BLIND BLEED-THROUGH REMOVAL FOR SCANNED HISTORICAL DOCUMENT IMAGES WITH CONDITIONAL RANDOM FIELDS

BIN SUN, SHUTAO LI
College of Electrical and Information Engineering
Hunan University
Changsha, China 410082
Email: sunbinxs@126.com; shutao_li@hnu.cn

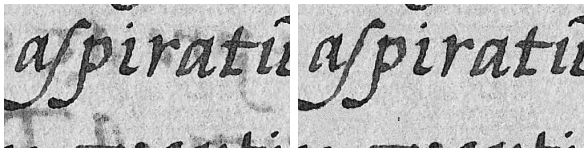
JUN SUN
Information Department
Fujitsu Research and Develop Center
Beijing, China 100025
Email: sunjun@cn.fujitsu.com

***Abstract*—Due to the quality of paper and long-time preservation, the ink on one side of the historical documents often seeps through and appears on the other side. In this paper, a new blind ink bleed-through removal method is proposed to deal with the scanned historical document images. The scanned historical document image generally consists of three components: foreground, bleed-through and background. In the proposed method, conditional probability distribution (CPD) models of the three components are firstly established by statistics. Then, conditional random fields (CRFs) are used to model the observed scanned image and the corresponding labels. For each input scanned image, parameters of the component-wise CPD models are estimated and belief propagation is performed on the CRFs model to determine the most possible labels. Once the bleed-through component is found, an inpainting algorithm is proposed to remove the ink bleed-through from the input historical image. Experimental results show that the proposed method preserves the foreground component very well and removes the bleed-through effectively.**

I. INTRODUCTION

Nowadays most historical documents are preserved by using a digital version. The paper pages are scanned to generate a picture of those historical documents, i.e., the scanned images of historical documents. Due to the paper quality and long preservation time, many double-sided historical documents suffer from ink bleed-through, which refers to that the ink seeped through the paper and reveals on the other side of the same page. This may make the historical documents hard to read manually or automatically. And it may also affect the aesthetic quality of some valuable manuscripts. Fig. 1 shows an example of the bleed-through removal.

To solve this problem, a variety of different methods are proposed to remove the ink bleed-through. The existing methods can be categorized into two categories: the blind methods [1], [2] and the non-blind methods [3], [4]. The non-blind bleed-through removal methods usually require two aligned images of both sides of one page [4], while the blind methods only take one. Since registration of the two scanned images of both sides is still an open issue, this task is usually accomplished manually. To avoid the registration problem, a lot of efforts have been devoted to the blind methods in the past decade.



(a) (b)

Fig. 1. An example of bleed-through removal. (a) the input image with bleed-through; (b) the bleed-through removal result.

A. Tonazzini *et al.* address the problem in the signal processing field by using blind source separation (BSS) [5]. The input scanned image with bleed-through is regarded as a mixed signal, i.e., a mixture of the foreground, bleed-through and background. The independent component analysis (ICA) algorithm is used to compute the weights of the component signals in the mixed signal (image pixel). This method requires color scanned image because it need signals collected by different sensors from the same objects. In the case of color scanned image, the input signals are the different channels of the input image. Therefore, the ICA method is also suitable for the non-blind bleed-through removal.

Apart from the ICA method, the same author also proposed to solve the BSS problem with Markov random fields (MRFs) and EM algorithm [6]. The mixing matrix and source vector are solved by MAP estimation on the MRFs. And the MAP estimation is performed in an iterative manner with EM algorithm where the mixing matrix and the source vector act as unobserved variables.

Different from A. Tonazzini's signal separation formulism, C. Wolf formulates blind bleed-through removal as an image segmentation problem [7]. A model with two hidden MRFs and single observation field is proposed, where the two hidden MRFs are binary. Each MRF and the observation field make a typical image segmentation problem which can be solved by maximum flow algorithm. Since both segmentation problems are limited by the same irregular nodes, an alternating optimization algorithm was proposed to estimate the hidden values, i.e., labels corresponding to the pixels.

