Pattern Recognition 45 (2012) 596–605

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Nonlinear model and constrained ML for removing back-to-front interferences from recto–verso documents

Francesca Martinelli [a], Emanuele Salerno [a], Ivan Gerace [b,a], Anna Tonazzini [a,*]

[a] National Research Council of Italy, CNR, Institute of Information Science and Technologies, Via G.Moruzzi, 1 - 56124 Pisa, Italy
[b] University of Perugia, Department of Mathematics and Informatics, Via Vanvitelli, 1 - 06123 Perugia, Italy

ARTICLE INFO

Article history:
Received 29 November 2010
Received in revised form
24 March 2011
Accepted 11 July 2011
Available online 23 July 2011

Keywords:
Document restoration
Nonlinear data model
Back-to-front interferences

ABSTRACT

In this paper, we approach the removal of back-to-front interferences from scans of double-sided documents as a blind source separation problem, and extend our previous linear mixing model to a more effective nonlinear mixing model. We consider the front and back ideal images as two individual patterns overlapped in the observed recto and verso scans, and apply an unsupervised constrained maximum likelihood technique to separate them. Through several real examples, we show that the results obtained by this approach are much better than the ones obtained through data decorrelation or independent component analysis. As compared to approaches based on segmentation/classification, which often aim at cleaning a foreground text by removing all the textured background, one of the advantages of our method is that cleaning does not alter genuine features of the document, such as color or other structures it may contain. This is particularly interesting when the document has a historical importance, since its readability can be improved while maintaining the original appearance.

1. Introduction

Historical documents are almost always affected by various types of degradations, which overstay or are even highlighted in their digital versions. These degradations compromise both human readability and automatic analysis tasks, such as feature extraction and optical character recognition. Back-to-front interferences (bleed-through and show-through) are perhaps the most common degradations observable in ancient manuscripts. Bleed-through is intrinsic to the physical document, and is due to seeping of ink from the opposite side of a page, as an effect of humidity or absorption by the paper fibers. Conversely, show-through mainly appears in the document image as an effect of the scanning process, combined to the partial transparency of the paper. For this reason, it often affects the scans of even modern documents.

Image processing techniques to remove these interferences have recently gained a great interest, since they are needed to enable other automatic analysis tasks, or to improve legibility by scholars. The methods proposed in the literature can be distinguished into two classes: the ones that process separately the scans from the two sides of the documents (adaptive thresholding [1,2], segmentation-classification [3–5]), and the ones that apply to double-sided scans (segmentation-classification [6–10], adaptive filters [11], nonlinear diffusion [12]). Although the latter require a very accurate spatial registration of the two sides, they usually perform better, since exploiting the information from both sides helps discriminating the main text from the back-to-front interference. Many methods from both classes tend to remove all the structured background and, besides canceling the strokes coming from the reverse side, they can also remove other patterns originally belonging to the front side (e.g. stamps). This may be undesirable, when these patterns are signs of the document history and authenticity.

A specific research line focuses on applying blind source separation algorithms, viewing the foreground and the interference as individual patterns that overlap in the document, and relying on multiple observations for their separation. In [13], we proposed a linear instantaneous mixture model for the data, and an independent component analysis strategy to analyze single-sided multispectral scans with several overlapping information layers. We then extended this approach to the reduction of the back-to-front interference in registered recto–verso grayscale [14] and color [15] scans. The same generative model is used in [16] to propose a joint segmentation–separation strategy including a Markov model for the overlapping patterns. In [17], we removed instantaneousness to propose a convolutional formulation involving unknown point spread functions. Further extensions were proposed in [18,19], to include known nonlinearities in the generative data model.

In this paper, we assume double-sided grayscale scans as our data sets, and adopt the nonlinear mixing model proposed by Sharma [11], which is physically plausible for the show-through

* Corresponding author.
E-mail addresses: francesca.martinelli@isti.cnr.it (F. Martinelli), emanuele.salerno@isti.cnr.it (E. Salerno), gerace@dmi.unipg.it (I. Gerace), anna.tonazzini@isti.cnr.it (A. Tonazzini).

0031-3203/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2011.07.016

effect produced by modern scanners. This model has the advantage of being nonlinear and convolutional, thus avoiding some of the common difficulties featured by linear and instantaneous models, but the constraints imposed and the approximated restoration algorithm proposed in [11] are only sufficient to remove moderate show-through. Moreover, we are also interested in images of ancient documents, which may have not been captured by scanners and normally show quite complicated features. First of all, ageing can cause the paper to be not "white" anymore, and the background colors can be different in the two sides. Second, both show-through and bleed-through can be observed in the same page, and their intensity can even be comparable to the one of the main text. Even though physically based models are also available for bleed-through (see, for example, [20]), to our knowledge, no general and comprehensive mathematical model has been proposed for all the possible back-to-front interferences.

To try solving this problem by a single algorithm, we conjecture that relaxing the constraints imposed in [11] and adopting an iterative optimization rather than an approximated restoration algorithm enables us to treat general back-to-front, even strong, interferences. Our experiments show how the new algorithm gives better results than the methods based on linear models. In the most severe cases, moreover, the advantages of our algorithm over the simplified strategy proposed in [11] are apparent. A further advantage is that the restored images, while cleansed of the unwanted interferences, keep substantially intact their original appearance. As specified in Sections 2 and 3, the essential differences between our proposal and the simplified strategy are that (1) the data model does not constrain the interference level to be moderate anymore, (2) the interference level of the back pattern over the front page is allowed to be different from the one of the front pattern over the back page, and (3) rather than an adaptive filtering with an approximated version of the interfering pattern, we adopt a constrained maximum likelihood method where the interfering patterns, along with the unknown point spread functions, are refined iteratively.

This paper is organized as follows. In Section 2, we describe our data model in detail, highlighting the aspects that distinguish it from its simplified version. Section 3 is devoted to the description and discussion of the solution technique adopted. In Section 4, experimental results on synthetic and real documents are shown and discussed. Finally, Section 5 concludes the paper and suggests some future prospects. Two appendices show some properties of the objective function to justify heuristically the optimization algorithm used, and explore a possible strategy to relax the stationarity assumption.

2. A nonlinear model for back-to-front interference

By a physical analysis of the show-through effect that a scanner introduces in a digital image of a double-sided document, Sharma [11] derives a generative model that takes into account the reflection, transmission and scattering parameters of the paper. Although extremely simplified with respect to the physical phenomenon, this model is quite complex, so that some approximations are taken to make it tractable. In particular, the assumption that the fraction of light transmitted is much smaller than the fraction scattered towards the sensor gives rise to a "linearized" data model, where the observed optical density in the recto of the document (that is, in its front side) is described as the sum of the ideal density and the ideal absorptance of the verso (that is, the document's back side) convolved with an unknown kernel, the *show-through point spread function* (*PSF*). A specular equation describes the observed optical density of the verso in terms of the absorptance of the printed pattern in the recto. Note that, although the model is linear and convolutional in density and absorptance, it is nonlinear when written in terms of all densities or all reflectances, due to the logarithmic relationship between optical density and reflectance. The model is symmetric and stationary, i.e. the same show-through PSF is assumed for both sides, and all the transmittance and reflectance parameters of the paper are considered uniform and independent of the side.

