AN INPAINTING TECHNIQUE BASED ON REGULARIZATION TO REMOVE BLEED-THROUGH FROM ANCIENT DOCUMENTS

I. Gerace, C. Palomba

Dipartimento di Matematica e Informatica
Università degli Studi di Perugia
Via Vanvitelli, 1, 06123 Perugia, Italy

A. Tonazzini

Istituto di Scienza e Tecnologie dell'Informazione
Consiglio Nazionale delle Ricerche
Via G. Moruzzi, 1, 56124 Pisa, Italy

ABSTRACT

In the techniques proposed so far to remove bleed-through from digital images of ancient documents, two critical aspects are the identification of the occlusion areas, i.e. those pixels where the bleed-through pattern overlaps with the main foreground text, and the inpainting of the areas to be removed with a pattern that is in continuity with the surrounding background, often inhomogeneous due to paper texture or noise. In this paper we propose a new method for bleed-through removal that aims at solving both the aforementioned issues. The method first exploits information from the accurately registered images of the manuscript recto and verso to locate, in each side, the pixels corresponding to the interfering text, no matter if they are pure bleed-through or occlusion pixels. Then, processing separately the two sides, the identified areas are filled in by interpolating, through a suitable regularization model, the surrounding regions. We show the promising results obtained with this method on manuscripts affected by a very strong bleed-through.

Index Terms— Document image enhancement, Bleed-through removal, Image regularization, Local smoothness

1. INTRODUCTION

In manuscripts written on both pages of the sheet, the aging process often makes the ink to seep through the support fiber, so that the text of one side appears also in the other side. This defect is called bleed-through, and impairs the fruition of the document contents, besides being highly unpleasant.

Unfortunately, physical restoration techniques cannot be applied, since they risk to remove also the informative ink. Conversely, in the last years, digital image processing techniques have been used with some success to remove this degradation from the scanned versions of the manuscripts, especially when the digital images of both sides, accurately aligned, can be exploited.

The many methods proposed so far can be based on segmentation-classification techniques, or in modeling the recto and verso images as two parametric mixtures of the uncorrupted front and rear sides.

To classify the pixels as foreground, bleed-through or background, in [1] a regularized energy function is defined whose data term derives from small sets of user-labeled pixels, and whose smoothness term is a dual-layer Markov Random Field (MRF). In [2] a 4-class classification approach is proposed, by segmenting the recto-verso joint histogram with the aid of available ground truths. Pixel misclassifications are then iteratively corrected by analyzing the connected components, and bleed-through is inpainted with background patterns. The work in [3] proposes to exploit the slanting writing style of the manuscripts considered.

In the modeling approach, the first attempts used instantaneous linear models, directly inverted through independent component analysis or data decorrelation [4], or compensated for the apparent non-linearity of the phenomenon by MRF-based regularization [5], and penalized Non-negative Matrix Factorization [6]. Non-linear models, incorporating also a convolutional kernel to describe the smearing of the seeping ink, have been proposed in [7], where total variation is used to regularize the problem and the non-linearity is known, and in [8] and [9], where the model parameters are estimated as well. In [10] the problem is modeled by non-linear diffusion and solved by using wavelet transforms, whereas, in [11] and [12], we proposed a non-stationary convolutional model, linear in the optical densities, with pixel-dependent coefficients representing the percentage of ink seeping from the opposite side and estimated from the data.

Two common problems of the above methods are the to distinguish the occlusion areas from bleed-through, and to obtain a befitting inpainting of the removed interferences.

In this paper we propose a two-step technique that aim at resolving these two drawbacks. The technique is designed for manuscripts acquired in color, but it can be simplified to treat grayscale manuscripts. In the first step, the bleed-through pixels in the recto and in the verso are roughly located by exploiting the information of both sides of a single color channel. In the second step, each side is processed separately, but the three channels of the side are reconstructed

978-1-5090-5524-1/16/$31.00

jointly. In this step, the problem to inpaint the bleed-through pixels is formulated as a problem of interpolation of missing data. Taking advantage of the typical thinness of the characters in a manuscript, we adopt a regularization technique that, by exploiting intrachannel and interchannel local smoothness constraints, "smooths" inside the missing strokes the color of the pixels of the surrounding area.

The paper is organized as follows. Section 2 describes the technique employed to locate the bleed-through areas. Section 3 is devoted to the regularization model for efficiently inpainting the bleed-through. In Section 4 the experimental results are shown and discussed, and, finally, Section 5 concludes the paper with suggestions for further improvements.