Considering the large variance of intensity and noise of different historical document images, a CRFs based method

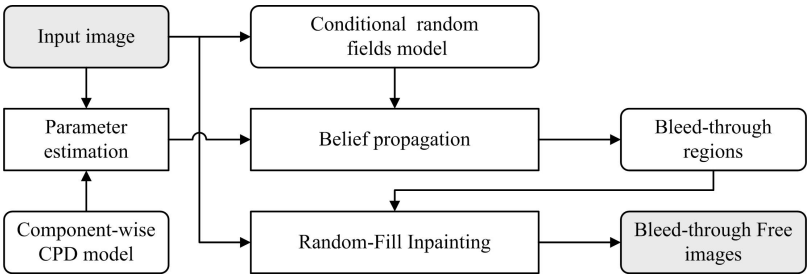


Fig. 2. The proposed conditional random fields based bleed-through removal method.

is proposed in this paper. The CPDs of the foreground, background and bleed-through components are modeled by using logistic function and Gaussian function respectively. The hidden layer in the graph model of the input image is also assumed to have the Markov property. Different from the MRFs method, the CRFs method maximizes the probability of each hidden node conditioned on the observation and the hidden fields [8]. The belief propagation (BP) algorithm [9] is used to estimate the probability of different values of the hidden node. Finally, a random filling (RF) algorithm is proposed to restore the bleed-through regions. The rest of this paper is organized as follows. Section II explains the proposed method in detail. Experimental results are shown in Section III. The conclusive remarks are made in Section IV.

II. CRFs based Blind Bleed-Through Removal

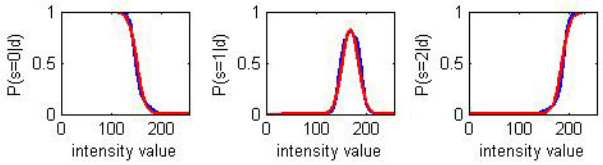
Fig. 2 is the overview of the proposed CRFs based blind bleed-through removal method. Architecturally, the proposed method consists of three blocks: component-wise CPD modeling, pixel-wise labeling with CRFs and random-fill inpainting. For the foreground, background and bleed-through components of the input image, logistic and Gaussian function are used to approximate the CPDs of the three components respectively. A K-means algorithm is used to generate a coarse labeling and the parameters of the component-wise model are computed accordingly. Then a CRFs model is established for the input image, where each node in the hidden Markov field has three label values corresponding to the three different components. The most possible candidate values for all the hidden nodes are chosen by using the BP algorithm. Finally, the detected bleed-through regions are inpainted by RF algorithm to ensure natural transition across boundaries of these regions.

A. Component-wise CPD Modeling

1) Component-wise Models: To model the CPD of each component, we firstly observe the CPDs of a small number of scanned images with groundtruth labels from the dataset, and then choose proper functions to approximate each component. Histograms of the whole image, foreground region, bleed-through region and background region are computed and denoted as H, H_{fg}, H_{bt} and H_{bg}, respectively. Then the CPD of each component can be computed as follows:

$$P(s=0|d) = \frac{H_{fg}}{H} \tag{1}$$

$$P(s=1|d) = \frac{H_{bt}}{H} \tag{2}$$



(a) (b) (c)

Fig. 3. An example of conditional probability distributions of the three components. (a) foreground component $P(s=0|d)$; (b) bleed-through component $P(s=1|d)$; (c) background component $P(s=2|d)$. The blue line is the CPDs calculated by (1), (2) and (3) on the real scanned image; The red line is the CPDs approximated by (4), (5) and (6).

$$P(s=2|d) = \frac{H_{bg}}{H} \tag{3}$$

where s is the label and d is the observed intensity value. $P(s=0|d)$, $P(s=1|d)$ and $P(s=2|d)$ are the conditional probability distributions of the foreground, bleed-through and background components accordingly. Fig. 3 shows an example of component-wise CPDs of the real image. The red lines in Fig. 3 are calculated on a real scanned image with groundtruth. Inspired by the shape of the three CPDs, we choose logistic function to model the foreground and background CPDs, and Gaussian function for the bleed-through. The model of the three CPDs are written as follows:

$$P(s=0|d) = \frac{1}{1+e^{\frac{d-u_0}{\sigma_0}}} \tag{4}$$

$$P(s=1|d) = \varphi e^{-\frac{(d-u_1)^2}{2\sigma_1^2}} \tag{5}$$

$$P(s=2|d) = \frac{1}{1+e^{\frac{u_2-d}{\sigma_2}}} \tag{6}$$

where φ, (u_0, u_1, u_2) and (σ_0, σ_1, σ_2) are the parameters of the CPD models. From Fig. 3, it can be found the CPD models fit the real-data CPDs very well.

