2012 International Conference on Frontiers in Handwriting Recognition

A Multilevel Text line Segmentation Framework for Handwritten Historical Documents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*—Text-line segmentation is considered as a crucial step of document analysis and recognition systems because its output is considered as the input of recognition systems. Due to the reason that the same handwritten image page has different characteristics, we propose in this paper a multilevel segmentation framework for handwritten historical documents. In this framework, one or many segmentation methods are selected according to the input document features. This framework is tested on the IAM historical database (60 images) and on images from the segmentation competition for handwritten document segmentation held at ICFHR 2010. The evaluation of the segmentation framework is based on several evaluation metrics. The tests show that the proposed framework gives promoting results.**

***Keywords*-Text line segmentation; evaluation metrics; text line features;**

I. Introduction

Several libraries have a huge amount of scanned historical documents. Recognition of manuscripts is essential for efficient automatic document exploitation. Text-line segmentation is a crucial step of document analysis systems, because its output is the input of recognition steps. Text-line segmentation for historical handwritten documents is considered as a big challenge, due to the fact that such documents are different from simple documents as newspapers, mails or cheques. Handwritten historical documents have many degradations as overlapping or touching components, smears, and holes which affect document analysis systems. Several segmentation methods have been presented in the literature and they have generally as input binary images [1]. These methods can be classified as top-down, bottom-up or hybrid methods. A survey of text-line segmentation for historical documents has been presented in [2]. Projection-based approaches use image histogram [3]. Smearing methods as RLSA (Run Length Smoothing Algorithm) transform the binary image to a smeared image, where text-lines are the bounding boxes of the connected components [4]. In KNN-based approaches, a pixel is classified according to its pixel neighbors [5]. Grouping methods are based on pixels or connected components features. These elements are joined together according to some features to form text-lines image [6]. Segmentation methods based on shredding document image using paths between two consecutive text-lines according to the blurred image [7].

Historical handwritten pages present different characteristics, e.g, some text-lines are horizontal while the others are not. Proposed methods for segmentation use the same approach for different documents' types, however it is difficult to find one approach which fits with different types of document. For that reason it is important to select one efficient method for each type of document [8].

In order to compare the performance of the segmentation methods, several contests for handwritten segmentation have been proposed [9], [10]. During these competitions only one evaluation metric was used. However several metrics were used in order to evaluate methods for page layout analysis [11], [12]. These methods rely on ground-truth images which are sometimes difficult to achieve. In order to evaluate text-line segmentation methods it is necessary to use different evaluation metrics and not only one metric.

This paper is organized as follows. Next section presents a description of the proposed framework for handwritten document segmentation. In section III the experimental results and a discussion are drawn. section IV presents the conclusion and future works.

II. Proposed Framework

Different segmentation methods have been proposed in the literature, most of these methods give very good results on some images and show weakness on other image types. This explains that for each type of document there is a specific method which fits with such documents. For that reason segmentation methods must be chosen according to the input document features. Due to the fact that in the same page of handwritten historical document image different features can be detected, we propose in this work a framework for segmentation which combines different segmentation methods for the same input image page. The proposed framework as shown in Figure 1 is multilevel and based

978-0-7695-4774-9/12 $26.00 © 2012 IEEE
DOI 10.1109/ICFHR.2012.159

on recursive application of different segmentation methods for text-line segmentation described in Section II-B.

During the proposed process, one or many segmentation methods are applied. First step is a pre-processing method, during which two features of the input image text-lines are extracted. According to these both features one method is selected. The method m_{a+1} is applied if the method m_a has failed. Regions where a segmentation method gives wrong results are called regions-of-problems. In order to check whether the selected method gives wrong text-lines, a method for error verification and regions-of-problems selection is developed. In this work Ib denotes the binary input image of our framework, p is an image pixel and (x,y) its coordinates, where $1 \leq x \leq Nx$ and $1 \leq y \leq Ny$. S denotes the result of the segmentation and S_i is the text-line having the label i, where $i \in \{1,\ldots,Nr\}$ and Nr is the number of the extracted lines

A. Pre-processing

According to the related works of segmentation methods, It is notable that two features are important in order to choose a text-line segmentation method.