We now present a generalized form of this model, written in terms of reflectances, from which we are able to both explain the constraints deriving from Sharma's physical assumptions and highlight the differences with the model assumed here. The image capture system provides two discrete $N\times M$ reflectance maps, from the front and the back side of the document, whose values at corresponding pixels are related to the same spatial locations, that is, the two maps are spatially registered. The intensity values are quantized and rescaled in some integer range, $\{0,1,2,\ldots R_{max}\}$. Each map presents the *ideal* reflectance pattern of one side, disturbed by traces of the *ideal* reflectance pattern of the opposite side. The observed and ideal maps from both sides are related through the following nonlinear system:

$$
\begin{aligned}
x_f^s(i,j) &= x_f(i,j)\cdot\exp\left\{-q_b\left[h_b(i,j)\otimes\left(1-\frac{x_b(i,j)}{R_b}\right)\right]\right\}\\
x_b^s(i,j) &= x_b(i,j)\cdot\exp\left\{-q_f\left[h_f(i,j)\otimes\left(1-\frac{x_f(i,j)}{R_f}\right)\right]\right\}\\
i &= 1,\ldots,N,\quad j=1,\ldots,M
\end{aligned}
\tag{1}
$$

where $x_{f,b}^s(i,j)$ and $x_{f,b}(i,j)$ represent the observed and ideal reflectances, respectively, at pixel (i,j), $R_{f,b}$ (in the range $[0,R_{max}]$) are the mean background reflectances, $h_{f,b}(i,j)$ are the PSFs, and the subscripts f and b denote the recto and verso sides, respectively. The symbol "$\otimes$" means convolution. Functions $h_{f,b}$ are assumed to have unit volume. The gains $q_{f,b}$ (*interference levels*) affect the intensities of the front-to-back and back-to-front interferences, respectively, and would generally depend on location. Since we want model (1) to be stationary, hereafter we treat $q_{f,b}$ as constant, unless otherwise specified.

As anticipated, the model introduced in [11] is uniform and symmetric, and is physically sound if only a small part of the incident light penetrates the paper and is reflected back by the scanner. In the notation introduced in Eq. (1), these conditions are expressed as

(i) $q_b = q_f = q$
(ii) $h_b = h_f = h$
(iii) $R_b = R_f = R$
(iv) $q\cdot h(i,j) \ll 1 \quad \forall(i,j)$

Any restoration algorithm derived from such a model can only be effective with scans of documents characterized by homogeneous supports and affected by moderate show-through. Working with ancient documents, however, can be more complicated: homogeneity is not ensured, show-through can be severe, bleed-through can be present, and the interference effect could not be symmetric.

Modeling such a situation could lead to relationships that are much more complicated than (1). To avoid unnecessary complications, we attempt to maintain this model in cases with heavy distortions, by somewhat relaxing constraints (i)–(iv) above. This choice is motivated by the fact that the data model, although unphysical, maintains some desirable properties. First, being convolutive, it accounts for a blur in the interfering patterns and, as demonstrated in [17], this improves the result obtained, even through a linear model. Second, being nonlinear, it can avoid the partial attenuation produced by the linear assumptions where the main text occludes the back-to-front pattern (this has also

been noted in [17]). Of course, the effectiveness of our attempt can only be demonstrated experimentally. Section 4 contains some discussion on this issue.

We assume that the back-to-front interference is not moderate, namely, since our PSFs have unit volume, that q_f and q_b can be large. In principle, q_f and q_b range from zero (the paper is completely opaque) to infinity, that is, the paper is infinitely transparent, and the ideal pattern on one side appears unmodified as show-through in the opposite side. As explained in the next section and in Appendix A, however, we need to assume a finite maximum for q_f and q_b. This can be obtained if we evaluate the maximum interference level in the pixels whose value is 1, rather than 0 (i.e. the black pixels). From Eq. (1) evaluated in a pure show-through pixel, it is easy to see that the maximum value for q_f or q_b depends on R_{max}. For $R_{max}=255$, for example, the maximum value for $q_{f,b}$ is 5.56. Note that this is only to evaluate an interference level that cannot be exceeded in any practical case, and does not entail any preprocessing on the data images.

By allowing $q_f \neq q_b$, $h_f \neq h_b$ and $R_f \neq R_b$, we relax the symmetry assumption. However, since we assume perfectly registered image pairs, we consider h_f and h_b radially symmetric and approximately Gaussian. In our views, this can reflect both the phenomena of light scattering and ink diffusion, thus modeling both a show-through and a bleed-through PSF. We set the supports of the PSFs to a reasonable maximum size, inferred by visual inspection. As mentioned, assuming different background levels R_f and R_b can be necessary to capture possible differences caused by ageing.

The last issue to be treated, which can be important when dealing with ancient documents, is nonstationarity. To account for this phenomenon, a data model could become numerically intractable. We do not add anything on this issue here, since this is out of our present scope. However, we are now approaching this problem by just admitting that q_f and q_b can be functions of (i,j), while maintaining constant supports and shapes for the PSFs. In Appendix B, we give a few details about this strategy.

3. A constrained maximum likelihood solution

Eq. (1), subject to constraints (i)–(iv), is assumed in [11] as a simplified physically based generative model for show-through. To derive a show-through cancellation algorithm, the model is further simplified, by approximating the unknown ideal back and front maps in the arguments of the exponentials by the corresponding observed scans. This allows the two sides to be processed independently and pixel by pixel. A least-mean-square adaptive filter is then used to restore each side in a single pass. This approach compensates for possibly nonstationary PSFs by tracking them across the pure show-through areas, and is also a way to relax the symmetry and stationarity of the original model. The algorithm is simple and effective, but needs to identify, on each side, the printed areas, the background, and the show-through strokes. This complicates the application of the procedure, since a number of thresholds must be set manually. Most importantly, assuming $x_{f,b}=x^s_{f,b}$ in the exponentials makes the requirement of having a mild show-through even more demanding. Conversely, we use Eq. (1) as is, by dropping all the constraints (i)–(iv), thus allowing the model to be fully asymmetric and admit a strong back-to-front interference. In this section, we describe a restoration scheme derived from this model.

