International Journal of Information Science & Technology – iJIST, ISSN : 2550-5114
Vol. 3 - No. 5 - September 2019

Virtual restoration and content analysis of ancient degraded manuscripts

Anna Tonazzini, Pasquale Savino, Emanuele Salerno, Muhammad Hanif, Franca Debole

***Abstract*—In recent years, extensive campaigns of digitization of the documental heritage conserved in libraries and archives have been performed, with the primary goal to ensure the preservation and fruition of this important part of the human cultural and historical patrimony. Besides protecting conservation, the availability of high quality digital copies has increasingly stimulated the use of image processing techniques, to perform a number of operations on documents and manuscripts, without harming the often precious and fragile originals. Among those, virtual restoration tasks are crucial, as they facilitate the traditional work of philologists and paleographers, and constitute a first step towards an automatic analysis of the written contents. Here we report our experience in this field, referring, as a case study, to the problem of removing one of the most frequent and impairing degradations affecting ancient manuscripts, i.e., the bleed-through distortion. We show that techniques of blind source separation are versatile tools to either cancel these unwanted interferences or isolate specific features for content analysis goals. Specialized algorithms, based on recto-verso models and sparse image representation, are then shown to be able to perform a fine and selective removal of the degradation, while preserving the original appearance of the manuscript.**

***Index Terms*—Ancient manuscript restoration; recto-verso registration; bleed-through removal; blind source separation; sparse representation inpainting**

I. Introduction

The virtual restoration of an ancient and degraded manuscript simulates, on the digital images, the process of physical and/or chemical restoration performed on the original, tangible manuscript. Degradations in the manuscripts are caused by several and diverse damages suffered during time for bad storage environment, careless usage, and the natural ageing and deterioration of the ink and the support. The aim of virtual restoration is twofold: i) to remove or attenuate interfering patterns, such as seeping inks, humidity spots and molds; ii) to enhance the primary text against faded or blurred ink, fragmented characters, and incomplete words.

The advantages of virtual restoration over physical restoration are evident, since a number of different and reversible techniques can be attempted without harming the original manuscript. As an example, in the case of the very common bleed-through degradation, physical restoration is impossible, since the chemical substances needed to remove the seeped ink would also destroy the ink of the foreground text. Again, the reconstruction of fragmented characters and words can take advantage of similarity search techniques using available dictionaries, both at the image and the textual level.

The authors are with the Institute of Information Science and Technologies, Italian National Research Council, Pisa, Italy. Email: anna.tonazzini@isti.cnr.it

On one side, virtual restoration may be the ultimate goal of manuscript digital processing. Indeed, it can serve for providing the scholars with a help to a better and easier reading of the text, during its manual transcription and the study of the manuscript origin, history and contents. In this sense, while removing interferences and enhancing the writing, it must maintain the original appearance of the manuscript intact as much as possible. Therefore, it is fundamental that virtual restoration preserves all marks and genuine features such as pencil annotations, stamps, paper watermarks and textures, miniatures, and so on.

On another side, virtual restoration can be the first preprocessing step toward the automatic analysis of the writings, needed for the automatic or user-assisted textual transcription, and/or for natural language processing purposes. In this sense, it must facilitate subsequent tasks such as layout analysis, text binarization, word spotting, OCR, and ICR.

A comprehensive survey of the most advanced technologies and computer science tools applied to the study of manuscripts can be found in [1].

In this paper, we consider the problem of bleed-through removal, which is one of the most urgent and challenging issues in the field of virtual restoration of ancient, degraded manuscripts. We compare two different strategies to cope with this kind of degradation, for the purpose of facilitating the manuscript analysis by both the scholars and the available automatic tools.

In the literature, bleed-through removal is mainly addressed as a classification problem, where the document image is subdivided into three components: background (the unwritten paper support), foreground (the main text), and bleed-through. The existing methods can be divided into two main categories: blind approaches [2], [3], [4], where the image of a single side is used, and the far more adopted non-blind approaches [5], [6], [7], [8], [9], [10], [11], [12], [13], where the information of both the recto and verso sides of the document is required.

Using both sides of the manuscript implies that an accurate matching between the information carried on by corresponding pixels should be ensured. In other words, the two pixel values must be the spectral signatures in the two sides of the same geometrical point. This means that the two images must be aligned. To digitally acquire ancient manuscripts, usually professional cameras are used, either high resolution CCD cameras or multispectral cameras, mounted on special mechanical equipment that guarantee a stable setup. Despite that, misalignments between the images of the two sides are likely to occur, due to the human intervention needed for turning around and repositioning the leaf. Thus, registration

algorithms must be used prior the restoration process.

