ORIGINAL PAPER

A non-stationary density model to separate overlapped texts in degraded documents

Anna Tonazzini · Pasquale Savino · Emanuele Salerno

Abstract We address the problem of the removal of a text superimposed to a more important one, in a document image, considering the two instances of canceling back-to-front interferences from recto and verso images of archival documents and of recovering the erased text in palimpsests from multispectral images. Both problems are approached through a model where the ideal images of the two texts are considered as individual source patterns, mixed through some parametric operator. To cope with occlusions, ink saturation, and space variability of the mixing operator, a data model for this problem should be nonlinear and space variant. Here, we show that if a pointwise non-stationarity is allowed, a linear model can compensate for the lack of a suitable nonlinearity and for other modeling errors.

Keywords Document restoration · Nonlinear data model · Non-stationary data model · Back-to-front interferences · Palimpsests

1 Introduction

One of the most common degradations affecting archival documents 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 caused by either transparency or seeping of the ink of the text

A. Tonazzini (✉) · P. Savino · E. Salerno
Istituto di Scienza e Tecnologie dell'Informazione-CNR,
Via G. Moruzzi, 1, 56124 Pisa, Italy
e-mail: anna.tonazzini@isti.cnr.it

P. Savino
e-mail: pasquale.savino@isti.cnr.it

E. Salerno
e-mail: emanuele.salerno@isti.cnr.it

printed in the reverse side of the page. Such distortion can significantly degrade the readability of the document.

Several approaches for see-through reduction have been already investigated, the most effective being those that exploit information from scans or images of both sides (*recto* and *verso*). The earliest solutions proposed were based on thresholding [1], wavelet techniques for enhancing the foreground strokes and smearing the interferences [2], or segmentation and classification [3]. Other more recent methods use energy minimization. In [4], the Chan-Vese active contour model is modified to incorporate information from both sides of the document, and a subsequent function minimization technique is then applied to correct broken strokes where the two texts overlap. In [5], to classify the pixels in the two sides as foreground, bleed-through or background, a regularized energy function is defined that uses as data term likelihoods derived from small sets of user-labeled pixels, and as smoothness term a dual-layer Markov Random Field (MRF). A 4-class classification approach is proposed in [6], by segmenting the recto–verso joint histogram with the aid of a smoothness term built from a learning set based on the availability of ground truths. Possible pixel misclassifications are then iteratively corrected by analyzing the connected components in the segmented image, and the regions classified as bleed-through are inpainted with suitable patterns extracted from the background texture.

Model-based techniques have also raised much interest. The recto and verso appearances of the degraded document are modeled as two parametric superimpositions of the uncorrupted front- and rear-side images, to be recovered by blind or semi-blind source separation techniques. When a linear mixing model is assumed, fast algorithms are available, such as those based on Independent Component Analysis (ICA) and data decorrelation [7,8]. Within this approach, compensations for the apparent nonlinear-

ity and/or non-stationarity of the physical phenomenon have used regularization techniques using MRF [9,10] or penalized nonnegative matrix factorization (NMF) [11]. In [12], a nonlinear convolutional mixing model is derived by approximating the physical model of the show-through in modern scanners, and adaptive linear filters are used to estimate the unknown model parameters. Based on the same model with known nonlinearity, in [13], the problem is regularized by a total variation stabilizer. In [14] and [15], we experimented this model for bleed-through as well, employing a constrained maximum likelihood technique for blindly estimating both the mixing parameters and the source images, or estimating off-line the model parameters. A quadratic mixing model accounting for a blur kernel on the interfering patterns is assumed for show-through in [16] and solved through maximum likelihood. Still for show-through, a nonlinear but invertible model is proposed in [17], whose parameters are learned by nonlinear ICA. Finally, in [18], variational approaches, based on nonlinear diffusion and wavelet transforms, have been proposed to model and remove bleed-through.

Another issue in ancient document analysis is revealing the whole contents of a document, which can aid scholars in dating or authenticating. Often, interesting document features are hidden or barely detectable in the original color document. Although these features can sometimes be highlighted by multispectral acquisitions in the non-visible range (e.g., faint traces of the erased text in palimpsested manuscripts are often enhanced by ultraviolet fluorescence), they are still overlapped to other paper contents. It is thus important to look for digital image processing strategies to make the pattern of interest more readable and possibly free from interferences. Thus, as for the case of see-through in recto–verso pairs, also multispectral images of palimpsested manuscripts can be viewed as mixtures of two individual patterns, to be separately extracted. According to this point of view, a unified approach to the two seemingly different problems of see-through removal and underwriting extraction can be taken, where the choice of a suitable mixing model is crucial.

Nevertheless, we believe that a comprehensive and feasible model for text overlapping in ancient documents is still far from being available. Whereas nonlinearity can be useful to describe text superposition in the occlusion areas, non-stationarity is important as well. Indeed, whereas in the scans of modern documents it is expected that the percentage of the rear-side text shining through the front-side (and vice versa) is almost constant throughout the page, the seeping of ink in bleed-through can be highly variable in space, due, for example, to humidity spots. Similarly, in palimpsests, usually very old manuscripts, the aging process is likely to have produced irregular erosion and/or diffusion of the materials. We argue that the use of accurately estimated pixel-dependent parameters, even within an additive or whatever (reasonable) model, could compensate not only for non-stationarity, but also for the lack or the imprecise knowledge of the nonlinearity. By using the same nonlinearity as in [15], we tested this possibility in [19] and in [20] within a regularization framework.

Here, we propose a non-stationary linear model combining the ideal optical densities of the two individual texts, to describe two of the available observations. In the see-through case, these observations will be the recto and verso images, both acquired in the same wavelength range, and the model includes blur kernels to account for the smearing of the interfering patterns, due to light transmission or ink diffusion through the medium. In the case of palimpsests, these observations will be two images of the manuscript, acquired at two different channels, and the model is assumed to be instantaneous. The challenge is to estimate the pixel-dependent coefficients of the linear mixture. We propose to estimate them from the data alone, based on simple and intuitive criteria. A few characteristics that are peculiar of each problem will entail the adoption of some minor modifications in the models and, consequently, in the related way to estimate those coefficients. The algorithms derived are very fast and only require a little user intervention to estimate, still from the data, mean background values and the blur kernels. However, as explained later on, the algorithms could be easily made fully automatic and even faster. Another interesting feature of our method is that it meets the requirement, fundamental when dealing with the restoration of ancient documents, to remove interferences while preserving the original appearance (i.e., color and texture) of the characters and the background. For the see-through case, we show experimental results that are competitive with those of more complex recent methods.