- Line touching: Text-lines are considered as separated lines or as lines containing touching or overlapping components.
- Line orientation: Text-lines are considered horizontal or skewed.

1) RLSA Method: This method is used in order to check whether the input image text-lines are separated. The RLSA method for segmentation is applied horizontally on the input binary image Ib. We suppose that foreground pixels (black pixels) are equal to 1 and background pixels (white pixels) are equal to 0. If the number of adjacent pixels from the background is equal to or higher than a value C, these pixels are classified as foreground. The smeared image Is is composed of blocks supposed to be text-lines. We suppose LM the mean of the height of black blocks in Is. If $LM < v$, where v is an input value, the text-lines are considered separated, otherwise they are considered touched or overlapped. In our method the parameters C and v are set to 300 and 50 respectively.

2) Skew Angle Correction: The input of the proposed framework are images containing only text-blocks. In order to calculate θ automatically, a set of skew angles θ_l are tested, such as $\theta_{l+1} = \theta_l + \alpha$ based on the method proposed in [13], where $\theta_l \in [-2,\ldots,2]$ and α is an input parameter. For each θ_l, the input image is rotated θ_l degree, the sum of foreground pixel for each y position is calculated and the number of interlines is estimated. Interlines are considered horizontal y positions having a limited number of foreground pixels. θ is equal to θ_l which gives the maximum number of interlines. θ is estimated in degree. After θ estimation, Ib is rotated using the bilinear interpolation. The rotated image is bigger than Ib, in order to contain all pixels of the rotated image. In order to retrieve pixels' positions of the original image, the result of the segmentation is rotated with $-\theta$ degree and it is resized to the same dimensions of the original image ($Nx \times Ny$).

B. Segmentation Methods

Three different methods for text-line segmentation are performed. The first one is a bottom-up method based on vertical projection. The second method is a top-down method based on the estimation of text-lines using the grouping of connected components. The third method is based on the detection of the nearest neighbor. While the first two methods are applied on the entire input image Ib, the third method is applied only on regions-of-problems. Regions-of-problems are formed using text-lines containing error(s) detected with the method described in Section II-C.

1) Projection-Based Method: In this method (m_1) the histogram H of the binary image is used. For each y position in Ib, foreground pixels are summed as it is shown in Equation 1.

$$H(y) = \sum_{x=1}^{Nx} Ib(x,y),\ \forall y \in \{1,\ldots,Ny\} \tag{1}$$

y positions, where $H(y)$ is very low or equal to 0 are considered as interlines, the remainders are considered text-line positions. The segmentation m_1 is considered as a bottom-up method and gives good results in the case where text-lines are horizontal and don't present any touching or overlapping connected components.

2) Text-line Estimation Method Based on Grouping of Connected Components: This segmentation (m_2) is a top-down method where text-lines are estimated based on the grouping of connected components o_k according to the y axis, where $1 \leq k \leq N$. The area a_k as well as the gravity centers of each connected component o_k are calculated. The o_k with very huge or a very small areas are removed. In order to estimate the positions of text-lines, the local means pos_i of the positions of the o_k gravity-centers are calculated. These local means are considered text-lines positions. Each o_k is afterwards associated to each text-line (pos_i). A connected component is associated to a specific text-line ($pos_{i'}$) if the condition calculated by Equation 2 is satisfied.

$$\ni i',\text{ such as } \|pos_{i'},o_k\| = min(\|pos_i,o_k\|)\forall i, 1 \leq i \leq Nr \tag{2}$$

Where Nr is the number of detected text-lines.