2) Parameter Estimation: In order to calculate the parameter values of the input scanned image without groundtruth, a K-means segmentation is introduced to generate coarse labels. The centers of the foreground, bleed-through and background components are denoted as c_0, c_1 and c_2, respectively. Since the coarse segmentation may contains too many incorrect labels, optimization algorithms like the least mean square error algorithm are not suitable to estimate the parameters in the present of too many outliers. Therefore, a simple parameter estimation strategy is developed in the proposed method.

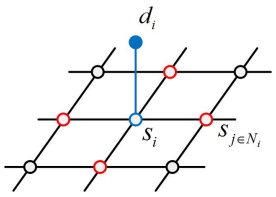


Fig. 4. The graphic unit of the CRFs model of the scanned image.

By fixing the conditional probability at the centers of each component, the parameters are calculated as follows:

$$u_0 = \frac{c_1 ln(\frac{1}{\delta_0^0} - 1) - c_0 ln(\frac{1}{\delta_1^0} - 1)}{ln(\frac{1}{\delta_0^0} - 1) - ln(\frac{1}{\delta_1^0} - 1)} \tag{7}$$

$$\sigma_0 = \frac{c_1 - u_0}{ln(\frac{1}{\delta_1^0} - 1)} \tag{8}$$

$$u_2 = \frac{c_1 ln(\frac{1}{\delta_2^2} - 1) - c_2 ln(\frac{1}{\delta_1^2} - 1)}{ln(\frac{1}{\delta_2^2} - 1) - ln(\frac{1}{\delta_1^2} - 1)} \tag{9}$$

$$\sigma_2 = \frac{u_2 - c_1}{ln(\frac{1}{\delta_1^2} - 1)} \tag{10}$$

where δ_0^0 and δ_1^0 are the CPD values of the foreground at c_0 and c_1, δ_1^2 and δ_2^2 are the CPD values of the background at c_1 and c_2. Since the bleed-through component is modeled with Gaussian function, the parameters u_1 and σ_1 are the mean and standard deviation of the bleed-through component.

B. Pixel-wise Labeling with CRFs

The CRFs model of the input scanned image consists of one hidden Markov field and one observation field. Fig. 4 illustrates one graphic unit of the CRFs model. In the CRFs model, one observation node and one hidden node correspond to one pixel of the input scanned image. The value of the observation node is the pixel value, and the hidden node value is one of three values, i.e., 0(foreground), 1(bleed-through) and 2(background). The value of the hidden node is determined by solving the following optimization function

$$\hat{s_i} = \arg\max_{s_i} P(s_i|D, S) \tag{11}$$

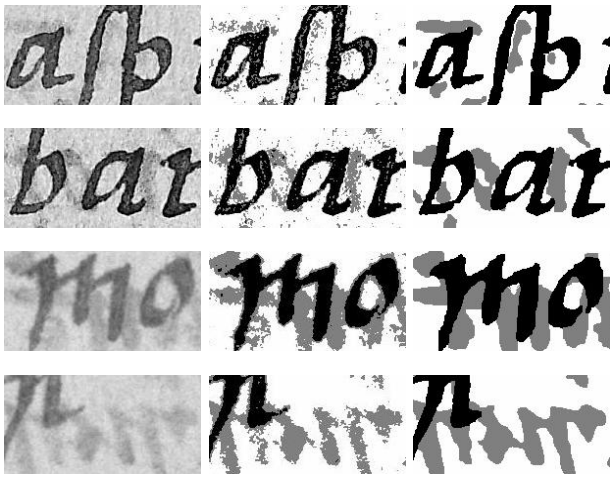
where s_i is the hidden node to be estimated, D is the observation field, S is the hidden label field. Because of the Markov property of the hidden field, it yields

$$\begin{aligned} P(s_i|D,S) &= P(s_i|d_i, N_i) \\ &= P(s_i|d_i)P(s_i|N_i) \\ &= P(s_i|d_i) \prod_{j\in N_i} P(s_i|s_j) \end{aligned} \tag{12}$$

where d_i is the corresponding observation node value of s_i and N_i is the neighbor set of s_i in the hidden field. The solution to this problem can be obtained approximately by the BP algorithm [9].