Registration of recto-verso images is not an easy task, since the intensity of corresponding foreground and bleed-through areas are usually very different, bleed-through might only occur sparsely across the page, and the binding of the page in case of books may have different degrees of curvature in the two sides. The earliest recto-verso registration methods in the literature considered global affine transformations [14], [15], [16], [17], [18]. More recently, projective transformations have been proposed [19], as well as non-rigid registration methods [20], [21], [22], to cope with binding in the page.

Although many works address bleed-through removal for grayscale digital manuscripts, we tackle the problem for misaligned recto-verso manuscripts acquired as RGB digital images. As a matter of fact, after the extensive and high quality digitization campaigns recently carried out or ongoing in the majority of libraries and archives, it is very likely that the digital version of a manuscript affected by ink seeping from the reverse side comes out in at least the RGB modality. On one hand, diversity of acquisition increases information and, as such, should be exploited, when available. On the other hand, the color of a manuscript is an important cue, both for a more pleasant and authentic representation, and for the information that it can provide about the chemical nature of the substances (i.e. ink pigments), and the kind of degradation that the manuscript has undergone.

The paper is organized as follows. In Section II we describe the registration algorithm that we designed for this specific application. Section III is devoted to the description and the analysis of the results of a first virtual restoration technique, based on linear models of patterns overlapping in the two views, and blind source separation algorithms. In Section IV we discuss the results of a second virtual restoration technique based on non-stationary pattern overlapping models, and image inpainting via sparse representation. We also show that blind source separation, exploiting the spectral diversity of the restored images, can facilitate subsequent layout analysis. Section V concludes the paper.

II. Recto-verso registration

In this paper, we consider flat manuscripts (e.g. letters), so that we assume a global rigid deformation of the horizontally reflected verso with respect to the recto. However, we extend the transformation to be projective, since, besides translations and rotation caused by turning around the leaf, accidental movements of the camera may also cause scale changes and projective deformations.

If (x, y) and $(x + \Delta x, y + \Delta y)$ are two geometrically corresponding locations of the recto and verso sides of the physical manuscript, a projective transformation between the two locations, unique for the three pairs of homologous channels, has a form like the following one:

$$\begin{bmatrix} x+\Delta x \\ y+\Delta y \\ 1 \end{bmatrix} = \\ = \begin{bmatrix} s_x \cdot cos(\theta) + p_x b_x & -s_y \cdot sin(\theta) + p_y b_x & b_x \\ s_x \cdot sin(\theta) + p_x b_y & s_y \cdot cos(\theta) + p_y b_y & b_y \\ p_x & p_y & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \quad (1)$$

where homogeneous coordinates are used, b_x and b_y, s_x and s_y, p_x and p_y represent translations, scale factors and projective deformations, respectively, along the x and y axes, and θ is the angle of rotation. Thus, given the three RGB digital images of the verso, their versions registered on the homologous images of the recto will be built by applying to each pixel of an ordered grid the above transformation, and then computing its graylevel by interpolating the values of those pixels that, in the original verso, surround the possibly non-integer location. In this work, we adopt bicubic interpolation.

To determine the seven parameters of the transformation in eq. (1) by using least mean squares (LMSE), a minimum of four pairs of corresponding locations are needed. Of course, a larger number of pairs will ensure more accurate estimates, since possible small mismatches of the corresponding locations can be compensated.

In general, finding geometrically corresponding pixels in recto-verso images affected by bleed-through is not easy, since, as said, corresponding strokes may have very different intensities, or even be absent in one of the two sides. Thus, rather than searching for matching points among a set of singular points, such as corners or crosses, we consider as matching points the centers of small patches that correspond, i.e. depict the same physical scene, in the two sides. In order to find such centers, we apply the following procedure [23]. For a given patch in the recto side, we first select the homologous patch in the verso side, that is the patch having same size and same location. Owing to the global misalignment of the recto and verso images, these two patches will be misaligned too, but, if their size is sufficiently small, their mutual misalignment can be approximated by a translation only. Thus, to find the actual corresponding verso patch of the given recto patch, the amount of the translation must be estimated. To this purpose, we exploit the shift property of the Fourier Transform (FT), through the computation of the cross power spectrum of the two patches.

