RESTORATION OF RECTO-VERSO ARCHIVAL DOCUMENTS THROUGH A REGULARIZED NONLINEAR MODEL

Ivan Gerace[(1)(2)], Francesca Martinelli [(1)], Anna Tonazzini [(1)]

[(1)]Consiglio Nazionale delle Ricerche, Istituto di Scienza e Tecnologie dell'Informazione,
Via G. Moruzzi, 1, I-56124 PISA, Italy
{name.surname}@isti.cnr.it
[(2)]Università degli Studi di Perugia, Dipartimento di Matematica e Informatica,
Via Vanvitelli, 1, I-06123 PERUGIA, Italy
gerace@dmi.unipg.it

ABSTRACT

We approach the removal of back-to-front interferences from recto and verso scans of archival documents as a blind source separation problem, considering the front and back ideal images as two individual patterns that overlap in the observed scans through some mixing operator. The nonlinear mixing model and the related restoration algorithm proposed in [1] are efficient for modern documents affected by mild show-through, but are not fully adequate to cope with ancient documents often degraded by the heavier and non-stationary bleed-through distortion. We then propose to modify this data model to account for non-stationarity of the degradation, and resort to the genuine concept of source separation for deriving the restoration algorithm. Within a regularization approach, we joint estimate the ideal images and the model parameters, by minimizing an energy function of all the unknowns, accounting also for local autocorrelation of the the ideal images. We derive a fully deterministic algorithm that is computationally efficient, and analyze its performance against documents heavily degraded by either show-through or bleed-through.

Index Terms— Document restoration, nonlinear data model, non-stationary data model, back-to-front interferences

1. INTRODUCTION

One of the most common degradations affecting archival documents that are written or printed on both sides of the page is the presence of back-to-front interferences (or see-through, show-through/bleed-through). These are undesired patterns in the background, caused either by transparency or seeping of the ink of the text printed in the reverse side of the page. Such distortion can significantly degrade the readability of the document by the interested scholars, make difficult the analysis of the document content, or prevent the successful application of OCR techniques.

Several approaches for see-through reduction have been already investigated, the most effective being those methods that exploit the information from scans of both sides (*recto* and *verso*). Basically, they are based on thresholding [2], wavelet techniques for enhancing the foreground strokes and smearing the interferences [3], and segmentation-classification [4].

Recently, the interest in applying Blind Source Separation (BSS) techniques for solving this problem has increased noticeably. The degraded document appearance is first modeled as a parametric superimposition of the uncorrupted front and back side images, and then a separation algorithm is used to estimate both the mixing parameters and the ideal images. The assumption of a linear mixing model has led to Independent Component Analysis (ICA) or data decorrelation techniques, which can be applied also when multispectral scans of a single side only are available [5][6][7], and to Non-negative Matrix Factorization (NMF) techniques, which minimize a data term plus a term compensating for the apparent nonlinearity of the show-through phenomenon, where occlusions between the texts occur [8].

Some works have also addressed more realistic nonlinear and/or convolutional mixing models. In [1], the physical model of the show-through in modern scanners is simplified for deriving a tractable mathematical nonlinear convolutional mixing model. Then the two recto and verso equations are decoupled, in order to design an adaptive linear filter. In [9], we modeled both multispectral single-sided and double-sided scans as convolutional mixtures of several component layers, and proposed edge-preserving regularization, performed via mixed deterministic-stochastic optimization. For the grayscale double-side case, paper [10] proposes a convolutional BSS formulation, which also accounts for a known nonlinearity, derived from [1], and the total variation stabilizer for the images. In [11], variational approaches, based

The work of Ivan Gerace has been supported by PRIN 2008 N. 20083KLJEZ.

on nonlinear diffusion and wavelet transforms, have been proposed to model and then remove bleed-through.

In this paper, we experiment the effectiveness of a modified version of the nonlinear convolutional mixing model proposed in [1], coupled with a totally different estimation algorithm, based on image regularization. Our aim is fourfold: i) relax the assumptions that make the original model mostly suited to describe mild show-through effects, in such a way to describe documents affected by even heavy see-through, or bleed-through, or both; ii) overcome the main limitation of most of the methods surveyed above, that is the stationarity of the mixing model; iii) include a regularity model for the images; iv) remove the simplifying assumptions leading to the restoration algorithm of [1], and, within a genuine BSS approach, derive a more effective separation algorithm.

Due to the nonlinearity of both the data model and the image model adopted, the energy function we design is non-convex in all its terms. To avoid costly stochastic algorithms, we then derive a family of approximations of the original energy function, to be used within a fully deterministic continuation method [12]. The resulting algorithm is computationally efficient and its effectiveness is demonstrated on both numerical and real examples.