The paper is organized as follows. In Sect. 2, we describe the data model we consider when the application is see-through removal, discuss in detail our strategy to estimate the model parameters, and derive the scheme employed to estimate the two ideal texts. The adaptation of the data model and the restoration algorithm to the instance of underwriting recovery from palimpsests is summarized in Sect. 3. Section 4 analyzes some experimental results for both applications. Finally, Sect. 5 concludes the paper, by discussing some ideas that could help to overcome the present limitations of the method.

2 See-through interference removal: data model and estimation strategy

Assuming preregistered recto and verso images, the recto–verso data model we consider is the following non-stationary, linear convolutional model:

$$D_r^{obs}(t) = D_r(t) + q_v(t)\,[h_v(t) \otimes D_v(t)]$$
$$D_v^{obs}(t) = D_v(t) + q_r(t)\,[h_r(t) \otimes D_r(t)]$$
$$t = 1, 2, \ldots, T \qquad (1)$$

with

$$D(t) = -log\left(\frac{s(t)}{R}\right) \qquad (2)$$

where $D_r^{obs}(t)$ and $D_v^{obs}(t)$ are the observed optical densities, and $D_r(t)$ and $D_v(t)$ are the ideal optical densities, of the front and back side, respectively, at pixel t. Each density is related to the corresponding (ideal or observed, recto or verso) reflectance s through Eq. (2), involving a suitable constant R that represents, in the two instances R_r and R_v, the mean reflectance values of the background in the recto and verso side, respectively. The model also includes two unit volume point spread functions (PSF), h_r and h_v, describing the smearing of ink that penetrates or shines through the paper, thus allowing a pattern in a side to match the corresponding one in the opposite side. These PSFs are stationary throughout the image, but characterized by different gains $q_r(t)$ and $q_v(t)$, which, while representing our way to account for non-stationarity, have also the physical meaning of interference levels from the front to the back and from the back to the front, respectively, at each pixel.

In order to invert system (1) for D_r and D_v, the model parameters R_r and R_v, $q_r(t)$ and $q_v(t)$, $\forall t$, and h_r and h_v, need to be estimated. We do this off-line, by exploiting a minimum intervention by the user. R_r and R_v are computed by averaging the pixel intensity values within two selected pure background areas. Note that fully automatic procedures to estimate mean background values could also be easily devised for document images, usually characterized by dark texts on light backgrounds. As per the PSFs, normally, if the two sides are perfectly aligned, it is reasonable to assume that both h_r and h_v have the form of a centered, gaussian-like function. To estimate h_v, we isolate a small area in the recto side, where pure see-through is present ($D_r = 0$) and the interference levels can be assumed almost constant. Using also the corresponding area of pure foreground text in the opposite verso side, h_v can be estimated from the two areas by constrained least squares [21], as detailed in [15], and h_r can be estimated in specular way. In the presence of a possible small residual displacement between the two sides, due to registration errors, the maximum of h_v will appear shifted of a corresponding value, which can be used to correct the misalignment, by just translating one of the two sides. We have experimentally verified that, in the majority of the cases, it is $h_r = h_v$, so that the above procedure can be executed for a side only. Furthermore, for documents belonging to the same class, e.g., same ink and paper used, even a fixed

Gaussian PSF can be effectively used, thus avoiding the need for user intervention.

As per the estimation of the space-variant maps $q_r(t)$ and $q_v(t)$ at each pixel t, we adopted the following formulas:

$$q_r(t) = \frac{D_v^{obs}(t)}{h_r(t) \otimes D_r^{obs}(t) + \epsilon}$$
$$q_v(t) = \frac{D_r^{obs}(t)}{h_v(t) \otimes D_v^{obs}(t) + \epsilon} \qquad (3)$$

where ϵ is a small positive number to avoid indeterminacies or infinity.

It is apparent that Eq. (3) make both sense when the ideal density is zero, that is in the pixels corresponding to the background in both sides. For the pixels of pure see-through in one side and foreground text in the opposite side only one of the two equations makes sense. Finally, for pixels corresponding to occlusion areas none of the two equations is valid. More specifically, given a pixel t, four situations can occur: 1) t is a background pixel in both the recto and verso side, 2) t is a pure show-through pixel in the recto side and foreground text in the verso side, 3) t is a foreground text pixel in the recto side and pure show-through in the verso side, and 4) t is an occlusion pixel. Unfortunately, it is in general impossible to determine a priori which of the four situations a pixel corresponds to. Therefore, we first compute q_r and q_v at each pixel t according to Eq. (3), and then, based on simple heuristic criteria, correct the possible wrongly computed interference levels, thus also discriminating the various situations.

Ideally, in the background pixels (case 1), we would like Eq. (3) to give null interference levels, so that the reconstructed images retain there the same values of the observed images, thus preserving at best the original appearance. This should occur automatically, since in those pixels both densities should be zero. Nevertheless, due to small fluctuations around the mean background values, and to the fact that the density is always close to 0 in those areas, even high values of $q_r(t)$ and $q_v(t)$ can be obtained. When inverting the system of Eq. (1), this might lead to negative densities, which however simply means reconstructed background values greater than R. Alternatively, the negative values of the estimated density can be set to zero, or to the density of the lightest pixel in the data.

For the pixels of foreground text in one side and see-through in the opposite side (cases 2 and 3), we would like to obtain $q_v > 0$, $q_r = 0$ in case 2 and $q_v = 0$, $q_r > 0$ in case 3. In other words, q_r and q_v should be mutually exclusive there. Nevertheless, we will obtain $q_r > q_v > 0$ in case 2, since the first of Eq. (3) does not hold true, and, symmetrically, $q_v > q_r > 0$ in case 3. However, correcting this error is trivial: it suffices, at each t, to maintain the smallest between the two computed interference levels, and set the other to zero, thus automatically discriminating cases 2 and 3.