In the mathematical formalism, given two patches f_1 and f_2 such that, in a common support, it is $f_1(x + \Delta x_0, y + \Delta y_0) = f_2(x, y)$, the following relationship between their FTs holds true:

$$F_2(\omega_x, \omega_y) = F_1(\omega_x, \omega_y) e^{j(\omega_x \Delta x_0 + \omega_y \Delta y_0)} \quad (2)$$

from which:

$$\frac{F_2(\omega_x, \omega_y)}{F_1(\omega_x, \omega_y)} = e^{j(\omega_x \Delta x_0 + \omega_y \Delta y_0)} \quad (3)$$

Then, theoretically, the inverse FT of the ratio in eq. (3) will return a Dirac impulse located in $(\Delta x_0, \Delta y_0)$. Nevertheless, due to the presence of random noise, dissimilar parts and gain changes, inverse filtering produces a very noisy map, where a trustable peak cannot be located. A more robust estimate of the cross correlation between f_1 and f_2 is given by the inverse FT of the cross power spectrum:

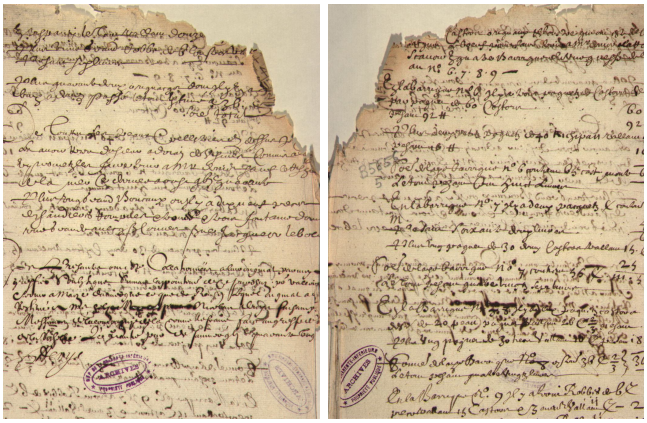
$$\frac{F_2(\omega_x, \omega_y) \cdot F_1^*(\omega_x, \omega_y)}{|F_2(\omega_x, \omega_y)| \cdot |F_1^*(\omega_x, \omega_y)|} = e^{j(\omega_x \Delta x_0 + \omega_y \Delta y_0)} \quad (4)$$

where $*$ denotes the complex conjugate. The location of the well emerging peak of the cross correlation function computed

using eq. (4) defines the relative translation between the two patches. Once this displacement is estimated, the center of the recto patch and the shifted center of the verso patch are taken as a pair of geometrically corresponding locations, whose coordinates will be used to estimate the coefficients of the projective matrix. Using FFT, this computation is very fast, allowing a quick detection of a very large number of relative translations and then corresponding points, which makes LMSE more robust [23].

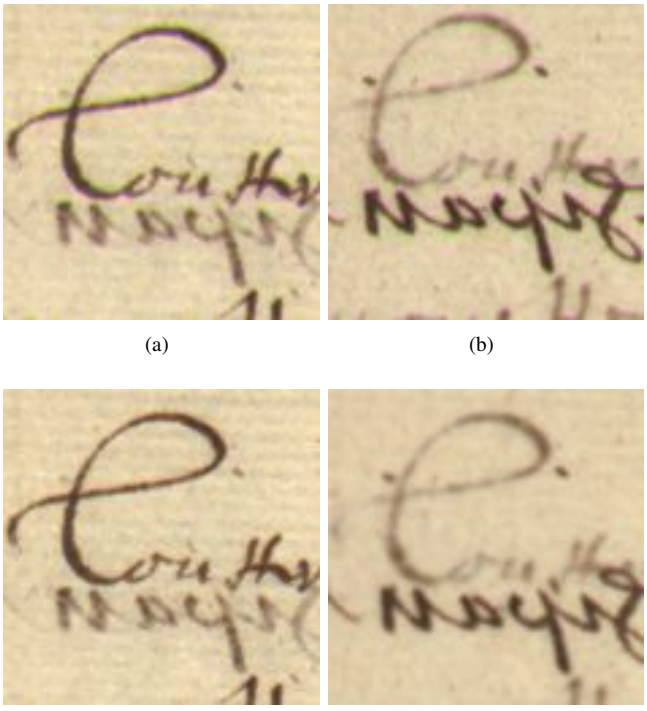
In practice, to detect the patches, a square window is simultaneously moved across the two sides so to span the whole image domain. The moving step can be smaller than the window size, to increase the number of detected corresponding points. The pairs of patches that are good candidates to be used for robustly estimating their relative displacements are those that both contain some text (either bleed-through in one side and foreground in the opposite side, or mixed bleed-through and foreground in both sides), and patches of pure background, i.e. with small values of standard deviation, are discarded. On average, with this criterion, we are able to locate a large number of patch pairs, equally distributed across the image. A uniform distribution of the patches in the image is fundamental to capture the different displacements caused by a projective transformation in the different areas of the image.