The paper is organized as follows. In Section 2, we discuss the nonlinear convolutional model of [1], and propose some generalizations, to make it non-stationary and suitable to describe strong interferences levels. In Section 3, we derive the regularization method and algorithm proposed to solve the data model in a fully blind manner. Section 4 describes some experimental results on recto-verso pairs both numerically generated and real. Finally, Section 5 concludes the paper.

2. NONLINEAR, NON-STATIONARY CONVOLUTIONAL DATA MODEL

The data model we adopt in this paper is inspired to the show-through model for modern scanners proposed in [1]. This distortion is driven by a very complex physical model, depending on the paper scattering and transmittance parameters, the spreading of light in the paper, and the reflectance of the backing. In [1] a simplified, yet complex, mathematical model was first derived, and then further approximated to make it tractable. In particular, the assumption that the fraction of light transmitted is much smaller than the fraction of light scattered permits to "linearize" the model in such a way that the observed optical density in the recto/verso side is described as the sum between the ideal recto/verso optical density and the absorptance of the verso/recto written pattern convolved with an unknown kernel, the "show-through point spread function (PSF)". The model is symmetric and stationary, and a single parameter is taken for the reflectance of white paper, unprinted on either sides. To derive the show-through cancellation algorithm, the model is further simplified, by approximating the unknown show-through corrected absorptances with the corresponding absorbances from the observed scans. In this way the two sides can be processed independently in a pixel-by-pixel fashion, and the PSF tracked across the areas where the currently processed side has no printing and the other has printing, through least-mean square adaptive filtering. The resulting algorithm is very simple and effective, however it requires a number of thresholds to be set manually, and the PSF is assumed small in comparison to unity. This makes the algorithm suitable for modern scanned documents, affected by a relatively mild show-through. On the other hand, the simplifying assumptions adopted to derive the restoration algorithm make it not a proper separation algorithm, with some lacks in effectiveness. Indeed, substituting the ideal absorbance of the opposite side with the observed one can only work for mild distortion, in that, when the interference is strong, it produce erosion of the foreground characters in the side being restored.

When bleed-through is considered, its physical nature is ink seeping through the paper fiber and chemical transformations of the materials, rather than pure light transmission. In addition, the bleed-through pattern is usually an interference much more strong than show-through, and it is likely to be highly non-stationary, due to unpredictable causes, such as accidental humidity undergone by some parts of the sheet or inhomogeneity of the support and ink. From all the considerations above, it is clear that a general and comprehensive mathematical model for generic back-to-front interferences in ancient documents, accounting for these large variability of degradation effects, cannot be easily formulated. We then attempt to generalize the original show-through model in [1] to allow for non-stationarity and higher levels of interference, and propose to solve the data model through regularization, in order to include known a priori information on the solution.

Assuming, for simplicity sake, a single observation channel (or a grayscale observation), the nonlinear, non-stationary recto-verso model we consider is the following:

$$\begin{aligned} r(t) &= s_1(t) exp\left\{-q_2(t)\left[h_2(t)\otimes\left(1-\frac{s_2(t)}{R_2}\right)\right]\right\}, \\ v(t) &= s_2(t) exp\left\{-q_1(t)\left[h_1(t)\otimes\left(1-\frac{s_1(t)}{R_1}\right)\right]\right\}, \\ & t = 1, 2, \ldots, \mathrm{T} \end{aligned} \tag{1}$$

where $r(t)$ and $v(t)$ are the observed reflectances, and $s_1(t)$ and $s_2(t)$ are the ideal reflectances, of the front and back side, respectively, at pixel t, and R_1 and R_2 are the mean reflectance values of the background in the recto and verso side, respectively. We assume two different PSFs, h_1 and h_2 for the two sides, stationary across the image but characterized by different gains $q_1(t)$ and $q_2(t)$, possibly higher than 1, which represent the space-variant interference level from the front to the back and from the back to the front, respectively, at each pixel. With this notation, h_1 and h_2 are intended to be of unitary sum.

Let be $\mathbf{x}(t) = (r(t), v(t))^T$, $\mathbf{s}(t) = (s_1(t), s_2(t))^T$, $\mathbf{q}(t) = (q_1(t), q_2(t))^T$, and $\mathbf{H} = \{h_1, h_2\}$, the data model of eq. (1) can be rewritten in compact matrix form as:

$$\mathbf{x}(t) = B(\mathbf{q}, \mathbf{H}, \mathbf{s}; t)\mathbf{s}(t), \qquad t = 1, 2, ..., \mathrm{T} \qquad (2)$$

where $B(\mathbf{q}, \mathbf{H}, \mathbf{s}; t)$ is a 2×2 diagonal matrix, whose diagonal elements are

$$B_{11}(t) = exp\left\{-q_2(t)\left[h_2(t) \otimes \left(1 - \frac{s_2(t)}{R_2}\right)\right]\right\},$$

and

$$B_{22}(t) = exp\left\{-q_1(t)\left[h_1(t) \otimes \left(1 - \frac{s_1(t)}{R_1}\right)\right]\right\},$$

respectively.