2. THE METHOD: LOCATION OF THE BLEED-THROUGH PIXELS

We assume that, except at the occlusions, the foreground text is darker than bleed-through, so that, in order to locate the bleed-through areas in one side, the basic idea is to look for those pixels that are lighter than the corresponding pixels in the opposite side. However, based on some other considerations and assumptions, this strategy can be made more specific and efficient. First, since usually the ink in ancient manuscripts is black or brown, we can assume that the foreground is darker simultaneously in all the three color channels, and then select the bleed-through pixels using a single recto-verso pair of channels rather than the RGB pair. Second, we take into account the typical diffusion of the ink that penetrates through the paper. In other words, rather than simply compare the two recto and verso intensities, we compare the intensity in one side with the "smeared" version of the intensity in the opposite side. Therefore, we apply the following formulas to compute, at each pixels, the two ratios q_r and q_v:

$$\begin{aligned} q_r(i,j) &= \frac{1-\frac{s_v(i,j)}{b_v}}{h_r(i,j)\otimes\left(1-\frac{s_r(i,j)}{b_r}\right)+\epsilon} \\ q_v(i,j) &= \frac{1-\frac{s_r(i,j)}{b_r}}{h_v(i,j)\otimes\left(1-\frac{s_v(i,j)}{b_v}\right)+\epsilon} \end{aligned} \tag{1}$$

where the subscripts r, v indicate recto and verso, respectively, $s(i,j)$ is the intensity of the light of the chosen color reflected at pixel (i,j), b is an estimate of the predominant graylevel in the background, and ϵ is a small positive constant to avoid indeterminacies or infinity. These formulas include two unit volume Point Spread Functions (PSF), h_r and h_v, which, as said above, have the role to describe the smearing of ink that seeps through the paper, thus allowing a pattern in a side to better match the corresponding one in the opposite side. We assume that h_r and h_v have the form of a centered, Gaussian function, whose size and standard deviation can be approximately estimated from the extent of the character smearing in the bleed-through pattern (see Figure 1).

Based on the assumption of interferences less dark than the foreground, we will retain the smallest between the two

(a) (b)

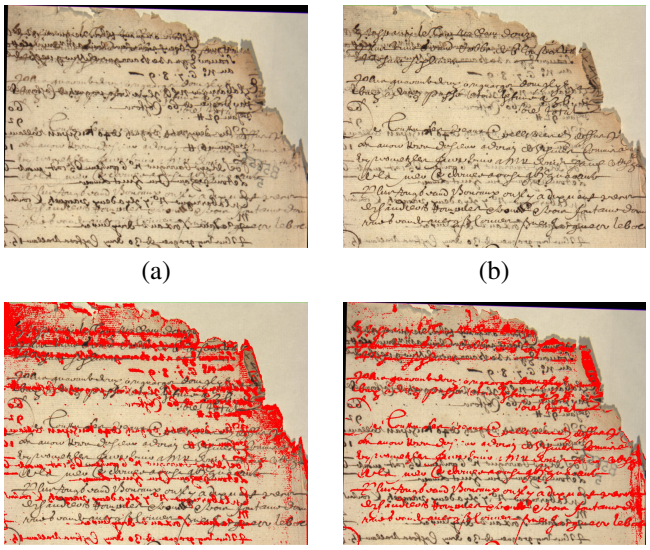
Fig. 1. Example of the typical smearing of the ink that seeps through the paper fiber: (a) foreground text; (b) corresponding bleed-through text.

computed ratios in eq. 1, to indicate a bleed-through pixel in the related side, and set to zero the other, to indicate a foreground pixel in the opposite side. Unfortunately, since the occlusion pixels cannot be individuated a priori, the application of the criterium above make them to be recognized as bleed-through in one side and as foreground in the other. A partial solution to this drawback can be found provided that the two inks reflect similarly under the same wavelength, and the observed densities at the occlusions are the highest across the manuscript and close to each other. If so, the occlusions can be located with the aid of suitable thresholds, and then q_r and q_v set both to zero, to indicate foreground in both sides.

In summary, with the strategy described above, the pixels with $q_v > 0$ are considered as bleed-through pixels in the recto side, while those with $q_r > 0$ are considered as bleed-through pixels in the verso side. The subsequent restoration phase will then consist in inpainting, independently in each side, the areas where q is greater than zero, whereas the areas where $q = 0$ are left unchanged.

As the two ratios are computed for all pixels in the two images, it would then be desirable to obtain zero values for both q_r and q_v is in correspondence of pixels of background in both sides. However, due to small fluctuation around b, this is usually hard to achieve. Again, a partial remedy is to set q_r and q_v to zero when the intensity of the pixel il greater than b minus some percentage of the standard deviation of a small patch of pure background, manually selected.

