-DIBCO 2010 the proposed method is classified 1st and 4th respectively.

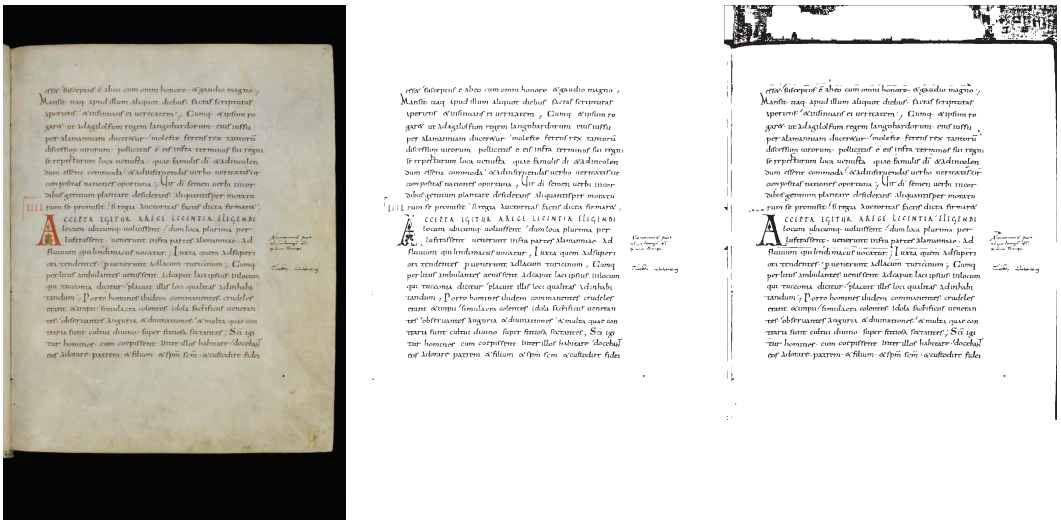
In the experiment applied on IAM-HistDB the proposed approach gives the best results according to the evaluation metrics *FM*, *PSNR* and ρ, which are considered as good evaluation metrics for binarization. Otsu's method gives on average the lowest results, because this method shows weakness when it deals with images having variable background as it is shown in Figure 4(a).

VI. CONCLUSIONS

We have proposed in this work a parameterless framework for binarization adapted for handwritten historical documents. This framework is based on the detection of regions-of-interest. A method for the detection of regions-of-interest according to the connect-components position is

Table III
COMPARISON BETWEEN THE PROPOSED APPROACH FOR BINARIZATION AND THE STATE-OF-THE-ART METHODS APPLIED ON HANDWRITTEN IMAGES FROM THE IAM-HISTDB

	FM (%)	*PSNR*	*NRM* ($\cdot 10^{-2}$)	*MPM* ($\cdot 10^{-3}$)	*GA* ($\cdot 10^{2}$)	ρ ($\cdot 10^{2}$)
Otsu	18.28	5.95	28.03	212.28	62.07	19.62
Bernsen	24.01	6.09	15.46	212.37	72.48	30.42
Niblack	28.44	7.01	12.23	123.33	77.86	35.25
Sauvola	60.27	12.76	**3.66**	24.12	93.63	63.44
Gatos	58.45	12.43	3.97	**6.45**	**93.14**	61.85
Proposed Method	**69.32**	**15.50**	10.54	9.57	87.92	**68.93**



(a) (b) (c)

Figure 4. (a): Sample of the IAM-HistDB database, (b): the corresponding ground-truth, (c): the binary image using the proposed method for binarization

applied. The thresholding methods used during our work are well evaluated for handwritten historical documents. The binarization results returned using the proposed framework are very promising. This framework will be extended by the improvement of the method for parameters selection.

ACKNOWLEDGMENTS

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

REFERENCES

[1] C. Tsai and H. Lee, "Efficiently extracting and classifying objects for analyzing color documents," *Machine Vision and Applications*, vol. 22, no. 1, pp. 1–19, January 2011.

[2] B. Gatos, I. Pratikakis, and S. Perantonis, "Adaptive degraded document image binarization," *Pattern Recognition*, vol. 39, no. 3, pp. 317–327, September 2006.

[3] 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.

[4] M. Valizadeh and E. Kabir, "Binarization of degraded document image based on feature space partitioning and classification," *IJDAR*, vol. 15, no. 1, pp. 57–69, February 2010.

[5] S. Huang, M. Sid-Ahmed, M. Ahmadi, and I. El-Feghi, "A binarization method for scanned documents based on hidden Markov model," in *IEEE International Symposium on Circuits and Systems*, Island of Kos, May 2006, pp. 4309–4312.

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

[7] 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.

[8] 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*, Beijing, China, September 2011, pp. 177–183.

[9] ——, "New binarization approach based on text block extraction," in *ICDAR*, Beijing, China, September 2011, pp. 1205–1209.

[10] I. Pratikakis, B. Gatos, and K. Ntirogiannis, "ICDAR 2011 document image bianrization contest (DIBCO 2011)," in *ICDAR*, Beijing, China, September 2011, pp. 1506–1510.

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

[12] B. Su, S. Lu, and C. Tan, "Binarization of historical document images using the local maximum and minimum," in *DAS*, Boston, Massachusetts, USA, June 2010, pp. 159–165.

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

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

[15] 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)*, Beijing, China, September 2011, pp. 3–10.

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

[17] B. Gatos, K. Ntirogiannis, and I. Pratikakis, "DIBCO 2009: document image binarization contest," *IJDAR*, vol. 14, no. 14, pp. 35–44, May 2011.

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

[19] A. Fischer, V. Frinken, A. Fornés, and H. Bunke, "Transcription alignment of latin manuscripts using hidden Markov models," in *Historical Document Imaging and Processing*, Beijing, China, September 2011, pp. 29–36.