The last possible situation occurs in the occlusion pixels (case 4), i.e., when the densities in the two sides are both high since the two texts overlap. According to the data model (1), a correct estimation of the interference levels should account for the ideal density, which is however unknown. Assuming instead the ideal density to be zero, as done in (3), would overestimate the interference level thus producing "holes" in correspondence of the occlusion areas. However, provided that the two inks in the two sides reflect similarly under the same wavelength, we can expect that the densities of the occlusion areas in the two sides are high and close to each other. If the degradation is not the strongest possible, i.e., interferences as dark as foreground text, only the pixels in the occluding areas have this property, so that they can be located. In the most obvious way, this can be accomplished with the aid of manually selected thresholds. However, in many cases, we have found that a simple procedure allows us to automatically discriminate the occlusion pixels: if the gray levels of the two foreground texts are similar, we compute the absolute difference between the reflectance maps of the recto and verso observations, and then binarize (e.g., through the Otsu algorithm [22]) the difference map. This will return a binary map where the zero pixels correspond either to the occlusion pixels or to the background pixels, where both interference levels will be set to zero.

Once the model parameters are known, D_r and D_v are computed through the following single step:

$$
\begin{aligned}
D_r(t) &= D_r^{obs}(t) - q_v(t)\left[h_v(t) \otimes D_v^{obs}(t)\right] \\
D_v(t) &= D_v^{obs}(t) - q_r(t)\left[h_r(t) \otimes D_r(t)\right]
\end{aligned}
\tag{4}
$$

3 Recovery of underwriting in palimpsests: data model and estimation strategy

In the case of palimpsests, we assume the availability of multiple observation channels. Let us consider two such channels, one centered at wavelength λ_1 and the other at wavelength λ_2. Assuming preregistration of the two images, and neglecting the support (paper or parchment), i.e., considering that its density is zero at the two wavelengths, we can write:

$$
\begin{aligned}
D_{\lambda_1}^{obs}(t) &= D_{\lambda_1}^{u}(t) + D_{\lambda_1}^{o}(t) \\
D_{\lambda_2}^{obs}(t) &= D_{\lambda_2}^{o}(t) + D_{\lambda_2}^{u}(t)
\end{aligned}
\tag{5}
$$

where $D_{\lambda_1}^{obs}(t)$ and $D_{\lambda_2}^{obs}(t)$ are the observed maps of the optical density at wavelengths λ_1 and λ_2, respectively, and at pixel t, and D^o and D^u are the ideal optical densities of the overwriting and underwriting texts, observed at wavelength λ_1 in the first equation and at wavelength λ_2 in the second equation. Let us consider two positive, space-variant coefficients, $q^o(t)$ and $q^u(t)$, such that the following relationships are satisfied:

$$
\begin{aligned}
D_{\lambda_1}^{o}(t) &= q^o(t)D_{\lambda_2}^{o}(t) \\
D_{\lambda_2}^{u}(t) &= q^u(t)D_{\lambda_1}^{u}(t)
\end{aligned}
\tag{6}
$$

Hence, Eq. (5) can be rewritten as:

$$
\begin{aligned}
D_{\lambda_1}^{obs}(t) &= D_{\lambda_1}^{u}(t) + q^o(t)D_{\lambda_2}^{o}(t) \\
D_{\lambda_2}^{obs}(t) &= D_{\lambda_2}^{o}(t) + q^u(t)D_{\lambda_1}^{u}(t)
\end{aligned}
\tag{7}
$$

As in Eq. (1), the pixel-dependent parameters $q^o(t)$ and $q^u(t)$ account for the modeling errors, even if in this case, they do not have any precise physical meaning. Note that no blur kernels are included in the above data model. In fact, the filters used for imaging the palimpsested manuscript at the two wavelengths might have different transfer functions, which means that the maps D^o and D^u do not match perfectly in the two equations above. However, since, differently from the see-through case, here the blur kernels affect the whole observations and are known, they can be removed off-line. Thus, without losing generality, we can assume D^o and D^u to be at the same resolution in both equations.

To invert system (7) and estimate the ideal overwriting and underwriting texts, the same scheme of (4) can be used, once the two maps q^o and q^u are known. In particular, we expect to obtain the ideal underwriting image as depicted at wavelength λ_1 from the first of Eq. (7), i.e., from the first data image, and the ideal overwriting image as it appears at wavelength λ_2 from the second of Eq. (7), i.e., from the second data image.

To estimate the two maps q^o and q^u, we first observe that, given a pixel t, one of four situations occurs in both the observations: 1) t is a background pixel, 2) t is an underwriting pixel, 3) t is an overwriting pixel, and 4) t is an occlusion pixel. In general, to determine a priori which of the four situations a pixel corresponds to is more difficult than in the see-through case. However, the two wavelengths λ_1 and λ_2 can be chosen to drive the estimation of the correct $q^o(t)$ and $q^u(t)$.

In palimpsests, the two different writings normally feature different spectral signatures, thus tending to fade over different wavelength ranges. Often, there exists a wavelength where the underwriting is more visible than at the others, although still mixed to the overwriting. We thus choose this wavelength as λ_1, to obtain the first data image. Furthermore, there exists another wavelength where the underwriting almost disappears. We choose this second wavelength as λ_2; it is apparent that from this second data image, the overwriting can easily be located. Neglecting for the moment the fact that the overwriting includes also the occlusions, we can assume that, at the locations of the overwriting pixels, the ideal first image should have null density, so that q^o can be computed at those locations as the ratio between $D_{\lambda_1}^{obs}$ and $D_{\lambda_2}^{obs}$, and set to zero elsewhere. Now we know that the

underwriting, except the occlusions, is located somewhere outside the overwriting, in both images. Consequently, since the density of the ideal second image should be zero in those pixels, we set q^u as the ratio between $D_{\lambda_2}^{obs}$ and $D_{\lambda_1}^{obs}$ where q^o is zero, and zero elsewhere, i.e., in correspondence of the overwriting. In formulas, it is:

$$q^o(t) = \begin{cases} \dfrac{D_{\lambda_1}^{obs}(t)}{D_{\lambda_2}^{obs}(t)+\epsilon} & \text{if } t \text{ is overwriting} \\ \\ 0 & \text{elsewhere} \end{cases} \tag{8}$$

$$q^u(t) = \begin{cases} \dfrac{D_{\lambda_2}^{obs}(t)}{D_{\lambda_1}^{obs}(t)+\epsilon} & \text{if } q^o(t) = 0 \\ \\ 0 & \text{elsewhere} \end{cases} \tag{9}$$

where ϵ is a small positive number to avoid indeterminacies or infinity.