TABLE I. PARAMETER SETTINGS OF THE PROPOSED METHOD

Block	Parameter	Value
Foreground model	δ_0^0	0.85
	δ_1^0	0.15
Background model	δ_1^2	0.15
	δ_2^2	0.85
Random filling	r	30



(a) (b) (c)

Fig. 5. Labeling results of different methods. (a) original image; (b) K-means clustering; (c) The proposed CRFs based method.

C. Random-Fill Inpainting

After the label of each pixel is determined by CRFs based algorithm, an RF inpainting algorithm is proposed to restore the bleed-through regions. Each pixel in the bleed-through regions is replaced by the background pixel, which is randomly selected within the local window. This process can be described as following

$$d_k^{bt} = R(W_k \cap D^{bg}) \tag{13}$$

where R is the random selection operation, d_k^{bt} is the bleed-through pixel to be processed, W_k is the local window around d_k^{bt} and D^{bg} is the set of all the background pixels.

III. EXPERIMENTAL RESULTS

To evaluate the performance of the proposed method, an Irish historical manuscript scanned image dataset is used [10]. The dataset contains 25 image pairs. Each image pair consists of recto and verso image of the same page. In the experiment, only the recto images are used. Table I shows the parameter values of the proposed method used for all the images in the experiments.

Fig. 5 compares the labeling results of the proposed CRFs based method with the K-means clustering method. From Fig. 5(b), it can be found that the K-means method generate labels badly corrupted by the noise. On the contrary, the proposed CRFs based method produces label images with accurate and smooth boundaries even though the input scanned historical document images vary a lot in contrast and intensity.

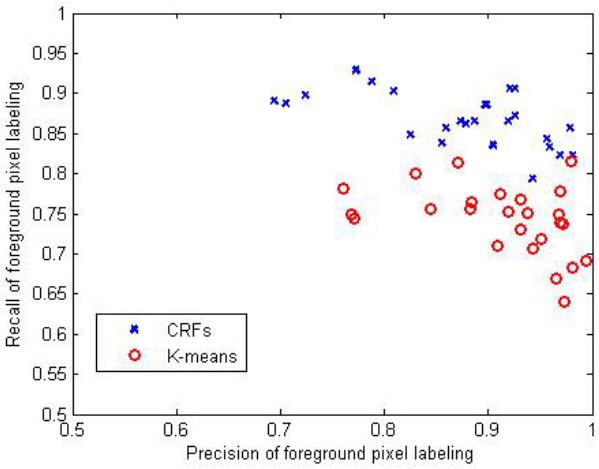


Fig. 6. Recall vs Precision of the foreground pixel labeling.

To objectively evaluate the labeling performance of the proposed method, the metrics including Precision and Recall are computed according to the foreground groundtruth provided in the dataset. The two metrics are computed as following

$$\text{Precision} = \frac{\text{sum}(\hat{S} = 0 \cap S^{gt} = 0)}{\text{sum}(\hat{S} = 0)} \tag{14}$$

$$\text{Recall} = \frac{\text{sum}(\hat{S} = 0 \cap S^{gt} = 0)}{\text{sum}(S^{gt} = 0)} \tag{15}$$

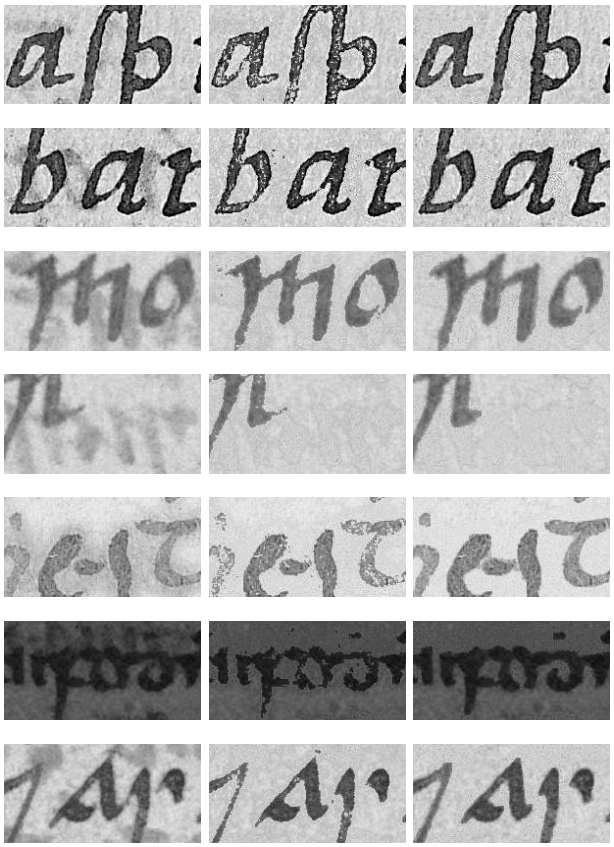
where $\hat{S}$ is the label set obtained by the proposed method, and S^{gt} is the groundtruth label set. Since the bleed-through regions in the recto images are usually different from the foreground regions in the verso image, these metrics of the bleed-through regions are not available.