3. SOLUTION THROUGH REGULARIZATION

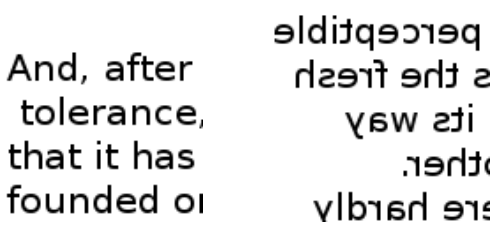
We propose the solve the system of eq. (2) with a regularization strategy, employing an edge-preserving autocorrelation model for the source images and the interference level maps. The fully blind problem we must solve is thus:

$$(\hat{\mathbf{s}}, \hat{\mathbf{q}}, \hat{\mathbf{H}}) = \arg\min_{\mathbf{s},\mathbf{q},\mathbf{H}} \left\{ \sum_{i=1}^{2} U_i(\mathbf{s}_i) + \sum_{i=1}^{2} \mathbf{q}_i^T W^T W \mathbf{q}_i + \sum_{t=1}^{\mathrm{T}} \left(B(\mathbf{q}, \mathbf{H}, \mathbf{s}; t)\mathbf{s}(t) - \mathbf{x}(t)\right)^T \left(B(\mathbf{q}, \mathbf{H}, \mathbf{s}; t)\mathbf{s}(t) - \mathbf{x}(t)\right) \right\} \qquad (3)$$

where $U_i(\mathbf{s}_i)$ are the image stabilizers, in the form of generic local smoothness models, augmented to account for constraints about the regularity features of realistic edge maps, such as penalization of broken or multiple edges [13]. This kind of model is particularly suitable for describing images of texts, both printed and handwritten. W is the matrix computing the finite differences of the first order, so that smoothness of the interference levels is enforced, by assuming that the intensity of the damage is somehow slowly varying across the document page.

As usual, we approach problem of eq. (3) by iteratively alternating componentwise minimizations with respect to subsets of homogeneous variables in turn, each minimization being subject to proper constraints of positivity and ranges of values allowed. In particular, for reasons that will become clear later on, we consider the subsets of the six homogeneous variables $\mathbf{s}_1$, $\mathbf{s}_2$, $\mathbf{q}_1$, $\mathbf{q}_2$, h_1, and h_2, respectively.

Since, in general, it is not possible to derive analytical formulas for the various variables viewed as functions of the others, and the energy is not convex in all its variables, our six minimization problems remain very difficult. The use of stochastic algorithms, such as simulated annealing, has already proven to be inefficient from a computational point of view, for a similar problem [9]. In this paper, we propose to exploit efficient deterministic non-convex optimization algorithms, and in particular we adapt to our specific problem the Graduated Non-Convexity (GNC) algorithm [12]. The GNC algorithm optimizes a non-convex function by tracking the minima of a sequence of functions generated by the variation of a control parameter, where the first member of the sequence ia a smoother, i.e. convex, version of the original non-convex cost function.



(a) (b)

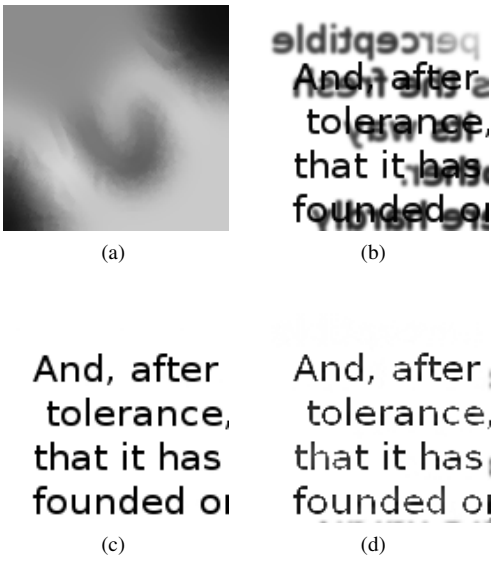
Fig. 1: Uncorrupted recto and verso images: (a) recto source; (b) verso source.