Since we do not let the ideal recto and verso maps be replaced by their observed versions, a solution to Eq. (1) cannot treat the two equations separately. Our approach is to build a maximum likelihood solution in the set of all the unknowns, $\Omega=\{x_f,x_b,h_f,h_b,q_f,q_b\}$. If we include a white, zero-mean, Gaussian and signal-independent noise component in the model, this problem reduces to finding the minimizer to

$$\begin{aligned}E(x_f,x_b,h_f,h_b,q_f,q_b) &= \sum_{i=1}^{N}\sum_{j=1}^{M}\left(x^s_f(i,j)-x_f(i,j)\cdot\exp\left\{-q_b\left[h_b(i,j)\otimes\left(1-\frac{x_b(i,j)}{R_b}\right)\right]\right\}\right)^2 \\ &\quad+\left(x^s_b(i,j)-x_b(i,j)\cdot\exp\left\{-q_f\left[h_f(i,j)\otimes\left(1-\frac{x_f(i,j)}{R_f}\right)\right]\right\}\right)^2 \end{aligned} \tag{2}$$

subject to the constraints (assuming $R_{max}=255$):

$$0\le x^s_f(i,j)\le 255,\quad 0\le x^s_f(i,j)\le 255,\quad \forall(i,j)$$

$$0<q_f<5.56,\quad 0<q_b<5.56$$

$$\sum_{i,j}h_f(i,j)=\sum_{i,j}h_b(i,j)=1$$

$h_f(i,j)$ and $h_b(i,j)$ are single-peaked and circularly symmetric

As always happens in blind estimation, the problem of identifying model (1) and estimating the ideal recto and verso images, if unconstrained, admits an infinity of solutions. Indeed, if we fix any subset of Ω, it is possible to derive the remaining parameters so that the energy function E reaches its global minimum, which is zero. From heuristic considerations on the behavior of function (2), detailed in Appendix A, we have found that the constraints listed above are able to avoid most spurious solutions. We then derived a stochastic scheme able to reach the desired solution. This is made of two nested cycles. The outer cycle employs a Monte Carlo strategy to minimize E with respect to q_f, q_b, h_f and h_b, with x_f and x_b fixed. The inner cycle is deterministic and, at each step of the outer cycle, clamps x_f and x_b to the minimum reached by E when the remaining parameters are fixed to their current values. In formulas:

$$(\hat{h}_f,\hat{h}_b,\hat{q}_f,\hat{q}_b)=\arg\min_{h_f,h_b,q_f,q_b}E(x_f(h_b,q_b),x_b(h_f,q_f),h_f,h_b,q_f,q_b) \tag{3}$$

with

$$(x_f(h_b,q_b),x_b(h_f,q_f))=\arg\min_{x_f,x_b}E(x_f,x_b,h_f,h_b,q_f,q_b) \tag{4}$$

where the problem in (4) is solved by iterating the following deterministic scheme:

$$\begin{aligned}x^{n+1}_f(i,j) &= x^n_f(i,j)+\mu\left(x^s_f(i,j)-x^n_f(i,j)\right. \\ &\qquad\left.\cdot\exp\left\{-q_b\left[h_b(i,j)\otimes\left(1-\frac{x^n_b(i,j)}{R_b}\right)\right]\right\}\right) \\ x^{n+1}_b(i,j) &= x^n_b(i,j)+\mu\left(x^s_b(i,j)-x^n_b(i,j)\right. \\ &\qquad\left.\cdot\exp\left\{-q_f\left[h_f(i,j)\otimes\left(1-\frac{x^{n+1}_f(i,j)}{R_f}\right)\right]\right\}\right) \\ i&=1,\ldots,N,\quad j=1,\ldots,M\end{aligned} \tag{5}$$

where μ is a small constant, and, at each iteration, the constraints on (x_f,x_b) are enforced. A practical procedure to implement this scheme is

Algorithm 1.

(a) $k=0$, $q^0_f=5.56$, $q^0_b=5.56$, $h^0_f=$ uniform, $h^0_b=$ uniform, $x^0_f=x^s_f$, $x^0_b=x^s_b$
(b) Compute E^k
(c) Propose a random update $(h^{k+1}_f,h^{k+1}_b,q^{k+1}_f,q^{k+1}_b)$
(d) Find the corresponding minimum (x^{k+1}_f,x^{k+1}_b) by scheme (5)
(e) Compute E^{k+1}
(f) Accept the new estimate $(h^{k+1}_f,h^{k+1}_b,q^{k+1}_f,q^{k+1}_b,x^{k+1}_f,x^{k+1}_b)$ if $E^{k+1}<E^k$
(g) Set $k=k+1$
(h) Repeat from step (c) until convergence

where, at step (c), the parameters are perturbed by small random, Gaussian-distributed increments.

4. Discussion of the experimental results

In this section, we report some of the results obtained by testing the model of Eq. (1) and the related constrained ML solution algorithm.

In all the experiments, we assume h_f and h_b to have the same square support, of size $L \times L$ (L odd). This support is estimated by inspection of the data, and depends on both the extent of the ink diffusion through the paper, and the acquisition resolution. In our experiments, however, we observed that a certain tolerance is admitted for the size of the blur mask. As per the shape, we experimented both a pure Gaussian shape and a Gaussian-like shape. With the pure Gaussian shape, the parameters to be estimated reduce to the only standard deviations, independently of the support size chosen. This choice has the further advantage that just two shift parameters would be sufficient to correct possible residual misalignments between recto verso, due to an imperfect registration. The results we report here, however, always assume zero shifts, that is, perfectly registered recto–verso pairs. As anticipated, by "Gaussian-like shape" we mean a circularly symmetric blur mask, whose values decrease monotonically from the central peak to the periphery. By these assumptions, the number of coefficients to be estimated for each mask increases to $(L^2-1)/8+(L+1)/2$, and, although feasible, accounting for a residual misalignment would be less immediate. However, such a shape is straightforwardly compatible with both a uniform blur and a unit pulse (no blur). R_f and R_b are estimated by averaging a small background patch in each side. In most experiments, rather than forcing each estimate to saturate at 255, we preferred to saturate x_f to R_f and x_b to R_b. Indeed, a saturation value equal to 255 would produce "imprints" of the removed interferences when the background is significantly dark. To implement Algorithm 1, we still need the constant factor μ to be used in (5), and the standard deviations of the random increments $\delta_{h,q}$ for $h_{f,b}$ and $q_{f,b}$. In all the experiments shown here, we set $\mu = 0.1$ and $\delta_h = \delta_q = 0.01$. At each iteration of the outer cycle, the deterministic update (5) is performed just once.