Note that, in this way, q^u is forced to zero at the occlusion pixels, thus preserving the occlusion areas of the overwriting, but can be positive and high in the background pixels. By managing this latter situation as done for the see-through case, scheme (4) will return a map where the overwriting is extracted cleansed from the faint traces of the older text, also retaining the values of the data in the occlusions. Conversely, q^o is forced to zero in the background pixels, thus preserving the textures of the background outside the underwriting, but has wrong positive and possibly high values at the occlusion pixels. Indeed, in those areas, the density observed at wavelength λ_1 is normally higher than that observed at wavelength λ_2. Thus, scheme (4) will return a map where the underwriting is extracted cleansed from the even strong strokes of the overwriting, but with "holes" in the occlusion areas. To solve this problem, we assume herein that, at wavelength λ_1, the occlusion pixels are significantly darker than the other pixels, which leads to the highest values of q^o. Thus, in those pixels where q^u is zero and q^o exceeds the average of its positive values, we set q^o to its inverse value. In this way, the occlusion pixels of the cleansed underwriting map are attenuated without reaching the background value. More efficiently, the occlusions so located can be filled in both maps with the average value of the recovered texts outside them, or, equivalently, these average values can be used to refine q^o and q^u through inversion of the complete data equation.

4 Discussion of the experimental results

We experimented the method proposed above for the separation of overlapped texts in recto–verso documents and for the extraction of the underwriting from multispectral images of palimpsests. The algorithms derived are very fast since the data model can be inverted in a single step.

As per the see-through problem, we tested the method on several grayscale real recto–verso pairs, affected by either show-through, bleed-through, or both. In the following, we analyze in detail the results obtained in some of such experiments, by also providing comparisons with the results obtained by a stationary linear model, the stationary nonlinear model in [15], and other state-of-the-art methods.

Figure 1a, e shows the original degraded recto and verso images. It is not easy to decide whether the degradation is caused by show-through, bleed-through, or both. As apparent, the interfering patterns are highly non-stationary. The results of the application of ICA, improved by histogram clipping, are shown in Fig. 1b, f, whereas Fig. 1c, g shows the results obtained with the stationary nonlinear model of [15]. The results of the method proposed in this paper are shown in Fig. 1d, h. The stationary linear model produces lower ink density in the occlusion areas and leaves some interferences as well. On the other hand, the results of the stationary nonlinear model highlight that a single value of the interference level is not sufficient to remove all the interference, although "holes" are not present anymore in the occlusions, for this image. The superior performance of the method proposed here is apparent. In other words, the adoption of a stationary data model, even nonlinear and convolutional, necessarily entails a compromise between see-through removal and preservation of the occlusions. Transforming the model, either linear or nonlinear, from stationary to non-stationary allows us to meet both requirements.

Figure 2 shows the maps of the estimated interference levels q_r and q_v and of the estimated occlusion pixels. Darker pixels in the interference level maps indicate higher values of q_r and q_v. In the third map, black pixels represent occlusion areas.

Recently, a database of high-resolution grayscale images of ancient documents affected by bleed-through has been published online [23]. This database comprises 25 registered recto–verso sample grayscale image pairs, taken from larger high-resolution manuscript images, with varied degrees of bleed-through. In addition, for each image, a binary ground truth mask of the foreground text is provided. Although these ground truth images are synthetic, i.e., created manually, they can be useful for a quantitative analysis of the results. This database is diffusely described in [24].

In Fig. 3, we present the results obtained on one of such images, a manuscript belonging to the Allan and Maria Myers Academic Centre, University of Melbourne (Fig. 3a, d). The results of using the stationary nonlinear model of [15] are shown in Fig. 3b, e, whereas Fig. 3c, f shows the much better results obtained by using the non-stationary, linear model proposed in this paper. From a visual point of view, the higher quality of the reconstructions in this case can be appreciated especially in the recto side. Indeed, in Fig. 3b, the lower ink

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

(e)   **(f)**   **(g)**   **(h)**
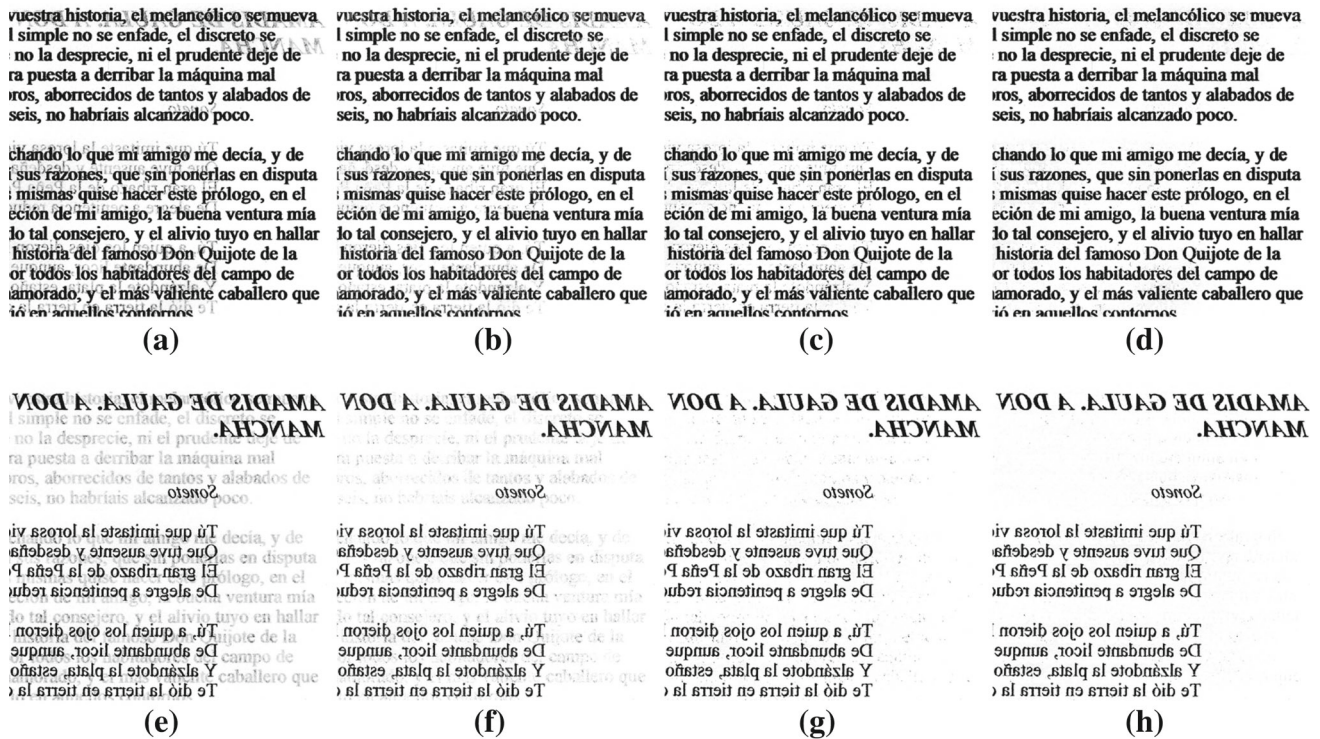
Fig. 1 Restoration of a recto–verso pair of a real document: **a** degraded recto, **b** recto restored with ICA, **c** recto restored with the method in [15] (stationary nonlinear model), **d** recto restored with the proposed method (non-stationary linear model), **e** degraded verso, **f** verso restored with ICA, **g** verso restored with the method in [15] (stationary nonlinear model), **h** verso restored with the proposed method (non-stationary linear model)