In [9], for a linear data model and the sources alone as variables, i.e. for a quadratic data term, the sequence of functions was generated by suitably approximating only the stabilizers U_i, which were the same we adopt herein. Nevertheless, the peculiarity, in the present case, is that also the data fidelity term is non-convex in all the parameters, and we must devise different approximating sequences for the energy function depending on the subset of variables considered.

In particular, building the first convex approximations for the data term is not immediate, if the homologous parameters in the two document sides (e.g. the two sources $\mathbf{s}_1$ and $\mathbf{s}_2$) were simultaneously considered as variables, since, basically, one variable multiplies the exponential of the other. Conversely, things becomes tractable when we only linearize the exponential, by considering fixed the multiplicative parameter. This is why we propose to estimate each pair of homologous parameters in two separated steps. Thus, the derivation of the convex approximation for the data term is based on approximating the exponential part by exploiting the polynomial of best approximation, which must be of degree 1 in this case, through the Gram-Schmidt orthogonalization [14].

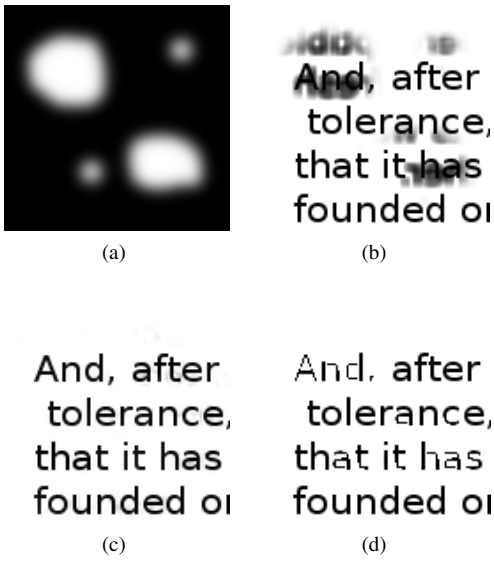
4. DISCUSSION OF THE EXPERIMENTAL RESULTS

We first tested the proposed method on two synthetic examples, where the data were generated numerically from the two real recto and verso source images shown in Figure 1. The aim is to quantify the performance of the method against very strong and highly non-stationary degradations. Without loss of generality, we used equal PSFs $h_1 = h_2$ and interference



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

Fig. 2: Experiment one on the pair of Figure 1: (a) interference level map; (b) observed recto; (c) recto restored by our method; (d) recto restored by the supervised Sharma method.



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

Fig. 3: Experiment two on the pair of Figure 1: (a) interference level map; (b) observed recto; (c) recto restored by our method; (d) recto restored by the supervised Sharma method.

levels $\mathbf{q}_1 = \mathbf{q}_2$ for the recto and the verso data images.

In the first experiment, we used a Gaussian PSF with standard deviation $\sigma = 1$, and the severe, arbitrary interference level map shown in Figure 2(a) (white pixels correspond to a very high interference level of value 3). Figures 2(b) and (c) show the observed and restored recto, respectively. In the second experiment we used a Gaussian PSF with standard deviation $\sigma = 2$, and the interference level map shown in Figure 3(a), simulating a damage caused by humidity. Figures 3(b) and (c) show the observed and restored recto, respectively. Table 1 reports the RMSEs between the restored images and the sources, the ideal and estimated interference level maps, and the ideal and estimated PSFs, respectively, in the two experiments. For comparison, Figures 2(d) and 3(d) show the solution of a supervised version of the Sharma algorithm in [1], i.e. assuming the exact knowledge of the PSF and the interference level map. The related RMSEs are reported in the last two columns of Table 1.

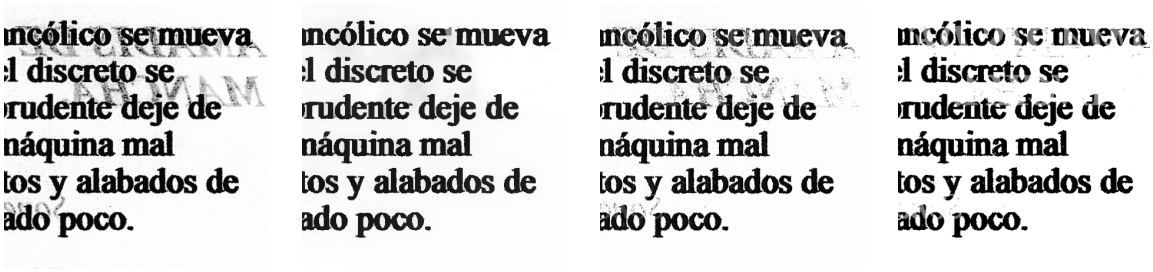
In another experiment, shown in Figure 4, the method is applied to the recto and verso scans of a real document. As it is apparent, it is not easy to decide if the degradation is caused by show-through, bleed-through, or both. In any case, the interfering patterns are highly non-stationary. The result of our method is very satisfactory (Figure 4(b)). For comparison, the result of the application of ICA, improved by histogram clipping, is shown in Figure 4(c), while Figure 4(d) shows the result of the method in [9]. The superior performance of the present method is apparent. Indeed, the other two methods, based on linear models, produce lower ink intensity in the occlusion areas, although the method in [9], accounting for a convolutional rather than instantaneous data model and regularity constraint for the solution images, is able to remove most of the interfering patterns.