In a first set of experiments, we consider four synthetic recto–verso pairs, generated so as to fit the model perfectly. The four interference levels adopted have been chosen so as to make our images similar to the ones used in [12], although the latter were generated by a different model, namely, a nonlinear diffusion. Since these experiments are only intended to evaluate the ideal performance of our method, we generated the test data through a perfectly symmetric model with a uniform 3×3 PSF. In the left-hand column of Fig. 1, our degraded recto sides are shown, with increasing interference levels from top to bottom. In these artificial cases, rather than initializing the PSFs by uniform masks as stated in Algorithm 1, we always adopted unit pulses as our initial guesses. In the center column of Fig. 1, we show the recto sides of the images restored by our method. In the right-hand column, we show the corresponding results obtained from system (1) by using $x_f = x_f^s$ and $x_b = x_b^s$ in the exponentials, and the true values of $q_{f,b}$ and $h_{f,b}$. Since these results can be considered as the best attainable by the method proposed in [11], we call them "Sharma's results". Table 1 shows the estimated $\hat{q}$ and $\hat{h}$, and the root mean square errors (*RMSE*) obtained by our fully blind algorithm and Sharma's algorithm in the four cases. As apparent, although the interference is completely removed by Sharma's method, the characters of the printed text are somehow "eroded".

Fig. 1. Left, from top to bottom: synthetic recto images with increasing interference levels ($q = 0.5,1,2,3.18$, $h =$ uniform 3×3). Center: corresponding restoration results from Algorithm 1. Right: corresponding "Sharma's results".

Table 1
Quantitative evaluation of the results in Fig. 1.

$q=0.5$	$q=1$	$q=2$	$q=3.18$
$\hat{q}=0.507$	$\hat{q}=1.010$	$\hat{q}=2.027$	$\hat{q}=3.224$
$\hat{h}=\begin{bmatrix}0.110 & 0.111 & 0.110\\ 0.111 & 0.118 & 0.111\\ 0.110 & 0.111 & 0.110\end{bmatrix}$	$\hat{h}=\begin{bmatrix}0.111 & 0.111 & 0.111\\ 0.111 & 0.115 & 0.111\\ 0.111 & 0.111 & 0.111\end{bmatrix}$	$\hat{h}=\begin{bmatrix}0.110 & 0.110 & 0.110\\ 0.110 & 0.117 & 0.110\\ 0.110 & 0.110 & 0.110\end{bmatrix}$	$\hat{h}=\begin{bmatrix}0.110 & 0.111 & 0.110\\ 0.111 & 0.115 & 0.111\\ 0.110 & 0.111 & 0.110\end{bmatrix}$
$RMSE=1.18$	$RMSE=1.48$	$RMSE=2.80$	$RMSE=9.26$
$RMSE_{\text{Sharma}}=2.76$	$RMSE_{\text{Sharma}}=8.65$	$RMSE_{\text{Sharma}}=20.74$	$RMSE_{\text{Sharma}}=29.92$

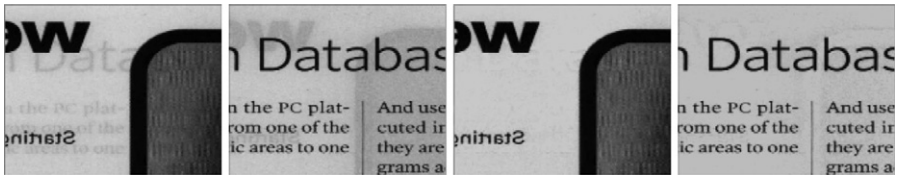

Fig. 2. Removal of back-to-front interferences from a real recto–verso pair (from [18]). From left to right: degraded recto; degraded verso; restored recto; restored verso. The effects of a pure Gaussian and a Gaussian-like PSFs are visually indistinguishable.

We believe that this effect is due to the substitution of the ideal reflectances by the observed images. Indeed, the result of this choice is that the "interference map" subtracted in each side contains a fraction of the main text. This is reflected in the $RMSE_{\text{Sharma}}$ values in Table 1, which are always worse than the corresponding values obtained by the iterative method. The table refers to the results obtained by a symmetric model. The same experiments have been performed without binding the model to be symmetric, and assuming both pure Gaussian and Gaussian-like PSFs. The results are visually undistinguishable from the ones shown in Fig. 1, and the quantitative indices are very similar to the ones reported in Table 1. We also attempted a restoration by guessing a symmetric model with a 5×5 Gaussian-like PSF. Significantly, the result was a 3×3 almost uniform mask surrounded by a frame with almost vanishing elements. In the case $q=2$, we obtained

$$\hat{h}=\begin{bmatrix}0.000 & 0.000 & 0.004 & 0.000 & 0.000\\ 0.000 & 0.104 & 0.113 & 0.104 & 0.000\\ 0.004 & 0.113 & 0.116 & 0.113 & 0.004\\ 0.000 & 0.104 & 0.113 & 0.104 & 0.000\\ 0.000 & 0.000 & 0.004 & 0.000 & 0.000\end{bmatrix}$$

and an estimated interference level $\hat{q}=2.04$. This result suggests that the Gaussian-like PSF can be used to make the algorithm independent of the size guessed for the PSFs.

We also performed an extensive experimentation on real cases. This does not allow for a quantitative evaluation, but is necessary nonetheless. Indeed, our unphysical data model cannot account for all the degradations, so a number of real tests are essential to appreciate the practical applicability of our method. Here we report some indicative examples.

In the first (see Fig. 2), we test both the pure Gaussian and the Gaussian-like PSFs within the asymmetric model. Table 2 summarizes the parameters obtained in the two cases, assuming the size of the PSFs to be 3×3. Despite the small differences in the estimated masks, the reconstructed images are nearly identical (see the two rightmost panels in Fig. 2). From this experiment, it is possible to appreciate the salient feature of the source separation approach from recto–verso data, that is, only the strokes and patterns coming from the reverse side are canceled, whereas other structures contained in the background are preserved, as well as its original color.

The data images and our restoration results from a second real experiment are shown in Fig. 3. A detail thereof (Fig. 4) is used to compare our method with the one proposed in [19], based on a quadratic mixing model and a fixed show-through PSF. Whereas the restored images would probably be equally suitable for subsequent processing tasks (e.g binarization, segmentation), it is to note that our method reconstructs better the original features of the document, such as the contrast and the background color. The same data were also used to test a method based on a linear instantaneous model. A direct comparison can be made from the results shown in [14], where it is apparent that the linear method is not able to remove all the interference and also introduces additional distortions to the main text strokes.

Table 2
Parameters obtained for the real example shown in Fig. 2. $\hat{\sigma}_f$ and $\hat{\sigma}_b$ are the standard deviations of the estimated Gaussians.