When a manuscript is acquired through a multispectral camera, it is likely that, besides the geometrical misalignment between recto and verso images, also the color channels of each single side appear misaligned, as shown in Fig. 2 (a), where a detail of the manuscript recto in Fig. 1 (a) is shown. Consequently, the misalignment of the mirrored color planes of the verso is "inverted" (see Fig. 2 (b)). To correct the misalignment of the color planes and geometrically register the two sides, we start correcting the color channel misalignment of the recto (Fig. 2 (c)). This is trivial if we assume that this misalignment is caused only by global displacements between pairs of channels. Hence, choosing a reference channel, the relative displacements of the other two channels can still be computed via the shift property of the FT, by using a single pair of patches. Subsequently, we separately estimate and apply as described before the individual projective transformations from each channel of the corrected recto and the homologous channel of the acquired verso. The final result is shown in Fig. 2 (d). It is evident that in the registered verso the correct RGB appearance has also been restored.



(a) (b)

Fig. 1. The manuscript used for our experiments: (a) recto; (b) verso.



(a) (b)

(c) (d)

Fig. 2. Illustration of the whole registration process on a small portion of the recto-verso pair in Fig. 1 : (a) original recto; (b) original verso after horizontal flipping; (c) recto after color plane alignment; (d) verso after geometrical registration on the recto.

III. Blind Source Separation

Any degraded manuscript, including those affected by bleed-through, appears as a mixture of layers of different texts and, possibly, of other information (paper texture and watermarking, stains, stamps, pencil annotations, etc.). Some of them are interferences (e.g. bleed-through and stains) and should be removed. Others, such as pencil annotations and stamps, may provide useful information to the scholars, and then should be preserved. In any case, the individual channels of the RGB acquisition, or possible multispectral/hyperspectral acquisitions, show the various layers, or some of them, still overlapped. This suggests that, at each pixel, such channels can be modeled as linear mixtures of the individual layers, weighted by coefficients that are related to the spectral responses of the corresponding object. If the different layers exhibit different spectral responses, that is, they are of different colors, the mixing matrix will be non-singular, and then invertible, if known. On the other hand, when viewed as spatial signals, while the channels are highly correlated, the layers are likely to be uncorrelated.

All the above observations lead to infer that the layers could somehow be separated from each other. Since the mixing matrix is not known, a layer uncorrelation constraint can be exploited to this purpose through statistical blind

source separation (BSS) techniques [24], [25]. Indeed, BSS techniques, such as principal component analysis (PCA) and independent component analysis (ICA), linearly combine the highly correlated spectral images to produce a different set of images that are uncorrelated and with decreasing variance, i.e. carrying on decreasing amounts of information. Furthermore, the output channels of ICA are statistically independent. Thus, they can actually produce images each showing one single layer separated from the others [26], [27]. Being the requirement of statistical independence of ICA a stronger condition than the assumption of uncorrelation of PCA, it may also happen that signals that are not well segmented by PCA may be separable by ICA, or by ICA applied on a set of principal components of highest eigenvalues, as done in [28].

The case considered here of manuscripts affected by the bleed-through degradation and acquired in the RGB modality is a particular instance of the data model and solution strategy described above. Below, this instance will be shortly formalized from a mathematical point of view, for application to bleed-through removal. We consider first the blind case, and then the non-blind approach.

Having said that a manuscript image affected by bleed-through can be modeled as the superposition of three different layers, in the BSS formalism this means that we have three different sources that combine someway to give the observed image. At the same time, we can assume to have three observed maps, obtained by splitting the mixture of the three sources into its red, green and blue components. Thus, we have the same number of sources and observations, which is the most classical assumption for BSS problems. Since we consider images of manuscripts containing texts, we can also reasonably assume that the color of each of the three sources is almost uniform, i.e we will have mean reflectance indices (r_1, g_1, b_1) for the background, (r_2, g_2, b_2) for the foreground and (r_3, g_3, b_3) for the bleed-through.