Fig. 6 shows the foreground metrics on different images. From Fig. 6, it can be found that the proposed method obtains better Recalls than the K-means method but on some images the precision is a little lower. To the scanned historical document images, the foreground pixels should be preserved as many as possible. Therefore, in this case, the Recall metric should be more important than the Precision metric.

Fig. 7 show the final results of the proposed bleed-through removal method. By comparing the Fig. 7(b) and 7(c), it can be found that the proposed methods preserves the foreground component better than the K-means method. And by using the RF inpainting algorithm, the inpainted bleed-through regions enjoy natural transition across the boundaries, even if the background is very noisy. Note that although the input scanned images vary a lot in contrast and intensity, the proposed method removes the bleed-through effectively.

IV. Conclusion

A CRFs based bleed-through removal method is proposed in this paper. Firstly, the input scanned historical document image is coarsely segmented into foreground, bleed-through and background regions. The parameters of the component-wise CPD model are estimated based on the coarse segmentation. Then the CRFs model is established for the input image and the BP algorithm is used to compute the probabilities of candidate values corresponding to the three components. Finally, a random filling algorithm is developed to inpaint the bleed-through regions. Experiments on real data show that the proposed method preserves the foreground well and removes the bleed-through region effectively.



(a) (b) (c)

Fig. 7. Results of different bleed-through removal methods. (a) the original image; (b) K-means+RF; (c) the proposed method.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (No. 61172161 and No. 61325007).

References

[1] R. Estrada and C. Tomasi, Manuscript Bleed-Through Removal via Hysteresis Thresholding, *in Proc. Int. Conf. Doc. Anal. Recog.*, pp. 753-757, Jul. 2009.

[2] C. Wolf, Improving Recto Document Side Restoration with An Estimation of the Verso Side from A Single Scanned Page, *in Proc. Int. Conf. Pattern Recog.*, pp. 1-4, Dec. 2008.

[3] R. Rowley-Brooke, F. Pitie and A. Kokaram, A Non-Parametric Framework for Document Bleed-Through Removal, *in Proc. IEEE. Conf. Comput. Vis. Pattern Recog.*, pp. 2954-2960, Jun. 2013.

[4] J. Wang, M. S. Brown and C. L. Tan, Accurate Alignment of Double-sided Manuscripts for Bleed-through Removal, *in Proc. Eighth IAPR Workshop on Doc. Anal. Syst.*, pp. 69-75, Sept. 2008.

[5] A. Tonazzini and L. Bedini, Independent Component Analysis for Document Restoration, *Int. J. Doc. Anal. Recog.*, vol. 7, no. 1, pp. 17-27, Jan. 2004.

[6] A. Tonazzini and I. Gerace, A Markov Model for Blind Image Separation by a Mean-Field EM Algorithm, *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 473-482, Feb. 2006.

[7] C. Wolf, Document Ink Bleed-Through Removal with Two Hidden Markov Random Fields and a Single Observation Field, *IEEE Trans. Pattern Anal. Mach. Intel.*, vol. 32, no. 3, pp. 431-447, Mar. 2010.

[8] J. Lafferty, A. McCallum, and F. Pereira, Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Data, *in Proc. Int. Conf. Mach. Learn.*, pp. 282-289, Jun. 2001.

[9] J. S. Yedidia, W. T. Freeman and Y. Weiss, Constructing Free-Energy Approximations and Generalized Belief Propagation Algorithms, *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2282 - 2312, Jul. 2005.

[10] R. Rowley-Brooke, F. Pitie, and A. Kokaram. A Groundtruth Bleed-Through Document Image Database. In P. Zaphiris, G. Buchanan, E. Rasmussen, and F. Loizides, editors, *TPDL*, volume 7489 of *LNCS*, Paphos, Cyprus, 2012. Springer.