Gaussian PSFs	Gaussian-like PSFs
$\hat{q}_f=0.166$, $\hat{q}_b=0.222$	$\hat{q}_f=0.167$, $\hat{q}_b=0.213$
$\hat{\sigma}_f=0.998$, $\hat{\sigma}_b=1.343$	
$\hat{h}_f=\begin{bmatrix}0.075 & 0.124 & 0.075\\ 0.124 & 0.203 & 0.124\\ 0.075 & 0.124 & 0.075\end{bmatrix}$	$\hat{h}_f=\begin{bmatrix}0.046 & 0.147 & 0.046\\ 0.147 & 0.229 & 0.147\\ 0.046 & 0.147 & 0.046\end{bmatrix}$
$\hat{h}_b=\begin{bmatrix}0.091 & 0.120 & 0.091\\ 0.120 & 0.158 & 0.120\\ 0.091 & 0.120 & 0.091\end{bmatrix}$	$\hat{h}_b=\begin{bmatrix}0.044 & 0.164 & 0.044\\ 0.164 & 0.170 & 0.164\\ 0.044 & 0.164 & 0.044\end{bmatrix}$

The last two examples we show here are intended to test the performance of our algorithm when dealing with nonstationary interference, in the presence of both show-through and bleed-through. In Fig. 5, we show the recto and verso sides of a manuscript affected by a strong interference, and the result obtained by Algorithm 1. The effectiveness of the algorithm is apparent, whereas some distortions due to nonstationarity are still present. In the last experiment (Fig. 6), we also compare the performance of model (1) with the one of the linear convolutional model proposed in [17], where a Markov random field prior was included in the energy function to regularize the solutions. Also in this case, the presence of both show-through and bleed-through is apparent. In the areas where the interference is occluded by the main text, the image reconstructed through the linear model is lighter than the observed image. This is caused by the linear model used, which cannot account for saturation. On the other hand, the interference is removed almost perfectly from the background, by virtue of both the convolutional nature of the

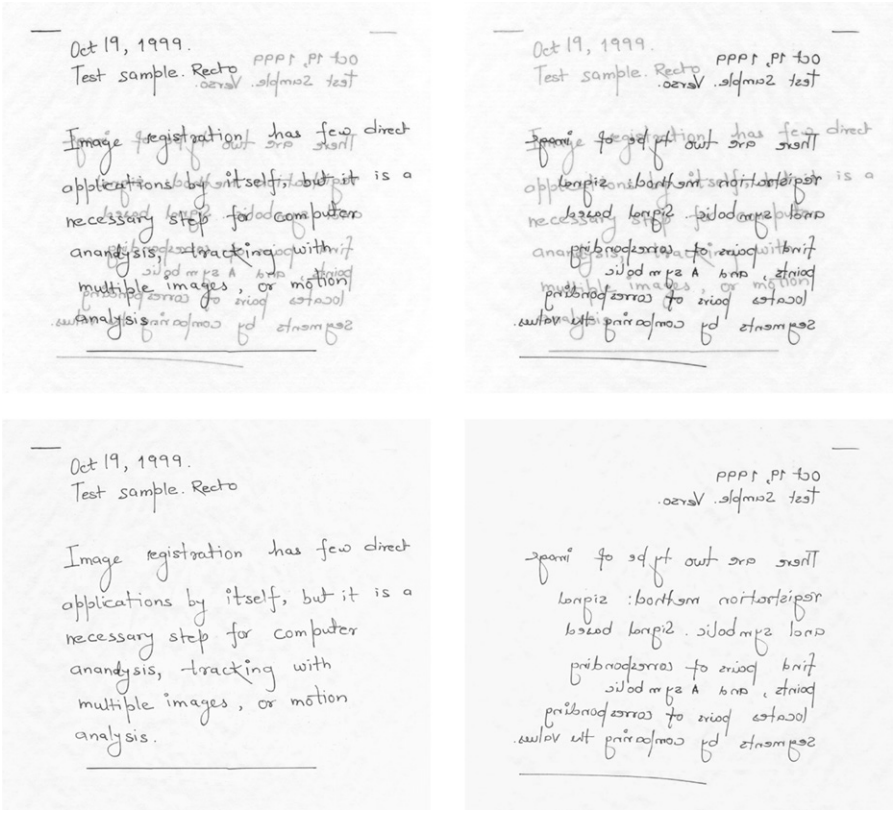

Fig. 3. Removal of back-to-front interferences from a real recto–verso pair (from http://www.site.uottawa.ca/~edubois/documents/). Top: degraded recto and verso; bottom: restored recto and verso by our method.

model and the Markov random field prior, which compensates for nonstationarity. Conversely, the nonlinear convolutional model we are proposing here is able to reconstruct the right intensity at occlusions, but leaves some interference in the background, that is, it cannot cope with a significant nonstationarity. Again, a comparison with a linear instantaneous model can be made by considering the results from the same test pair presented in [14]. The linear strategy is very efficient computationally, and is able to improve greatly the readability of the document, but the fidelity of the result to the original appearance is in any case worse than the ones attainable by more sophisticated (and costly) models.

The problem of nonstationarity is very important, since virtually all the degraded documents are affected by nonstationary distortions. A way to attack this problem could be to estimate the interference level pixel-by-pixel, as envisaged in Appendix B. Another possibility could be to leave the model as in Eq. (1), and provide the energy function with a penalty term that promotes locally smooth reconstructions, as done in [17]. We are currently studying both possibilities.

To have an idea of the computational cost of the strategy we are proposing, we show, in Table 3, the times per iteration needed for a 300×420 image. The code, written in a high-level interpreted language, is not optimized, and has been run on a Pentium D 3 GHz processor. With our synthetic images, and always assuming $q^0 = 5.56$, 1000 iterations at most were needed to converge to the ideal interference level. With real images, however, the model does not explain the data perfectly, and there is no "ideal" interference level to converge to. What we can observe is that the solution stabilizes, although the energy does not vanish. With all the real images shown here, less than 100 iterations are often enough to reach this result. A typical case is the image in Fig. 5, whose size is 2667×1657, for which the total elapsed time was 238 s, for 100 iterations with a 5×5 size PSF.

5. Conclusions

In this paper, we propose an iterative algorithm to restore document images affected by back-to-front interferences through a nonlinear generative model inspired by the one proposed in [11]. The interference in each side is accounted for through the unknown original absorptance of the opposite side, smoothed by an unknown kernel. Once a reasonable estimate of the kernel support is available, the procedure is automatic and fully unsupervised. This is an apparent advantage when a large number of documents are to be processed.

Our solution overcomes some of the drawbacks of the algorithms based on either segmentation–classification approaches or both instantaneous and convolutional linear models. Moreover, our strategy avoids the erosion effect that occurs when the observed images replace the unknown recto and verso patterns in a nonlinear data model. The choice of estimating the convolution kernel at each iteration extends the validity of this technique to documents that are heavily corrupted by back-to-front interferences of whatever origin. This is a fundamental feature to algorithms devoted to ancient document images. An extensive validation with real document images shows that the proposed method compares favorably with other methods proposed in the literature.

The possibility for this method to cope with nonstationary degradations has been envisaged and is now under experimentation. Other possibilities to exploit, in order to apply the method to as many real cases as possible, are to let the PSF account for small residual translational displacements between recto and verso due to registration errors. This is important, since recto–verso registration in general is not an easy task, and residual displacements are always to be expected. We are planning to address all these problems in the future.