To explain the superposition of the three layers we developed an approximated linear mixture model assuming that, at each point of the manuscript, the three reflectance indices of the three layers mix linearly to form the total reflectance indices. In this model, the reflectance indices $(x_r(t), x_g(t), x_b(t))$ of a generic point t of the manuscript can be seen as given by the following equation:

$$\begin{bmatrix} x_r(t) \\ x_g(t) \\ x_b(t) \end{bmatrix} = \begin{bmatrix} r_1 & r_2 & r_3 \\ g_1 & g_2 & g_3 \\ b_1 & b_2 & b_3 \end{bmatrix} \begin{bmatrix} s_1(t) \\ s_2(t) \\ s_3(t) \end{bmatrix} \quad (5)$$

where functions $s_i(t)$, $i = 1, 2, 3$, indicate the "quantity" of background, foreground and bleed-through, respectively, that concur to form the color at point t. Viewing the $s_i(t)$ as underlying sources, and the unknown reflectance indices r_i, g_i, b_i, $i = 1, 2, 3$, as the entries of a so called mixing matrix A, eq. (5) has the typical form of a linear and instantaneous BSS problem, which has infinitely many solutions. It has been proved that, if an assumption of mutual independence of the sources can be made, a unique solution to this problem can be found through independent component analysis (ICA). In our case, this assumption is reasonable, as we already observed that the individual layers are usually at least uncorrelated. Among the several algorithms proposed in the literature to perform ICA, that is to estimate both matrix A and the functions $s_i(t)$, we experimented the FastICA algorithm [25] [27].

For a correct application of this method, a fundamental constraint is that matrix A should be nonsingular. This means that the source reflectance indices must be linearly independent, i.e. the sources must have different colors. Note that, when the mixing matrix is ill-conditioned or singular, extra observations taken in non-visible spectral bands, used in conjunction or in place of the visible channels, could be a remedy. Similarly, additional observations can be used in the cases where the manuscript contains other interfering patterns besides bleed-through, and there is an interest in removing them as well, or recovering their separate maps. On the other hand, the model in eq. (5) can be easily extended to more than three sources and channels.

Simpler methods that only try to decorrelate the observed data can also be attempted besides ICA. Among the many data decorrelation methods, the most popular are PCA and symmetric whitening (SW). As is known, the uncorrelation requirement is weaker than independence, and, in principle, no source separation can be obtained by only constraining second-order statistics, at least if no additional requirement is satisfied. However, since the individual layers that compose the image are at least less correlated than the data channels, decorrelating the color components gives a different representation where the now uncorrelated components of the image could coincide with the single layers. Furthermore, the second-order approach is always less expensive than ICA algorithms, and due to the poor modeling or to the lack of independence of the layers, the results from decorrelation can also be better than the ones from ICA [26].

When one of the methods described above yields a perfect separation of the layers, each layer would dominate the grayscale range in the related output channel, while pixels belonging to the other layers would exhibit the same gray value and thus merge with the background in that channel. More realistically, PCA or ICA may not succeed in separating the layers if their statistics are not truly orthogonal. For example, when applying ICA to the image of Figure 1(a), since foreground and bleed-through exhibit the same spectral behaviour, they cannot be fully separated, whereas the stamps can be extracted as isolated patterns (Figure 3). However, the front stamp and the verso, seeping stamp remain mixed, because they have the same color.

In another manuscript, shown in Figure 4(a), the difference in color between the foreground and bleed-through patterns, even though slight, allows their separation, and one of the ICA outputs represents the foreground text free of interferences, as shown in Figure 4(b). Note that the ICA outputs are graylevel images. The color appearance of the restored manuscript has been recovered by virtue of the coefficients of the underlying mixing matrix, that are estimated along with the sources. Indeed, as said, according to our model of eq. (5), each column of the mixing matrix represents the mean reflectance indices in the Red, Green, and Blue band, respectively, of the corresponding source.

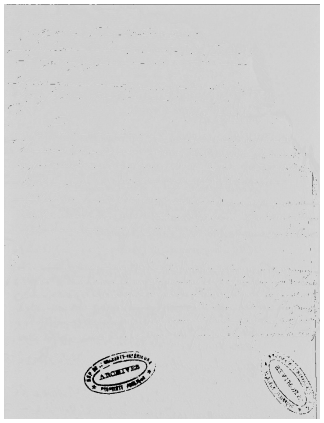


Fig. 3. The unique layer that ICA can isolate from the image in Figure 1(a).