Figure 2 shows the original recto and verso images of a manuscript affected by a strong bleed-through (panels (a) and (b), and the same images with marked in red the pixels identified as bleed-through with the method above. It is apparent that large portions of the background are classified as bleed-through, as well as some of the occlusion areas. This means that, instead of being left unchanged, they will be inpainted with intensity values in continuity with those of their surrounding areas. As per the background, this is not a big inconvenience, especially if the background is homogeneous enough or its texture is not too regular. As per the occlusion areas, due to the typical thinness of the strokes, these are usually small and enclosed in foreground text. Hence, the available information should be sufficient for the inpainting algo-



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

Fig. 2. Original recto and verso images of a manuscript affected by a strong bleed-through (panels (a) and (b)); the same images with marked in red the pixels identified as bleed-through (panels (c) and (d)).

rithm to smear into them the color of the close foreground.

3. THE METHOD: REMOVAL AND INPAINTING OF THE BLEED-THROUGH PATTERN

As said, the restoration step will be formulated as a problem of interpolation of missing data, and solved via regularization separately in the two sides of the manuscript. However, since we use a regularization technique that exploits interchannel dependencies, the three channels of each side are processed together. The method will be then described for a generic side of the manuscript, thus dropping the subscripts r and v from the variables, but introducing a superscript to indicate the specific color channel, i.e. R, G, or B. By defining the set $\Omega = \{(i,j), s.t. q(i,j) > 0\}$, our variables will be then the intensities $(s^R(i,j), s^G(i,j), s^B(i,j))$ such that $(i,j) \in \Omega$, whereas the intensities of all other pixels in the image will be our data. In order to write the equations in a compact form, we define in vector form as $\mathbf{s_\Omega} = (\mathbf{s^R_\Omega}, \mathbf{s^G_\Omega}, \mathbf{s^B_\Omega})$ the variables to be estimated, and in vector form as $\mathbf{s} = (\mathbf{s^R}, \mathbf{s^R}, \mathbf{s^R})$ the entire set of pixels in the image.

According to regularization, we define the solution of our problem as the minimizer in $\mathbf{s_\Omega}$ of an energy function that describes available information on the expected solution, in order to overcome the ill-posed nature of the problem. In our case, this energy function is designed to express the local regularity of the intensity of each channel, through stabilizers that implicitly address intensity discontinuities of first, second and third order. In this way, we will be able to reconstruct very complex scenes with fine details. The energy function accounts also for an interchannel local smoothness, which is enforced in correspondence of the image high frequency components, usually highly correlated in the three color channels, through stabilizers that promote the amplitude of the intensity discontinuities in the different channels to be equal almost everywhere. We defined the following energy function:

$$E(\mathbf{s_\Omega}) = \sum_{k=1}^{3} \lambda_k^N \sum_{c \in C_k} g_{k,N}\left(N_c^k \mathbf{s}\right) + + \sum_{k=1}^{3} \lambda_k^V \sum_{c \in C_k} g_{k,V}\left(V_c^k \mathbf{s}\right), \quad (2)$$

where the first term is the sum of stabilizers expressing local intrachannel smoothness, the second term is the sum of stabilizers expressing local interchannel smoothness, and λ_k^N and λ_k^V are positive parameters balancing the relative weight of the two terms. More specifically, the terms $N_c^k\mathbf{s}$ are the Euclidean norms of the vectors of the finite differences D_c^k of order k computed, for each pixel and for each color channel, over a suitable set c of adjacent pixels:

$$N_c^k \mathbf{s} = \left\| \left(\mathbf{D_c^k s^R}, \mathbf{D_c^k s^G}, \mathbf{D_c^k s^B} \right) \right\| \quad (3)$$

where C_k represents the union of all the sets c. To make this cumulative smoothness term local, i.e. edge-preserving, each $N_c^k\mathbf{s}$ is weighted by a suitable function $g_{k,N}$ that decreases quickly for already small values of the gradients, thus further inhibiting intensity discontinuities due, e.g., to noise, and promoting a smooth filling-in of the lacking pixels. Conversely, $g_{k,N}$ makes N_c^k to decrease less when it is greater than a certain threshold, thus preserving existing, truly intensity discontinuities. The interchannel smoothness is enforced by the terms $V_c^k\mathbf{s}$, which are the norms of the vectors of the three interchannel first order finite differences of the intrachannel derivatives:

$$V_c^k \mathbf{s} = \left\| \left(\mathbf{D_c^k s^R} - \mathbf{D_c^k s^G}, \mathbf{D_c^k s^R} - \mathbf{D_c^k s^B}, \mathbf{D_c^k s^G} - \mathbf{D_c^k s^B} \right) \right\| \quad (4)$$