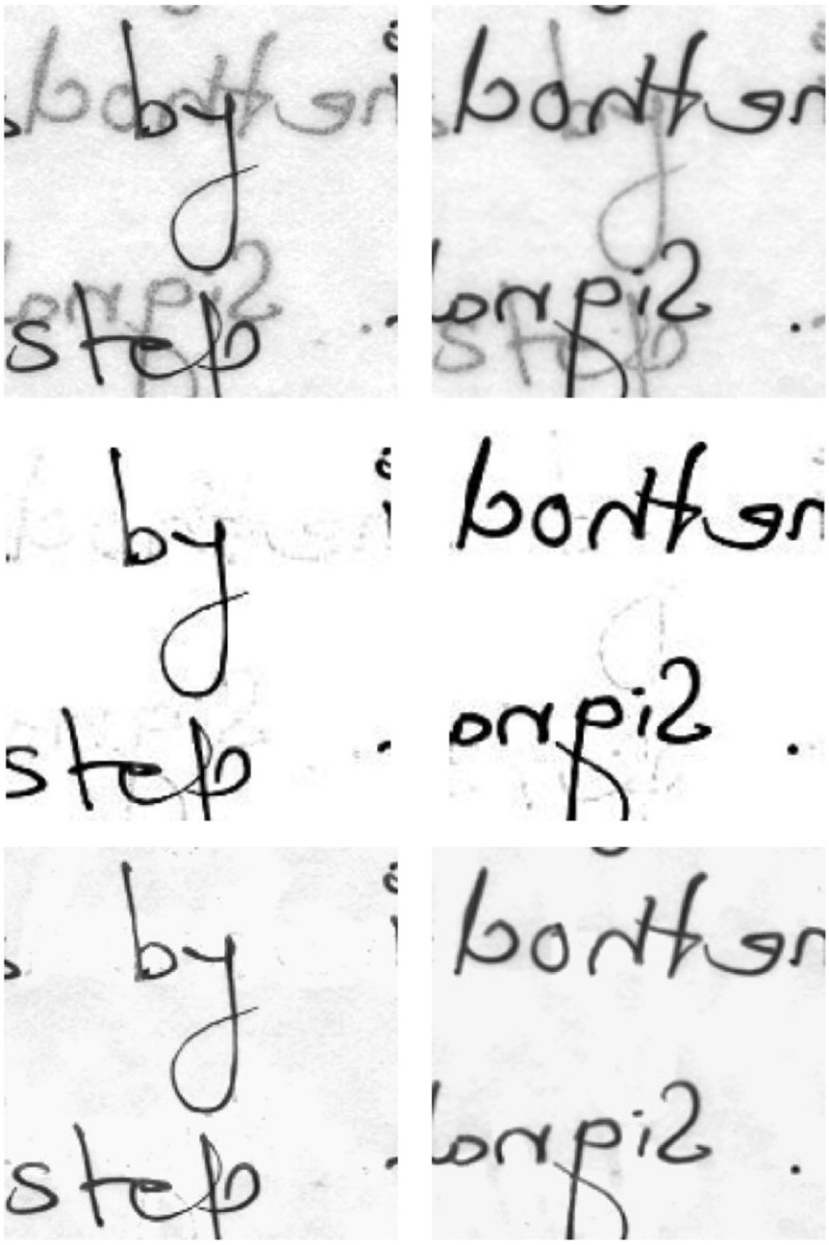

Fig. 4. Comparison between our method and the one proposed in [19]. Top: observed recto and verso data; middle: images restored through the model in [19]; bottom: images restored by our method (detail selected from Fig. 3).

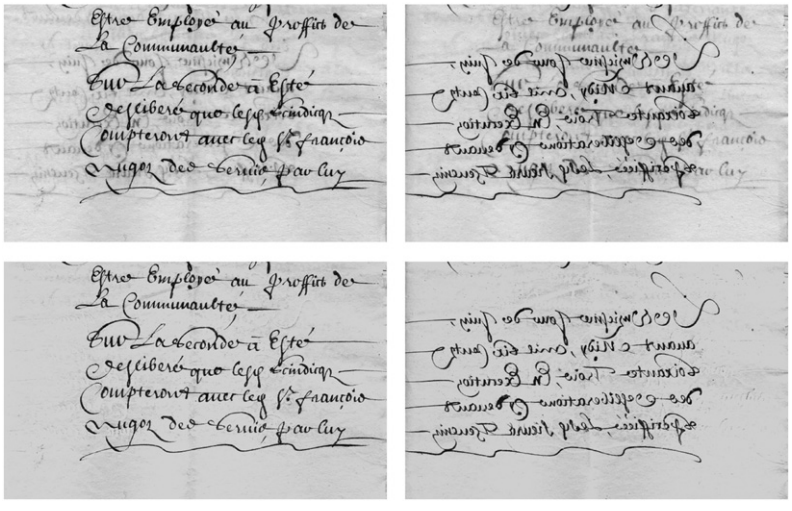

Fig. 5. Removal of back-to-front interferences from a real recto–verso pair. Top left: degraded recto; top right: degraded verso; bottom: restored recto and verso after 100 iterations of Algorithm 1.

Fig. 6. Removal of back-to-front interferences from a real recto-verso pair. Top left: degraded recto; top right: degraded verso; bottom left: restored recto with the nonlinear model; bottom right: restored recto with the linear convolutional model described in [17].

Table 3
Elapsed times per iteration for a synthetic 300 × 420 image.

PSF support size (pixels)	Elapsed time/iteration (ms)
3 × 3	53.4
5 × 5	66.2
7 × 7	85.2
9 × 9	106.6
11 × 11	136.7

Acknowledgments

This work has been supported by European funds, through Program POR Calabria FESR 2007-2013, Project 1220000119 *Ammira* (http://www.ammira.eu). Partners: TEA sas di Elena Console & C., Istituto di Scienza e Tecnologie dell' Informazione CNR, Dipartimento di Meccanica Università della Calabria.

Appendix A. The show-through cancellation algorithm

The first thing we must ensure in adopting the energy function (2) is that, as stated in Section 3, constraining the range of the result eliminates almost all the spurious solutions. Simple numerical experiments give a clear evidence of this fact. We performed these experiments on the example in Fig. 1 with $q=2$, and h uniform with size 3×3. Using the true PSF, fixing q, and minimizing E with no constraints, we get a pair (x_f,x_b) for which $E=0$. Then, if we project this solution onto the constraint set, we generally find a nonzero energy value. The energy values obtained for q decreasing from 5 to 0 (normalized for the total number of pixels) are plotted in Fig. A1, on the left. Note that the energy only vanishes for the true value of q (that is, $q=2$) and for $q=0$. Conversely, if we fix a different PSF, e.g. a random one, the energy exhibits a local minimum for some $q\neq 2$, which then gives wrong reconstructions, and again a global minimum for $q=0$ (see Fig. 2, right). For each q, we also computed the percentage of pixels whose unconstrained values are outside the admitted range, and the average deviations of those values from the bounds of the admitted range. The result is shown in Fig. A2. Note that the average deviation of the out-of-range pixels when $q=2$ is very small. It is apparent that, when the PSF is estimated correctly, the energy function admits two constrained global minimizers, one in the theoretical interference level and one in the trivial solution $q=0$. In practice, the minimum at $q=2$ is not exactly zero for numerical inaccuracies.