(a)   **(b)**   **(c)**
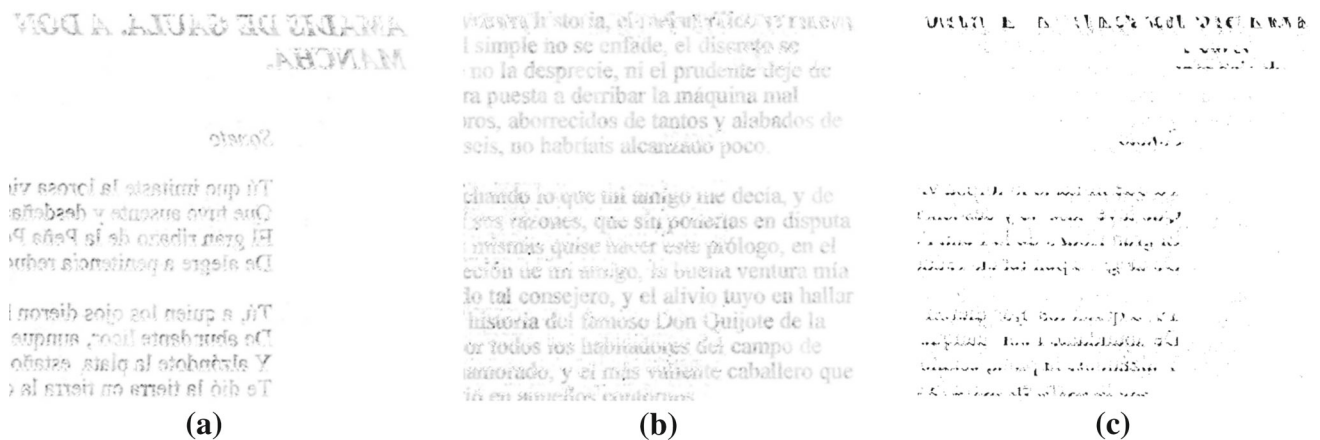
Fig. 2 Maps of the estimated interference levels q_v (**a**) and q_r (**b**) and of the estimated occlusion pixels (**c**), for the images in Fig. 1a, e. *Darker pixels* in the interference level maps indicate higher values; *black pixels* in the third map represent occlusion areas

density in the occlusion areas is well visible, whereas it does not affect the homologous image of Fig. 3c.

In Fig. 4, the maps of the estimated interference levels q_r and q_v and of the estimated occlusion pixels are shown.

Finally, Fig. 5 shows, for a detail of the images of Fig. 3a, d, the comparisons with recent state-of-the-art methods. The recto sides are depicted in the left and the verso sides in the right, respectively. From top to bottom, we see the original degraded images, the results from the methods in [5,10,18], and [6], respectively, and finally, the results from the method proposed herein. It is apparent that the results obtained with our method are fully qualitatively comparable to the best ones, produced by the method in [6].