3) Nearest Neighbor: This method has as input regions-of-problems pr_b. This method is applied to correct overlapping and touching problems. For each text-line S_i, which contains error(s), a region-of-problems pr_b is associated. pr_b contains generally more than one text-line as it is explain in the next section. We select in S_i, which contains error(s), the connected components $o_{k'}$, such as $o_{k'}$ heights are bigger than 0.9 of the maximum height of all the

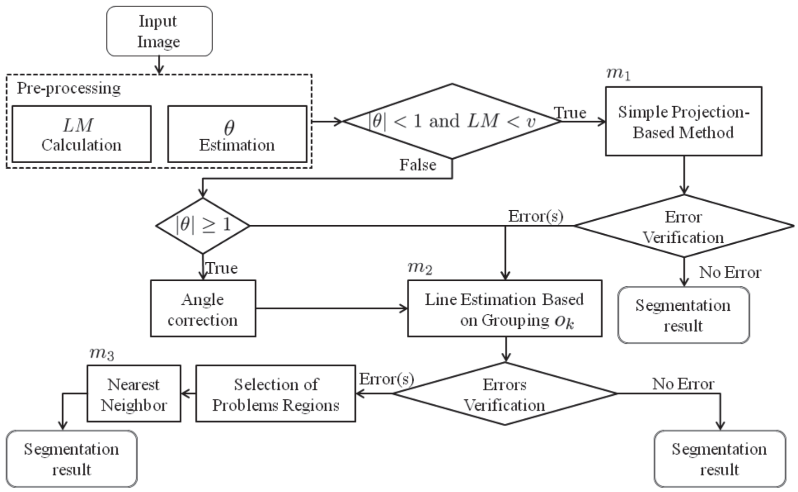

Figure 1. Overview of the proposed multilevel framework for text-line segmentation.

connected components in S_i. For each pixel in $o_{k'}$, $N^{\circ}(p)$ is calculated, where $N^{\circ}(p)$ is the nearest neighbor in pr_b to p and $N^{\circ}(p) \notin o_{k'}$. $N^{\circ}(p)$ is calculated in Equation 3.

$$\|N^{\circ}(p), p\| = min(p', p) \tag{3}$$

Where $p' \in pr_b \setminus o_{k'}$. As result each pixel $p \in o'_k$ takes the same label as its $N^{\circ}(p)$.

C. Error Detection and Regions-of-Problems Selection

The line height LH of the text-line is estimated. The interline height LI is given by the first local maximum of the FFT spectrum (Fast Fourier Transform) applied to H (Equation 1). The estimated line height LH is calculated by Equation 4.

$$LH = \frac{Ny}{LI} \tag{4}$$

We call R_i the smallest box containing the text-line S_i. S_i is classified as text-line containing error(s) if the condition shown in Equation 5 is satisfied.

$$|R_i| \geq \gamma \times LH \tag{5}$$

Where γ is an input parameter such as $\gamma \in]0.1, \ldots, 0.9]$ and $|R_i|$ denotes the height of R_i.

We denote with Within(R_i, S_q), the function which returns pixels in R_i and belonging to another text-line S_q, where $S_q \neq S_i$. The region-of-problem pr_b containing S_i is defined as it is shown in Equation 6.

$$pr_b = \begin{cases} S_i \cup S_q & \text{, if } \frac{\#\text{Within}(R_i, S_q)}{\#S_i} > \beta \\ S_i & \text{, otherwise} \end{cases} \tag{6}$$

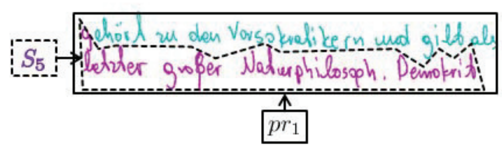

Figure 2. Sample of one detected region-of-problem of Figure 3(b): S_5 is a text-line containing error detected in the segmentation result shown in Figure 3(a) and the corresponding region-of-problem pr_1.

Where β is set to 0.8. Figure 2 shows a sample of a detected region-of-problem.

III. Experiments and Results

In this section we present the used image datasets as well as the experimental setup and the evaluation results. The evaluation of the segmentation methods is based on different evaluation metrics described in Section III-B.

A. Datasets

We have used for our tests the IAM historical database (IAM-HistDB) proposed in [14]. The IAM-HistDB includes about 60 images and transcriptions of handwritten Latin documents from the 9^{th} century written in Carolingian script. We have also used images from the ICFHR 2010 competition. These images present many degradations as overlapping and touching elements as well as skewed text-lines.