From these experiments, some empirical considerations can be drawn, for the case where the data model is fully satisfied. If we do not constrain the problem, our energy function has (i) a global minimum in the desired solution $(\overline{x}_f,\overline{x}_b,\overline{q}_f,\overline{q}_b,\overline{h}_f,\overline{h}_b)$, (ii) an infinity of global minima in the set of trivial solutions $(x_f=x_f^s,x_b=x_b^s,q_f=q_b=0,h_f=\text{any},h_b=\text{any})$, and (iii) an infinity of global minima, one for each arbitrary choice of any parameter subset. If we constrain the solution as shown in Section 3, the minima (iii) become local minima. The problem is thus to find a global minimization algorithm for Eq. (2) that avoids the trivial solutions. Since, as already mentioned, the energy value in the desired solution is not exactly zero, any truly global minimization algorithm invariably drives the solution towards one of the minima in

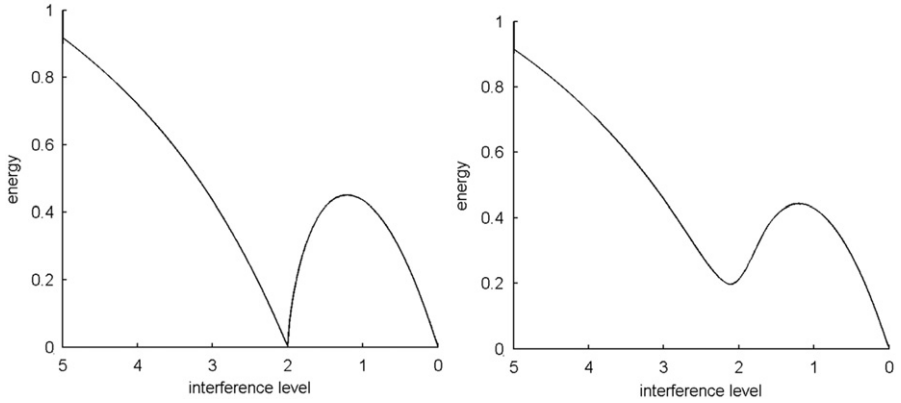

Fig. A1. Energy values obtained by projecting the unconstrained minimizers of function (2) onto the constraint set, as a function of q, adopting either the correct PSF (left) or a random PSF (right).

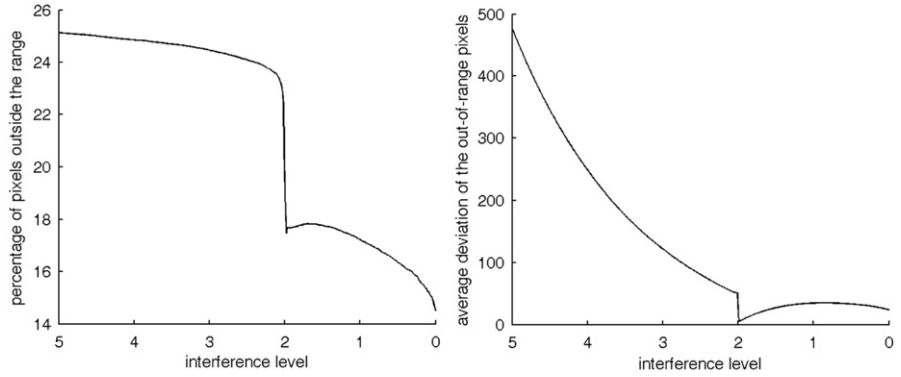

Fig. A2. Left: percentage of pixels out of the admitted range; right: average deviation of the out-of-range pixels.

(ii), independently of the initial guess. To avoid this behavior and enforce convergence to the correct solution, we need a way to condition the variations of q_f, q_b, h_f and h_b upon the solutions x_f and x_b they give rise to. Procedure (3) – (4) gets this result when it starts from the largest possible values for q_f and q_b, and the steps used to produce the successive guesses are sufficiently small. Indeed, each update of the interference level and the PSFs is made while the recto and verso estimates are clamped to their optimal values at the preceding iteration. This means that, once the correct values for q_f and q_b are reached, all the successive updates make the energy increase, and are not accepted by the Monte Carlo Scheme. Choosing good increment steps is not difficult, and is largely independent of the images being processed. Indeed, we used the same values in all our experiments, as shown in Section 4.

Appendix B. Dropping stationarity

As anticipated, the interference affecting the documents in most cases is nonstationary. In terms of our data model, this means that the interference levels are not constant throughout the image and that the show-through PSFs are replaced by space-variant kernels. Dealing properly with this problem is very important, as it is related to the practical applicability of any restoration strategy. As a first step towards a more complete solution, we are assuming stationary PSFs and nonstationary values for the interference levels q_f and q_b. In this way, the data model becomes

$$x_f^s(i,j) = x_f(i,j) \cdot \exp\left\{-q_b(i,j)\left[h_b(i,j) \otimes \left(1 - \frac{x_b(i,j)}{R_b}\right)\right]\right\}$$

$$x_b^s(i,j) = x_b(i,j) \cdot \exp\left\{-q_f(i,j)\left[h_f(i,j) \otimes \left(1 - \frac{x_f(i,j)}{R_f}\right)\right]\right\}$$

$$i = 1,\ldots,N, \quad j = 1,\ldots,M \tag{B.1}$$

and the energy minimization scheme could still be the one of (3)–(5). At step (c) of Algorithm 1, however, the random update of the scalars q_f and q_b should be replaced by a deterministic update of $q_f(i,j)$ and $q_b(i,j)$, $\forall(i,j)$:

$$\tilde{q}_f^{n+1}(i,j) = \tilde{q}_f^n(i,j) - \xi\left(\tilde{q}_f^n(i,j) - \frac{\log\left(\frac{x_b^k(i,j)}{x_b^s(i,j)}\right)}{h_f^k(i,j) \otimes \left(1 - \frac{x_f^k(i,j)}{R_f}\right)}\right)$$

$$\tilde{q}_b^{n+1}(i,j) = \tilde{q}_b^n(i,j) - \xi\left(\tilde{q}_b^n(i,j) - \frac{\log\left(\frac{x_f^k(i,j)}{x_f^s(i,j)}\right)}{h_b^k(i,j) \otimes \left(1 - \frac{x_b^k(i,j)}{R_b}\right)}\right)$$

$$i = 1,\ldots,N, \quad j = 1,\ldots,M \tag{B.2}$$

where $\tilde{q}_f^0 = q_f^k$, $\tilde{q}_b^0 = q_b^k$, and ξ is a scalar smaller than 1 to ensure convergence. The estimates q_f^{k+1} and q_b^{k+1} are then set to the values of $\tilde{q}_f^n$ and $\tilde{q}_b^n$ at convergence. We are now checking the effectiveness of (B.2) on real documents.