From the RGB image alone, full separation of all the layers contained in the manuscript has not been obtained even in this case, since the stamp at the top of the page remain mixed with the foreground text. Nonetheless, from the point of view of virtual restoration, this makes the result excellent. Indeed, only the degradation (the bleed-through) has been removed, while the other informative features have been preserved. Again, the impossibility of achieving full separation of the overlapped patterns can be explained by the fact that some of them, foreground and stamp, reflect similarly in the explored bands, which, on the other hand, are just three against a higher number of patterns (background, foreground, bleed-through, and stamp). As already said, acquisitions in bands of the non-visible range of the spectrum may help, if the mixed patterns exhibit different spectral responses in those bands. For this manuscript, three extra acquisitions in the non-visible spectrum were available: two infrared maps and an ultraviolet map. A pseudocolor image of these extra bands is shown in figure 5 (a). From this representation, it is evident that the stamp, at these wavelengths, reflects differently from the foreground, so that their separation is possible. Indeed, by applying ICA to the three non-visible acquisitions, we obtain the extraction of the stamp, as shown in Figure 5 (b).

Coming back to the unlucky situation in which foreground and bleed-through reflect similarly in all the available spectral bands, some other form of diversity of information must be sought, through other acquisition modalities. A natural, different acquisition modality, normally available in the digital archives, is the acquisition of the verso side of the page. In recto-verso images of manuscripts affected by bleed-through, diversity of information is constituted by the different intensities of a same text pattern in the two observations. A linear, instantaneous overlapping model, similar to that devised for a single-side RGB acquisition, can also be adopted in this case. This is a 2×2 model, whose mixing matrix is related to the percentage of ink seeping from one side to the other. Since it is reasonable to assume that the ink penetrates the paper in the same way from recto to verso and from verso to recto, the mixing matrix can be assumed to be symmetric, so that separation can be equivalently achieved by ICA or the faster symmetric whitening [29].

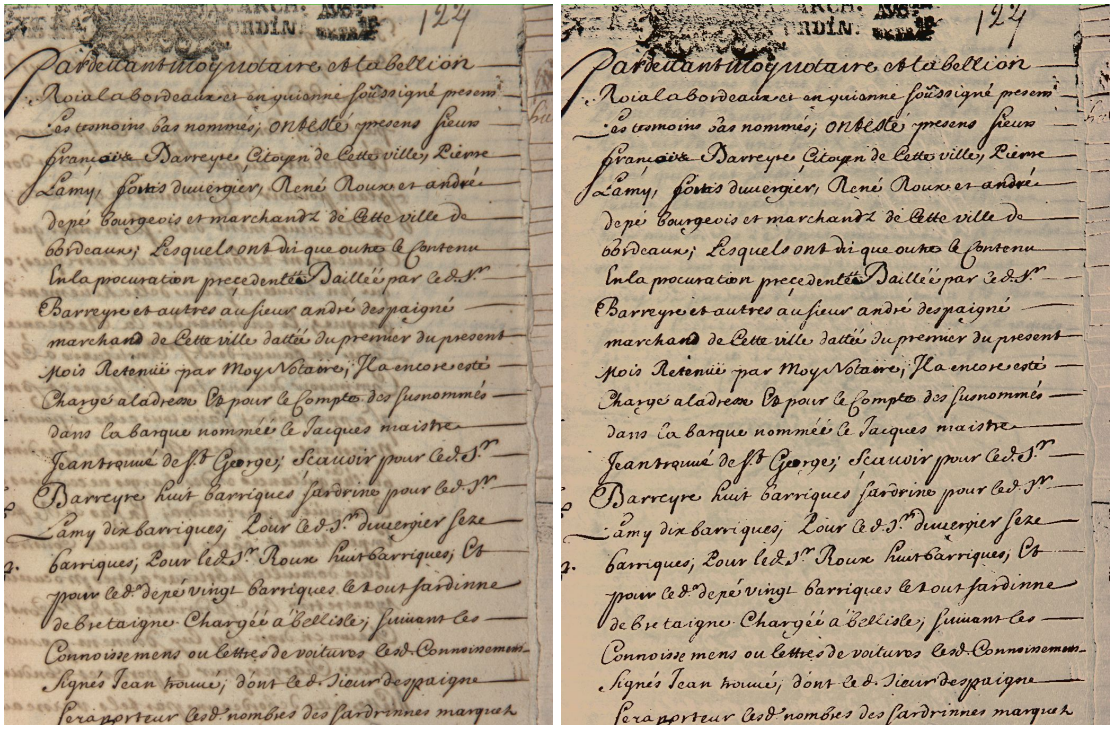
In the mathematical formalism, let $r(t)$ and $v(t)$, $t = 1, 2, ..., T$, be the grayscale images obtained by scanning the perfectly registered recto and verso pages of a manuscript, where t is a pixel index. We consider $r(t)$ and $v(t)$ as a linear combination of the two images $s_1(t)$ and $s_2(t)$, $t = 1, 2, ..., T$, representing the clean main texts in the recto and the verso, respectively. We can write:

$$\begin{aligned} r(t) &= A_{11}s_1(t) + A_{12}s_2(t) \\ v(t) &= A_{21}s_1(t) + A_{22}s_2(t) \end{aligned} \tag{6}$$

where A_{12}/A_{11} and A_{21}/A_{22} represent the intensity attenuations of the ink seeping from the verso to the recto and, respectively, from the recto to the verso. Such attenuations depend on the features of the transmission medium (paper, parchment, etc.) and on other factors, such as ink fading. Viewing s_1 and s_2 as the sources, and A as the unknown mixing matrix, eq. (6) turns out to be a 2×2 BSS problem. Again, this problem can be solved if the sources are mutually independent. In this hypothesis, both the sources and the mixing coefficients can be estimated from the data alone. In our case, however, some specific physical constraints allow us to relax the strict independence requirement, so that separation can be achieved by simply decorrelating the data images. We assume that the intensity contributions of the main texts are the same in the two pages, that is, $A_{11} = A_{22}$. We also assume that the attenuation of the bleed-through pattern in the two pages is the same, that is, $A_{12} = A_{21}$. Moreover, we expect that the contribution of the main text in each page is stronger than the contribution of the bleed-through, i.e. $A_{12} < A_{11}$ and $A_{21} < A_{22}$. In summary, we assume a symmetric and diagonal dominant mixing matrix. This is reasonable when dealing with printed pages or, in the case of manuscripts, when the two pages have been written at two close moments, with the same ink, by the same writer, and with the same pressure on the paper. Rather than taking this information into account explicitly, we exploit it to choose the most convenient among various BSS techniques available. Specifically, in [24], it is observed that decorrelating the data through symmetric whitening is equivalent to ICA when matrix A is symmetric. This is the property that we exploited in our experiments with recto-verso manuscripts, thus deriving a separation technique that, while performing similarly to ICA, is much faster and simpler.

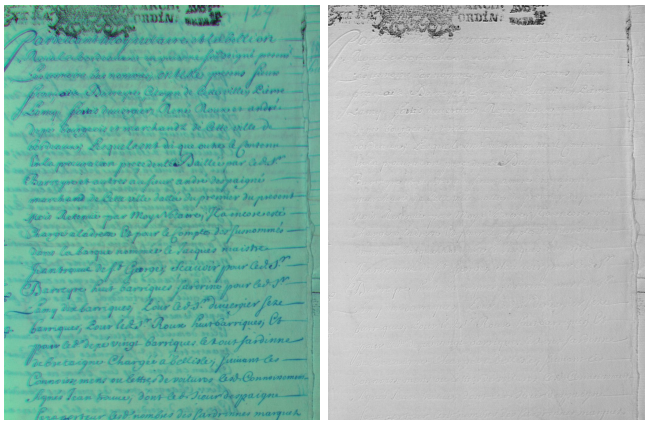
The superposition model of eq. (6) holds for each pair of recto-verso homologous channels, i.e. (R_r, R_v), (G_r, G_v) and (B_r, B_v), where the subscripts r and v stands for recto and verso, respectively, so that, once the three 2×2 systems have been inverted, the restored channels can be recomposed to give the restored RGB sides [17].

An example of the results of this procedure applied to the registered versions of the pair in Figure 1 is shown in Figure 6. Note how the bleed-through is reduced, especially in the verso side, and the color and other manuscript features are preserved. In particular, note how each side has maintained its own stamp only, and the pencil annotation in the verso side, highlighted by the red box, has not been removed.



(a) (b)

Fig. 4. Successful application of ICA for bleed-through removal: (a) RGB manuscript; (b) one output of ICA with color recovery.



(a) (b)

Fig. 5. Application of ICA to three acquisitions of manuscript of Figure 4 (a) in non-visible bands: (a) pseudocolor image encoding two infrared images plus one ultraviolet image; (b) one output of ICA.

Given the extremely high computational efficiency of such a procedure, this could be used as a routine in libraries and archives.