In another experiment, we processed a second pair from the same dataset [23]. The manuscript belongs to The James Hardiman Library, National University of Ireland, Galway.
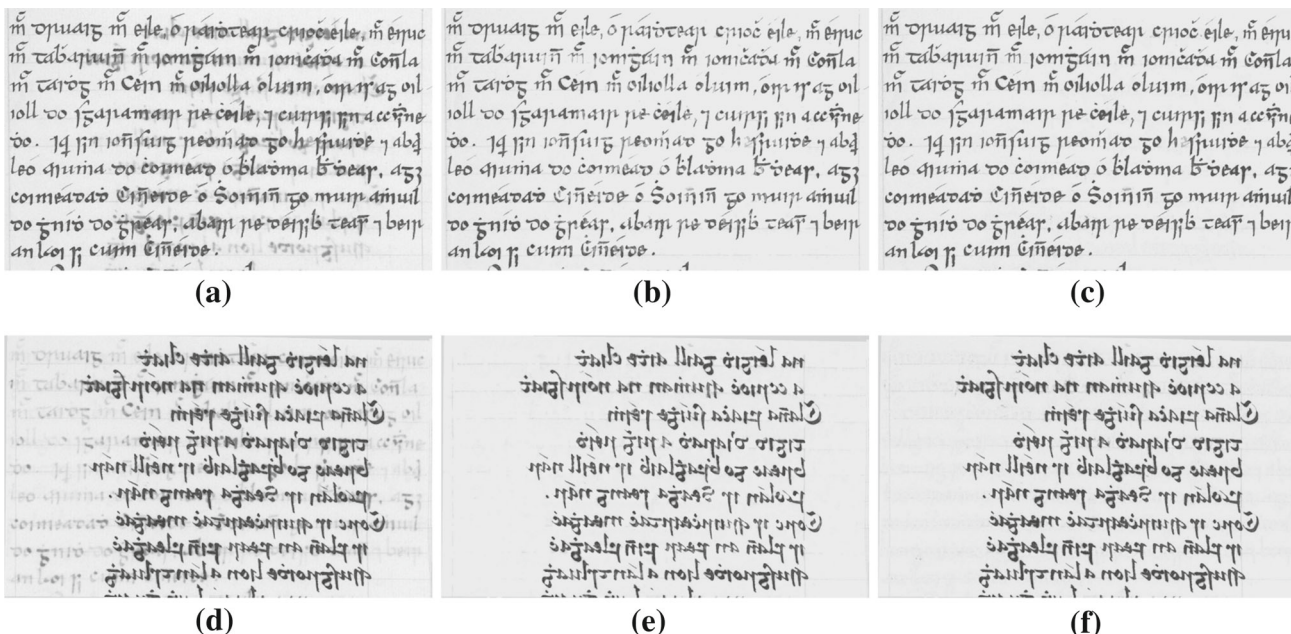
Fig. 3 Application of the proposed method to a real recto–verso pair: **a** original degraded recto, **b** recto restored with the method in [15] (stationary nonlinear model), **c** recto restored with the proposed method (non-stationary linear model), **d** original degraded verso, **e** verso restored with the method in [15] (stationary nonlinear model), **f** verso restored with the proposed method (non-stationary linear model). Original images (**a**) and (**d**): reproduction by courtesy of The Allan and Maria Myers Academic Centre, University of Melbourne, digitized by Irish Script On Screen (www.isos.dias.ie)
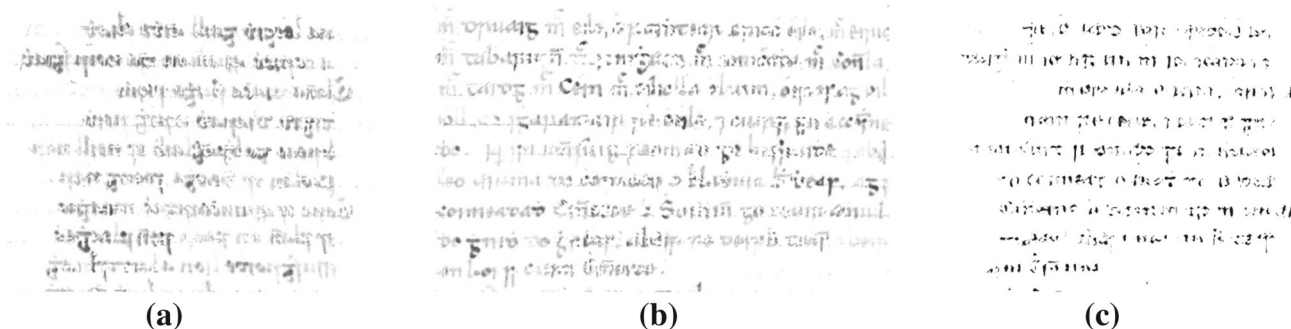
Fig. 4 Maps of the estimated interference levels q_v (**a**) and q_r (**b**) and of the estimated occlusion pixels (**c**), for the images in Fig. 3a, d. *Darker pixels* in the interference level maps indicate higher values; *black pixels* in the third map represent occlusion areas
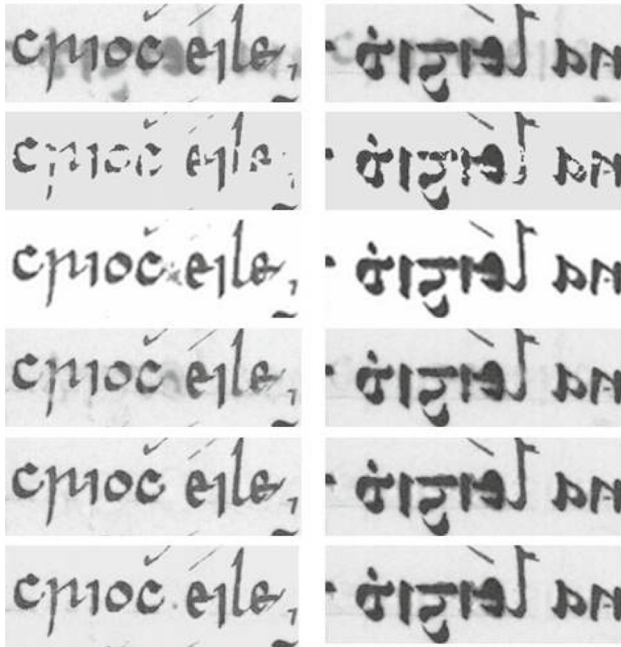
The original recto–verso pair and the corresponding reconstructions obtained with the method proposed in [6], which was the best performing for the previous experiment, and with our method are shown in Fig. 6. Again, our results are of very similar quality.

For a quantitative analysis, having available binary ground truth images, we binarized the reconstructions with the adaptive Sauvola algorithm [25] and then computed as quality indices the probability $FgError$ that a pixel in the foreground text was classified as background, the probability $BgError$ that a background or bleed-through pixel was classified as foreground, and the $WTotError$, that is, the weighted mean of $Fg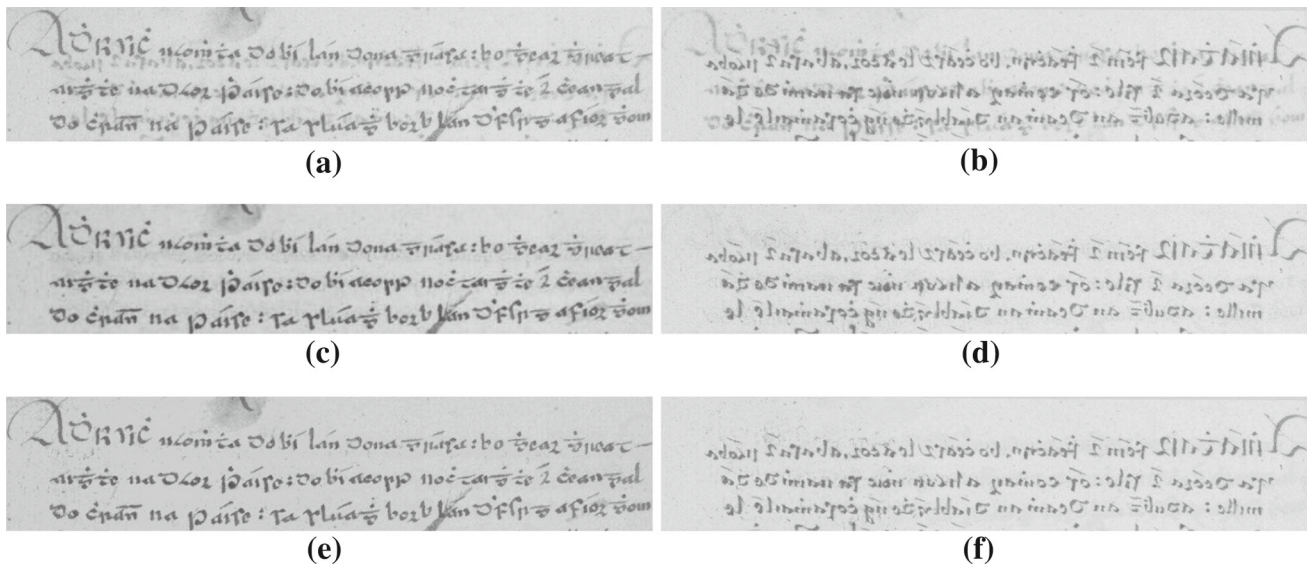Error$ and $BgError$, with the weights being the numbers of the foreground pixels and the background pixels as they result from the corresponding ground truth images, indicating the probability that any pixel in the image was misclassified. According to [24], these quality indices are defined as:

$$FgError = \frac{1}{N} \sum_{t \in GT(Fg)} |GT(t) - B(t)|$$

$$BgError = \frac{1}{N} \sum_{t \in GT(Bg)} |GT(t) - B(t)|$$

$$WTotError = \frac{N_{Fg} FgError + N_{Bg} BgError}{N} \tag{10}$$

where GT is the ground truth, B is the binarized restoration result, $GT(Fg)$ is the foreground region of the ground truth image constituted of N_{Fg} pixels, $GT(Bg)$ is the complementary background region of the ground truth image constituted of N_{Bg} pixels, and N is the total number of pixels in the

Table 1 Quality indices for the results in Figs. 3c, f, and 6e, f

Image	$FgError$	$BgError$	$WTotError$
Figure 3c	0.0097	0.0070	0.0074
Figure 3f	0.0071	0.0106	0.0100
Figure 6e	0.0151	0.0110	0.0035
Figure 6f	0.0143	0.0104	0.0108

image. The values of these indices for the two experiments above are summarized in Table 1.

The authors of [6] quantified the performance of their method and those of the methods in [5,18] and [10] by binarizing the reconstructions with Gatos binarization [26] and then computing the mean quality indices of Eq. (10) for the entire dataset. Their method was the best, with $FgError = 0.0696$, $BgError = 0.0085$ and $WTotError = 0.0196$. Though no comparison can be done on each single image, we observe that the quality indices for the two images shown above are below those mean values. We also computed the quality indices on the whole dataset. The measured errors are summarized in the following Table 2, where the mean errors are completed with the standard deviations and the best and worst quality indices. The mean errors reported in [6] are shown in the last column.

From Table 2, it appears that, on average, our results have lower FgError and WeightedTotError, and higher BgError, than those of [6]. From a check on some of our worst results, we observed that the wrong background pixels are usually confined at the character boundaries. This means that the

Fig. 5 Comparison among state-of-the-art methods on a detail of the images in Fig. 3a, d. From *top to bottom*: originals, results from the method in [5], results from the method in [18], results from the method in [10], results from the method in [6], and results from our method

Fig. 6 Application of the proposed method to a real recto–verso pair from dataset [23,24]: **a** original degraded recto, **b** original degraded verso, **c** recto restored with the method in [6], **d** verso restored with the method in [6], **e** recto restored with the proposed method, and **f** verso restored with the proposed method

Table 2 Quality indices for the entire dataset

	Mean	SD	Best	Worst	[6]
$FgError$	0.0176	0.011255	0.0049	0.0548	0.0696
$BgError$	0.0285	0.012949	0.0033	0.0669	0.0085
$WTotError$	0.0165	0.012115	0.0023	0.0542	0.0196
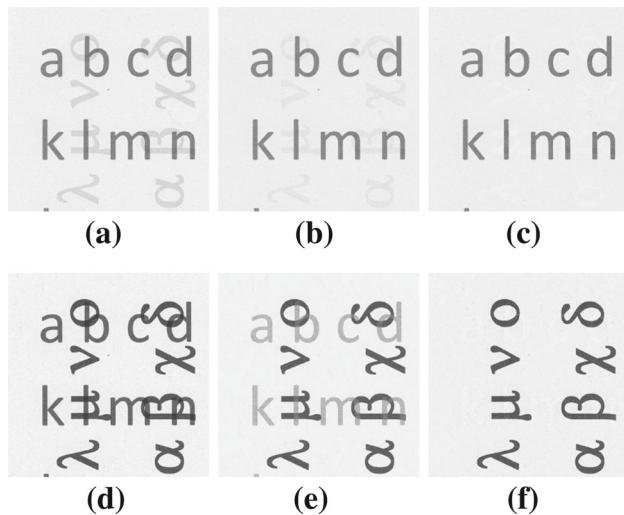
Fig. 7 Application of the proposed method to an RGB real-fake palimpsest: **a** green channel, **d** blue channel, **b** and **e** results of ICA applied to the density maps, **c** overwriting extracted with our method, and **f** underwriting extracted with our method

binarization algorithm underestimates the threshold between foreground and background. In general, we expect different values of the quality indices on the same image when using different binarization algorithms. Thus, for a fair quantitative comparison using this peculiar kind of ground truth images, the same binarization algorithm, with the same parameters, should be used. Finally, it is worth highlighting again the simplicity of our method, which leads to a very fast algorithm. Indeed, a non-optimized Matlab code takes 0.77 s. for restoring the $1,745 \times 1,070$ images of Fig. 3a, d, on an Intel core i7 3GHz CPU.

In the case of palimpsests, we show the potentialities of our model by presenting the results obtained on a real–fake palimpsest, i.e., a document created by printing a first vertical text in a light color and then overprinting it with a horizontal darker text. Scans of the documents have then been used for processing. With this choice of the colors, in the red channel, the underwriting (the vertical text) almost disappears, whereas it is slightly visible in the green channel and well visible in the blue channel. We chose as observation at wavelength λ_1 the blue channel, and as observation at wavelength λ_2 the green channel. Note that if we had chosen as second observation the red channel, we could have set $q^u(t) = 0, \forall t$, thus making the data system to reduce to a single equation.

Note also that none of the standard thresholding techniques applied to the blue channel alone could permit separation of the two texts. Figure 7a, d shows the green and blue channels of the RGB original scan. Figure 7b, e shows the results of applying ICA to the density data images, with the aim at showing that global coefficients for the linear model do not allow the two texts to be separated. The promising results of our method are shown in Fig. 7c, f.