B. Segmentation Evaluation

The evaluation of the segmentation is based on the comparison between the labeled image S and the corresponding ground-truth G for segmentation. G_j denotes the j^{th} text-line in the ground-truth image, where $j \in \{1,\ldots,Ng\}$ and Ng is the number of text-lines in the input image. We have used during our work FM defined at [10] and the error rate U proposed in [11]

- F-measure (FM)
  It was defined in [9]. The match between each S_i and each G_j is denoted with $Matchscore(i, j)$, where $1 \leq i \leq Nr$ and $1 \leq j \leq Ng$. A text-line G_j is considered well retrieved by the segmentation method if it exists one S_i such as $Matchscore(i, j)$ is equal or higher than a threshold Ta fixed during our tests. We call No the number of the counted match as it is shown in Equation 7.
  $$No = \#Matchscore(i, j) >= Ta \tag{7}$$
  FM is calculated based on No and using Equation 8, where DR and RA are the detection rate and the recognition accuracy equal to $DR = \frac{No}{Ng}$ and $NA = \frac{No}{Nr}$ respectively.
  $$FM = \frac{2 \times DR \times RA}{DR + RA} \tag{8}$$
- Split regions (Ns)
  G_j is a split region if it coincides at minimum with two different S_i, such as the percentage of the G_j pixels in these S_i is equal or higher than Tb.
- Merged Regions (Nm)
  G_j is a merged region if it exists at minimum one ground-truth region $G_{j'}$, $j \neq j'$, coincides with the same S_i as G_j. The percentage of G_j and $G_{j'}$ in S_i has to be equal or higher than Tb.
- Missed Regions (Ne)
  G_j is a missed region if it does not coincide with any S_i.
- Partial-Missed Regions (Np)
  G_j is considered as partial missed region, if the percentage of detected pixels of G_j in S_i is less than Tb.
- Error Rate (U)
  U is the error rate calculated using Equation 9 according to Ns, Nm, Ne, and Np.
  $$U = \frac{\#(\psi_1 Ns + \psi_2 Nm + \psi_3 Ne + \psi_4 Np)}{\#Ng} \tag{9}$$
  where the weight values ψ_1, ψ_2, ψ_3, and ψ_4 are fixed during the tests.

Segmentation methods are ranked according to FM and U, the best segmentation is that one returning the highest FM and the lowest U.

C. Results and Discussions

For the experiments, we have tested different parameters of the evaluation metrics. Based on parameters used during the competition [10] and the comparison of segmentation methods [11], we have used the following parameters for the test, $Ta = 0.9$, $Tb = 0.4$, $\psi_1 = 0.2$, $\psi_2 = 0.2$, $\psi_3 = 0.4$, and $\psi_4 = 0.3$. The first experiment is applied to images from the IAM-HistDB database. This database includes the corresponding text-lines ground-truth, the binarization of the input images are realized using [15], [16]. In order to evaluate the proposed framework, the achieved results using our framework are compared to the results of the application of both methods for segmentation, the projection-based segmentation (m_1) and text-lines estimation method based on grouping connected components (m_2). Table I shows the results of the comparison between the three methods for segmentation using the evaluation metric FM and the error rate U. Most text-lines of the IAM-HistDB are horizontal ($\theta \in [0,\ldots,1]$) and does not present a big number of touching connected components, which explains that according to the pre-processing step of the proposed framework, the method m_1 is applied, when an error is detected by the proposed framework the method m_2 is applied. The proposed method gives a value of FM higher than the one given by m_1, that means that the choice of the segmentation method realized during our framework fits with the input image features and that when m_1 gives errors the method m_2 is applied during our framework. The error rate U values given by the proposed framework as well as the other methods for segmentation (m_1 and m_2) are very low which explains that the cardinality of the merge and split regions is very low.