5. CONCLUSIONS

We proposed a non-stationary and nonlinear convolutional mixing model coupled with an edge-preserving regularization approach for the separation of the two texts overlapped in recto-verso images of archival documents, affected by show-through or bleed-through. We derived a fully deterministic continuation method to optimize the non-convex cost function with respect to all the parameters. The experimental results on documents numerically generated in perfect accordance with the data model show that all the parameters involved can be estimated with very high accuracy. Also preliminary experiments on real data give promising results. In particular, the method clearly outperform standard BSS techniques such as ICA, which is based on the hypothesis of a linear instantaneous data model and independence of the overlapped patterns. The three fundamental features of our method are the nonlinearity and non-stationarity of the data model, which allows to describe the saturation of ink in the occlusions and the variability of the degradation intensity, the inclusion of a PSF, which allows a pattern in a side to mach the corresponding one in the opposite side, and, finally, the use of feasible local autocorrelation models for the images. Currently, we are working to the extension of the method to the multispectral, e.g. RGB, case, by restoring the various pairs of recto-verso channels independently.

	Recto	Verso	$\mathbf{q}_1$	$\mathbf{q}_2$	h_1	h_2	Recto Sharma	Verso Sharma
Experiment 1	1.89	2.01	0.01	0.12	$2.26 \cdot 10^{-3}$	$5.67 \cdot 10^{-3}$	23.43	21.94
Experiment 2	6.08	6.01	0.04	0.04	0.06	0.08	21.67	27.78

Table 1: RMSE



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

Fig. 4: Application of the nonlinear convolutional model to the recto-verso pair of a real document: (a) degraded recto; (b) recto restored with our method; (c) recto restored with ICA; (d) recto restored with the method in [9].

6. REFERENCES

[1] G. Sharma, "Show-through cancellation in scans of duplex printed documents," *IEEE Trans. Image Process.*, vol. 10, no. 5, pp. 736–754, 2001.

[2] E. Dubois and A. Pathak, "Reduction of bleed-through in scanned manuscript documents," in *Proc. IS&T Image Processing, Image Quality, Image Capture Systems Conference*, 2001, pp. 177–180.

[3] C. L. Tan, R. Cao, and P. Shen, "Restoration of archival documents using a wavelet technique," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 10, pp. 1399–1404, 2002.

[4] Q. Wang and C. L. Tan, "Matching of double-sided document images to remove interference," in *Proc. IEEE CVPR 2001*, 2001, p. 1084.

[5] A. Tonazzini, L. Bedini, and E. Salerno, "Independent component analysis for document restoration," *Int. Journal on Document Analysis and Recognition*, vol. 7, pp. 17–27, 2004.

[6] 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.

[7] A. Tonazzini, G. Bianco, and E. Salerno, "Registration and enhancement of double-sided degraded manuscripts acquired in multispectral modality," in *Proc. 10th International Conference on Document Analysis and Recognition ICDAR 2009*, 2009, pp. 546 – 550.

[8] 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.

[9] A. Tonazzini, I. Gerace, and F. Martinelli, "Multichannel blind separation and deconvolution of images for document analysis," *IEEE Trans Image Processing*, vol. 19, no. 4, pp. 912–925, April 2010.

[10] 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.

[11] R. F. Moghaddam and M. Cheriet, "Low quality document image modeling and enhancement," *Int. Journal on Document Analysis and Recognition*, vol. 11, no. 4, pp. 183–201, Mar 2009.

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

[13] L. Bedini, I. Gerace, and A. Tonazzini, "A deterministic algorithm for reconstructing images with interacting discontinuities," *CVGIP: Graph. Models Image Process.*, vol. 56, no. 2, pp. 109–123, 1994.

[14] I. Gerace, F. Martinelli, and A. Tonazzini, "See-through correction in recto-verso documents via a regularized nonlinear model," Tech. Rep. CNR.ISTI/2011-TR-011, ISTI-CNR, 2011.