5 Conclusions

We propose a non-stationary, linear mixing model of the optical densities to describe text overlapping in recto–verso images of archival documents, affected by see-through, and in multispectral images of palimpsests. Based on this model, we derive two simple, partially supervised algorithms to separate the two texts in the two different applications. These algorithms act in two phases: In the first one, the model parameters are estimated off-line from the data; in the second, the restored images are recovered by inverting the data model in a single step. Both the algorithms we propose are very fast, apart from the trivial need of selecting one or two small areas in one of the two observations. The experimental results on real recto–verso printed documents and manuscripts show that the parameters involved in the data model can be estimated with satisfactory accuracy, so as to obtain clean images of the individual texts. The method outperforms standard blind source separation techniques such as ICA and produces results comparable to those of the best performing among recent state-of-the-art methods. Also, the preliminary results obtained for the application to the recovery of the erased text in palimpsests are promising.

The two fundamental features of the data model adopted are non-stationarity, which permits to describe both penetration and transparency of the ink through the medium, as well as the variability of the degradation strength, and the inclusion of a PSF, which allows a pattern in a side to match the corresponding one in the opposite side.

Currently, we are working at the full automatization of the method and at its extension to the RGB case. This simply entails restoring the three pairs of recto–verso channels independently and then recomposing them to recover the RGB appearance of the cleansed document. Other issues regard relaxing the present assumption of lower density of the interferences with respect to that of the foreground text and finding better strategies for a more accurate selection of the model parameters in correspondence with the occlusion areas. These could range from the iterative refining of an initial, rough estimate, to the use of the average density of the foreground text computed in the same area used to estimate the PSF, and so on. Another possibility could be to include priors on the ideal images and estimate jointly all the parame-

ters. Even if we already pursued this way with some success in [20], more specialized priors should be sought, and the computational costs are still high.

References

1. Dubois, E., Pathak, A.: Reduction of bleed-through in scanned manuscript documents. In: Proceedings of IS&T Image Processing, Image Quality, Image Capture Systems Conference, pp. 177–180 (2001)

2. Tan, C.L., Cao, R., Shen, P.: Restoration of archival documents using a wavelet technique. IEEE Trans. PAMI **24**(10), 1399–1404 (2002)

3. Wang, Q., Tan, C.L.: Matching of double-sided document images to remove interference. In: Proceedings of IEEE CVPR 2001, p. 1084 (2001)

4. Hanasusanto, G.A., Wu, Z., Brown, M.S.: Ink-bleed reduction using functional minimization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 825–832 (2010)

5. Huang, Y., Brown, M.S., Xu, D.: User assisted ink-bleed reduction. IEEE Trans. Image Process. **19**(10), 2646–2658 (2010)

6. Rowley-Brooke, R., Piti, F., Kokaram, A.: A non-parametric framework for document bleed-through removal. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2954–2960 (2013)

7. Tonazzini, A., Bedini, L., Salerno, E.: Independent component analysis for document restoration. Int. J. Doc. Anal. Recognit. **7**, 17–27 (2004)

8. Tonazzini, A., Salerno, E., Bedini, L.: Fast correction of bleed-through distortion in grayscale documents by a blind source separation technique. Int. J. Doc. Anal. Recognit. **10**, 17–25 (2007)

9. Tonazzini, A., Gerace, I., Martinelli, F.: Multichannel blind separation and deconvolution of images for document analysis. IEEE Trans. Image Process. **19**(4), 912–925 (2010)

10. Rowley-Brooke, R., Kokaram, A.: Bleed-through removal in degraded documents. In: Proceedings of SPIE 8297 Document Recognition and Retrieval XIX, 82970T–10 (2012)

11. Merrikh-Bayat, F., Babaie-Zadeh, M., Jutten, C.: Using non-negative matrix factorization for removing show-through. In: Proceedings of LVA/ICA, pp. 482–489 (2010)

12. Sharma, G.: Show-through cancellation in scans of duplex printed documents. IEEE Tans. Image Process. **10**(5), 736–754 (2001)

13. Ophir, B., Malah, D.: Show-through cancellation in scanned images using blind source separation techniques. In: Proceedings of International Conference on Image Processing ICIP. Volume III. pp. 233–236 (2007)

14. Martinelli, F., Salerno, E., Gerace, I., Tonazzini, A.: Non-linear model and constrained ml for removing back-to-front interferences from recto-verso documents. Pattern Recognit. **45**, 596–605 (2012)

15. Salerno, E., Martinelli, F., Tonazzini, A.: Nonlinear model identification and seethrough cancellation from recto-verso data. Int. J. Doc. Anal. Recognit. **16**, 177–187 (2013)

16. Merrikh-Bayat, F., Babaie-Zadeh, M., Jutten, C.: Linear-quadratic blind source separating structure for removing show-through in scanned documents. IJDAR **14**, 319–333 (2011)

17. Almeida, M.S.C., Almeida, L.B.: Nonlinear separation of show-through image mixtures using a physical model trained with ica. Signal Process. **92**, 872884 (2012)

18. Moghaddam, R.F., Cheriet, M.: A variational approach to degraded document enhancement. IEEE Trans. Pattern Anal. Mach. Intell. **32**(8), 1347–1361 (2010)

19. Tonazzini, A., Salerno, E., Savino, P., Bedini, L.: Removal of non-stationary see-through interferences from recto-verso documents. In: Proceedings of International Workshop on Intelligent Pattern Recognition and Applications WIPRA 2013, pp. 151–158 (2013)

20. Gerace, I., Martinelli, F., Tonazzini, A.: Restoration of recto-verso archival documents through a regularized nonlinear model. In: Proceedings of Eusipco 2012, Bucharest, August 27–31 pp. 1588–1592 (2012)

21. Luenberger, D.G.: Linear and Nonlinear Programming. Addison-Wesley, Reading (1984)

22. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man. Cybern. **9**(1), 62–66 (1979)

23. Irish Script On Screen Project: http://www.isos.dias.ie (2012)

24. Rowley-Brooke, R., Piti, F., Kokaram, A.: A ground truth bleed-through document image database. In: Adn, G., Buchanan, P.Z., Rasmussen, E., Loizides, F. (eds.) Theory and Practice of Digital Libraries. Volume 7489 of Lecture Notes in Computer Science, pp. 185–196. Springer, Berlin (2012)

25. Sauvola, J., Pietikäinen, M.: Adaptive document image binarization. Pattern Recognit. **33**, 225236 (2000)

26. Gatos, B., Pratikakis, I., Perantonis, S.J.: Adaptive degraded document image binarization. Pattern Recognit. **39**(3), 317–327 (2006)