Table I
EVALUATION OF THE TEXT-LINES SEGMENTATION METHODS ACCORDING TO FM AND U ON IAM-HISTDB DATABASE

	FM (%)	U ($\times 10^{-3}$)
Projection-Based Method	79%	**0.84**
Groupping Method	**86** %	3.8
Proposed Method	84 %	0.9

During the second experiment we have used some images from the ICFHR 2010 competition (about 250 handwritten text-lines). Due to the reason that we don't have access to ground-truth images for segmentation used during the handwritten segmentation contest held at ICFHR 2010, the evaluation is based on Nc, Ns, Nm, Ne and Np. Table II shows the results of the evaluation of the segmentation methods. According to these results, it is notable that the proposed framework returns better results than m_1 and m_2. Figure 3 shows samples of the different steps of the proposed framework applied to these images as well as the segmentation results. It is notable that the proposed framework retrieves most text-lines (comparing with m_1 and m_2) and the lower errors of merge and split text-lines. For the three segmentations compared methods $Ne = Np = 0$, because no text-lines are missed or partially missed. Figures 3(a) and 3(d) show the result of the first step after features extraction. Figures 3(b) and 3(e) show the detection of regions-of-problems, in the first one 4 regions are detected and in the second only one region is detected. Figures 3(c)

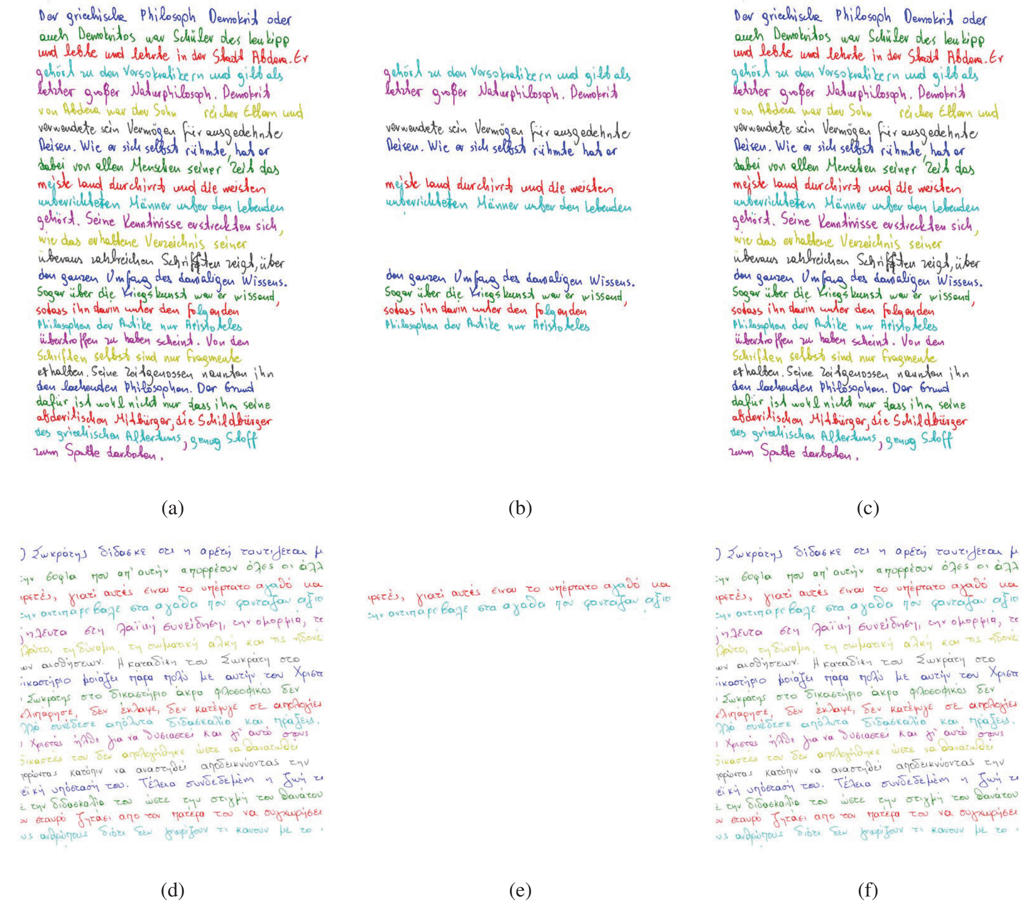

Figure 3. (a) and (d): Segmentation results after features extraction and pre-processing. (b) and (e): Regions-of-problems $pr_1, \dots, pr_4$ and pr_1 detected respectively in Figures (a) and (d). (c) and (f): Segmentation results of the proposed framework.