References

[1] H. Nishida, T. Suzuki, A multiscale approach to restoring scanned color document images with show-through effects, in: Proceedings of the ICDAR 2003, 2003, pp. 584–588.

[2] Q. Wang, T. Xia, L. Li, C.L. Tan, Document image enhancement using directional wavelet, in: Proceedings of the IEEE Conference on Computer Vision Pattern Recognition, vol. 2, 2003, pp. 534–539.
[3] Y. Leydier, F.L. Bourgeois, H. Emptoz, Serialized unsupervised classifier for adaptive color image segmentation: application to digitized ancient manuscripts, in: Proceedings of the International Conference on Pattern Recognition, 2004, pp. 494–497.
[4] F. Drida, F.L. Bourgeois, H. Emptoz, Restoring ink bleed-through degraded document images using a recursive unsupervised classification technique, in: Proceedings of the 7th Workshop on Document Analysis Systems, 2006, pp. 38–49.
[5] C. Wolf, Document ink bleed-through removal with two hidden Markov random fields and a single observation field', Technical Report RR-LIRIS-2006-019, Laboratoire d'Informatique en Images et Systémes d'Information, INSA, Lyon, France, November 2006.
[6] P. Dano, Joint restoration and compression of document images with bleed-through distortion, Master's Thesis, Institute for Electrical and Computer Engineering, School of Information Technology and Engineering, University of Ottawa, Ottawa-Carleton, June 2003.
[7] K. Knox, Show-through correction for two-sided documents, United States Patent 5 832 137, November 1998.
[8] Q. Wang, C.L. Tan, Matching of double-sided document images to remove interference, in: Proceedings of the IEEE CVPR 2001, 2001, p. 1084.
[9] E. Dubois, A. Pathak, Reduction of bleed-through in scanned manuscript documents, in: Proceedings of the IS&T Image Processing, Image Quality. Image Capture Systems Conference, 2001, pp. 177–180.
[10] R.F. Moghaddam, D. Rivest-Hénault, I. Bar-Yosef, M. Cheriet, A unified framework based on the level set approach for segmentation of unconstrained double-sided document images suffering from bleed-through, in: ICDAR, Barcelona, Spain, 2009, pp. 441–445. doi:10.1109/ICDAR.2009.108.
[11] G. Sharma, Show-through cancellation in scans of duplex printed documents, IEEE Trans. Image Process. 10 (2001) 736–754.
[12] R.F. Moghaddam, M. Cheriet, A variational approach to degraded document enhancement, IEEE Trans. Pattern Anal. Mach. Intell. 32 (8) (2010) 1347–1361.
[13] A. Tonazzini, L. Bedini, E. Salerno, Independent component analysis for document restoration, IJDAR 7 (2004) 17–27.
[14] A. Tonazzini, E. Salerno, L. Bedini, Fast correction of bleed-through distortion in grayscale documents by a blind source separation technique, IJDAR 10 (2007) 17–25.
[15] A. Tonazzini, G. Bianco, E. Salerno, Registration and enhancement of double-sided degraded manuscripts acquired in multispectral modality, in: Proceedings of the 10th International Conference on Document Analysis and Recognition ICDAR 2009, 2009, pp. 546–550.
[16] F. Su, A. Mohammad-Djafari, Bayesian separation of document images with hidden Markov model, in: Visapp-Visigrapp, 2007, pp. 151–156.
[17] A. Tonazzini, I. Gerace, F. Martinelli, Multichannel blind separation and deconvolution of images for document analysis, IEEE Trans. Image Process. 19 (4) (2010) 912–925.
[18] B. Ophir, D. Malah, Show-through cancellation in scanned images using blind source separation techniques, in: Proceedings of the International Conference on Image Processing ICIP, vol. III, 2007, pp. 233–236.
[19] F. Merrikh-Bayat, M. Babaie-Zadeh, C. Jutten, A nonlinear blind source separation solution for removing the show-through effect in the scanned documents, in: Proceedings of the EUSIPCO 2008, 2008.
[20] R. Moghaddam, M. Cheriet, Low quality document image modeling and enhancement, Int. J. Doc. Anal. Recognition 11 (4) (2009) 183–201.

Francesca Martinelli was born in Umbertide, Italy, in 1980. She received the degree in Mathematics from the University of Perugia, Italy, in 2005, and the Ph.D. degree in Mathematics and Computer Science for Information and Knowledge Management from the University of Perugia, in 2009. In 2009, she has been a Post Doctoral Fellow at the Dipartimento di Matematica e Informatica, University of Perugia, and is currently a research fellow at the Institute of Information Science and Technologies of CNR in Pisa, Italy. Her research interests are in blind image deconvolution, document analysis and recognition, blind source separation.

Emanuele Salerno graduated in electronic engineering from the University of Pisa, Italy, in 1985. In 1987, he joined the Italian National Research Council as a full-time researcher. At present, he is a senior researcher at the Institute of Information Science and Technologies in Pisa, Signals and Images Laboratory. He has been working in applied inverse problems, image reconstruction and restoration, microwave nondestructive evaluation, and blind signal separation, and held various responsibilities in research programs in nondestructive testing, robotics, numerical models for image reconstruction and computer vision, and neural network techniques in astrophysical imagery. He has been supervising various theses in Computer Science, Electronic and Communications Engineering, and Physics, and, since 1996, has been teaching courses at the University of Pisa. Dr Salerno is a member of the Italian society for information and communication technology (AICT-AEIT).

Ivan Gerace received the degree in Computer Science from the University of Pisa, Italy, in 1992, and the Ph.D. degree in Computational Mathematics and Operational Research from the University of Milan, Italy, in 1999. He has been an assistant professor in Numerical Analysis since May 2000 at the University of Perugia, Italy. His research interests include Image Processing, Numerical Linear Algebra and Computational Complexity.

Anna Tonazzini received the degree in Mathematics (cum laude) from the University of Pisa, Italy, in 1981. She is a researcher at the Signals and Images Laboratory of the Istituto di Scienza e Tecnologie dell'Informazione, Italian National Research Council (CNR) in Pisa. She cooperated in several projects for basic and applied research on image processing and computer vision and is co-author of over 80 scientific papers. Her present interests are in applied inverse problems, multichannel image restoration and reconstruction, document analysis and recognition, blind source separation, computational biology.