Analogously to the case of N_c^k, and with the same meaning, also operator V_c^k is weighted by a suitable function $g_{k,V}$, to permit the smoothness constraint to be local. We adopt functions $g_{k,N}$ and $g_{k,V}$ all having the same functional form of the truncated parabola [13], but containing a parameter κ that can instead vary from one to another:

$$g(t) = \begin{cases} t^2 & \text{if } |t| < \kappa \\ \kappa^2 & \text{otherwise} \end{cases} \quad (5)$$

Since the energy function of eq. (2) is non-convex, to obtain the solution we define a family of approximating energy functions $E^{(p)}$, $p = p_1, \cdots, p_0$, $p_1 > 0, p_0 = 0$, where the first approximation $E^{(p_1)}$ is convex and close enough to the original energy, and the successive ones gradually converge to it, i.e. $E^{(0)} = E$. The various approximations are obtained by acting on the form of the functions $g_{k,N}$ and $g_{k,V}$,

and are minimized in sequence, using the previous minimizer as starting point for the subsequent minimization. The minimization of each approximated energy is performed by means of a standard descent technique, namely the Non-Linear Successive Over-Relaxation (NL-SOR) [13] iterative algorithm.

To accelerate convergence, we try to start the process from a "good" point. We devised a fast, initial bilinear-type interpolation algorithm, to be applied independently on each channel, and whose result was proven to significantly reduce the computational time of the iterative process. Let us define the map M of missing pixels as follows:

$$M_{i,j} = \begin{cases} 0 & \text{if } (i,j) \in \Omega \\ 1 & \text{otherwise} \end{cases} \tag{6}$$

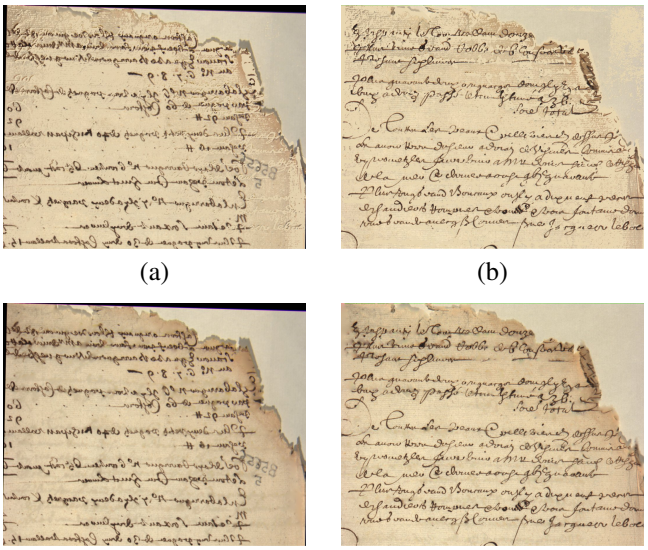
The algorithm scans the image from a corner and proceeds along adjacent pixels until a missing pixel (i.e. $M(i,j) = 0$) is reached. The value of such a pixel is then estimated as the average of the values of the non-missing pixels out of its four, vertically or horizontally aligned, neighbours, and the map M is updated to 1. The subsequent missing pixel is thus filled in by using also the values estimated for its previous, adjacent missing pixels. By dropping the superscript for the channel, in formulas it is:

$$\begin{aligned} s(i,j) &= \frac{s(i-1,j)M_{i-1,j} + s(i+1,j)M_{i+1,j}}{M_{i-1,j} + M_{i+1,j} + M_{i,j-1} + M_{i,j+1}} \\ &+ \frac{s(i,j-1)M_{i,j-1} + s(i,j+1)M_{i,j+1}}{M_{i-1,j} + M_{i+1,j} + M_{i,j-1} + M_{i,j+1}} \\ M(i,j) &= 1 \end{aligned} \tag{7}$$

The algorithm is repeated four times, each time starting from a different image corner, in such a way to avoid biases due to the direction, and the four solutions are averaged.