Table II
EVALUATION OF THE SEGMENTATION RESULTS APPLIED TO IMAGES FROM THE ICFHR 2010 HANDWRITTEN CONTEST

	Nc (%)	Ns (%)	Nm (%)
Projection-Based Method	0.81	0.11	0.11
Groupping Method	0.86	0.12	0.11
Proposed Framework	**0.97**	**0**	**0.01**

and 3(f) show the segmentation results given by the proposed framework.

IV. CONCLUSIONS

We have presented in this paper a new framework for segmentation based on the input document features. This framework fits with the handwritten historical documents as it is shown in the achieved results. Other features will be added to our work. New methods for text-line segmentation will be added during the selection and the evaluation will be performed according to each document type of the used databases.

ACKNOWLEDGMENTS

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

REFERENCES

[1] I. Ben Messaoud, H. El Abed, H. Amiri, and V. Märgner, "Collaborative access to ancient documents –towards a distributed comparison of pre-processing approaches," *International Journal of Mobile Computing and Multimedia Communications (IJMCMC)*, p. in press, 2012.

[2] L. Likforman-Sulem, A. Zahour, and B. Tacone, "Text line segmentation of historical documents: a survey," *International Journal on Document Analysis and Recognition*, vol. 9, no. 2-4, pp. 123–138, 2007.

[3] Z. Razak, K. Zulkiflee, and et al., "Off-line handwriting text line segmentation : A review," *International Journal of Computer Science and Network Security*, vol. 8, no. 7, pp. 12–20, 2008.

[4] N. Nikolaou, M. Makridis, B. Gatos, N. Stamatopoulos, and N. Papamarkos, "Segmentation of historical machine-printed documents using adaptive run length smoothing and skeleton segmentation paths," *Image Vesion Computer*, vol. 28, no. 4, pp. 590–604, April 2010.

[5] A. Zahour, B. Taconet, L. Likforman-Sulem, and W. Boussellaa, "Overlapping and multi-touching text-line segmentation by block covering analysis," *Pattern Analysis and Applications*, vol. 12, no. 4, pp. 335–351, 2010.

[6] M. Feldbach and K. D. Tönnies, "Line detection and segmentation in historical church registers," in *International Conference on Document Analysis and Recognition (ICDAR)*, Seattle, WA , USA, September 2001, pp. 743 –747.

[7] A. Nicolaou and B. Gatos, "Handwritten text line segmentation by shredding text into its lines," in *International Conference on Document Analysis and Recognition (ICDAR)*, Barcelona, Spain, July 2009, pp. 626– 630.

[8] I. Ben Messaoud, H. El Abed, H. Amiri, and V. Märgner, "Document preprocessing system – automatic selection of binarization," in *IAPR International Workshop on Document Analysis Systems (DAS)*, Gold Coast, Queensland, Australia, March 2012, pp. 85–89.

[9] B. Gatos, N. Stamatopoulos, and G. Louloudis, "ICDAR 2009 handwriting segmentation contest," *International Journal on Document Analysis and Recognition*, vol. 14, no. 1, pp. 25–33, March 2011.

[10] ——, "ICFHR 2010 handwriting segmentation contest," in *International Conference on Frontiers in Handwriting Recognition (ICFHR)*, Kalkutta, India, November 2010, pp. 737–742.

[11] F. Shafait, D. Keysers, and T. Breuel, "Performance evaluation and benchmarking of six-page segmentation algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 6, pp. 941–954, June 2008.

[12] A. Antonacopoulos, S. Pletschacher, D. Bridson, and C. Papadopoulos, "ICDAR 2009 page segmentation competition," in *International Conference on Document Analysis and Recognition (ICDAR)*, Barchelona, Spain, July 2009, pp. 1370–1374.

[13] S. Hinds and D. J.L Fisher, J.L.; D'Amato, "A document skew detection method using run-length encoding and the hough transform," in *International Conference on Pattern Recognition (ICPR)*, Atlantic Citys, USA, June 1990, pp. 464–468.

[14] 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.

[15] I. Ben Messaoud, H. El Abed, H. Amiri, and V. Märgner, "New binarization approach based on text block extraction," in *International Conference on Document Analysis and Recognition (ICDAR)*, Beijing, China, September 2011, pp. 1205–1209.

[16] ——, "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.