4. DISCUSSION OF THE EXPERIMENTAL RESULTS

We show the application of the method proposed above to the very degraded recto-verso pair of Figures 2 (a) and (b). Figures 3 compares the results of the new method (Figures 3 (c) and (d)) with those of the method proposed in [12] (Figures 3 (a) and (b)). Looking at the results of the method in [12], it is apparent that they are quite satisfactory, in terms of both bleed-through removal and preservation of the color, the paper texture and other possibly important features of the manuscript (note the pencil annotation in the middle of the right side of the recto). Furthermore, the algorithm is very simple and fast. Nevertheless, the price to be paid, as in this case, is that, for very inhomogeneous bleed-through patterns, whit some areas as strong as the foreground text, the full removal of the bleed-through implies also the removal, in one of the two sides, of the occlusion areas. In addition, the removed pattern is inpainted with a fixed color, i.e. the dominant color of the background. This means that, where the possible inhomogeneous background is darker, some imprints of the removed pattern can be visible. Looking instead at the results of the new method, it is apparent that, while all the good properties of the previous method are maintained, the occlusions are mostly preserved as well, and the bleed-through is inpainted with the natural texture of the close background.



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

Fig. 3. Restoration of the very degraded recto-verso pair of Figures 2 (a) and (b): (a) and (b) recto and verso restored with the method in [12]; (c) and (d) recto and verso restored with the proposed method.

5. CONCLUSIONS

We proposed a two steps algorithm for the removal of bleed-through from color manuscripts that is able to preserve the occlusion areas and to inpaint the bleed-through strokes with the background natural texture. In the first step bleed-through is roughly located by comparing the intensities in the recto and verso sides with the significant improvement that the typical smearing of the seeping ink is accounted for. In the second step, the inpainting of the bleed-through is formulated as a problem of interpolation of missing data, and solved via a regularization technique that exploits intrachannel and interchannel local smoothness constraints. The experimental results demonstrate the satisfactory performance of the method even against very degraded manuscripts.

Further improvements could arise from locating the bleed-through pattern more accurately. Indeed, at present, it often includes large portions of background, while, vice versa, it does not include some "halos" of the seeping strokes. This latter inconvenience might be due to the stationarity of the PSF adopted to describe ink smearing, and/or to an imprecise alignment of the two sides. In this respect, we plan to study efficient algorithms to solve the still challenging problem of the registration of recto-verso manuscript images.

6. REFERENCES

[1] Y. Huang, M. S. Brown, and D. Xu, "User assisted ink-bleed reduction," *IEEE Transactions on Image Processing*, vol. 19, no. 10, pp. 2646–2658, 2010.

[2] R. Rowley-Brooke, F. Piti, and A. Kokaram, "A non-parametric framework for document bleed-through removal," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2013, pp. 2954–2960.

[3] Jie Wang, Michael S. Brown, and Chew Lim Tan, "Accurate alignment of double-sided manuscripts for bleed-through removal," in *Proc. 8-th IAPR Workshop on Document Analysis Systems*, 2008, pp. 69–75.

[4] A. Tonazzini, E. Salerno, and L. Bedini, "Fast correction of bleed-through distortion in grayscale documents by a blind source separation technique," *Int. Journal on Document Analysis and Recognition*, vol. 10, pp. 17–25, June 2007.

[5] R. Rowley-Brooke and A. Kokaram, "Bleed-through removal in degraded documents," *Proc. SPIE 8297 Document Recognition and Retrieval XIX, 82970T-10*, 2012.

[6] F. Merrikh-Bayat, M. Babaie-Zadeh, and C. Jutten, "Using non-negative matrix factorization for removing show-through," in *Proc. LVA/ICA*, 2010, pp. 482–489.

[7] B. Ophir and D. Malah, "Show-through cancellation in scanned images using blind source separation techniques," in *Proc. Int. Conf. on Image Processing ICIP*, 2007, vol. III, pp. 233–236.

[8] F. Martinelli, E. Salerno, I. Gerace, and A. Tonazzini, "Non-linear model and constrained ml for removing back-to-front interferences from recto-verso documents," *Pattern Recognition*, vol. 45, pp. 596–605, 2012.

[9] E. Salerno, F. Martinelli, and A. Tonazzini, "Nonlinear model identification and seethrough cancellation from recto-verso data," *Int. J. on Document Analysis and Recognition*, vol. 16, pp. 177–187, 2013.

[10] R. F. Moghaddam and M. Cheriet, "A variational approach to degraded document enhancement," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, no. 8, pp. 1347–1361, 2010.

[11] A. Tonazzini, P. Savino, and E. Salerno, "A non-stationary density model to separate overlapped texts in degraded documents," *Signal, Image and Video Processing*, in press 2014.

[12] P. Savino and A. Tonazzini, "Digital restoration of ancient color manuscripts from geometrically misaligned recto-verso pairs," *Journal of Cultural Heritage*, vol. 19, pp. 511521, 2016.

[13] A. Blake and A. Zissermann, *Visual Reconstruction*, MIT Press, Cambridge, MA, 1987.