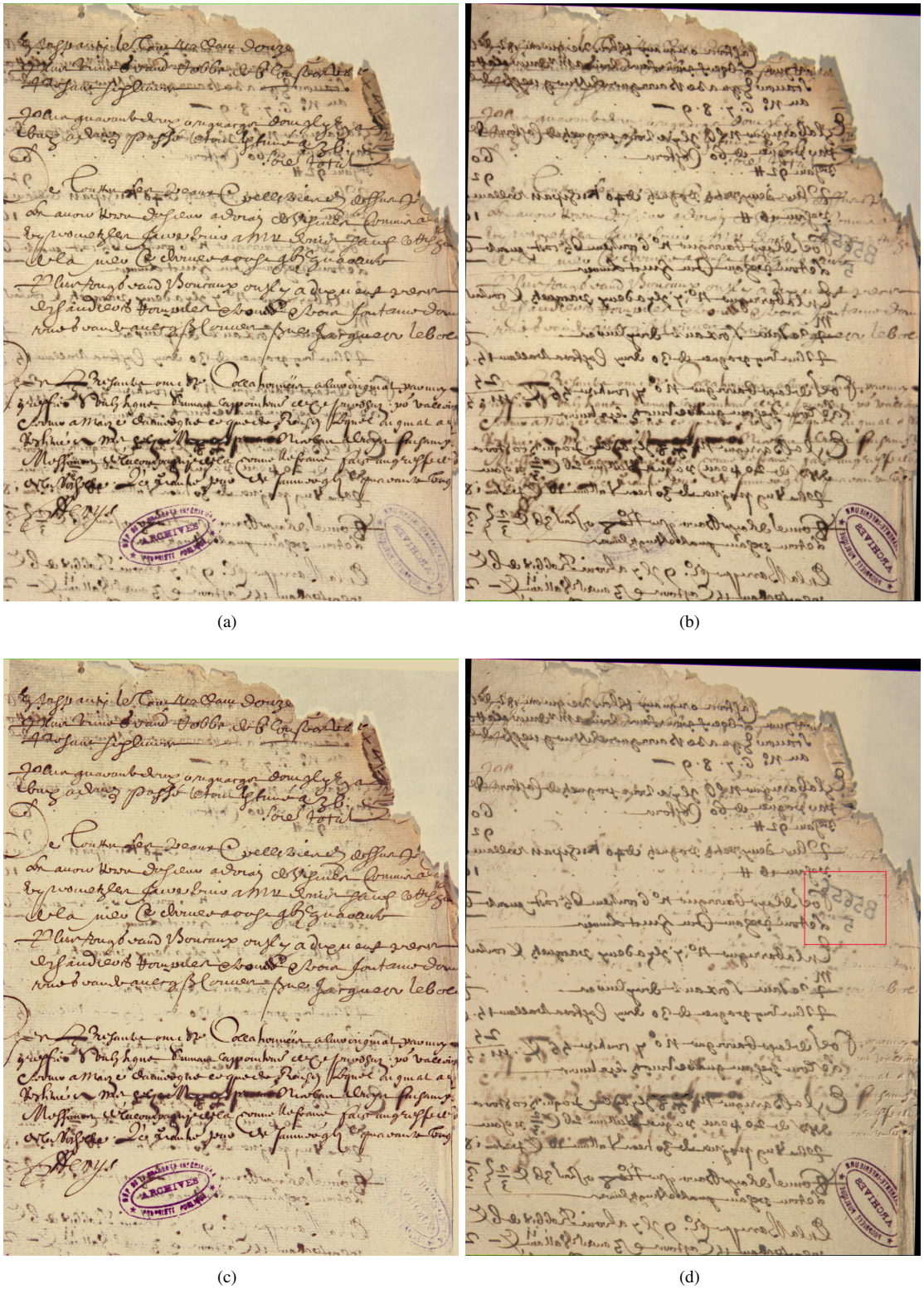
IV. Pixel-wise bleed-through identification and image inpainting

It is evident that the linear, instantaneous model in eq. (6) is a rough approximation of the phenomenon of interfering texts in recto-verso manuscripts, which derives from complicated processes of ink diffusion and paper absorption.

For instance, in the pixels where the two texts are superimposed to each other, the resulting intensity is not the sum of the intensities of the two components, but it is likely to be some nonlinear combination of them. For the phenomenon of show-through, a nonlinear model is derived in [30], but this must be linearized to have a tractable problem. Another simplification we adopted in (6) is to neglect the blur in the bleed-through pattern, due to the spreading of ink when passing through the paper fiber.

In [31], we proposed another bleed-through removal algorithm based on a recto-verso data model where the observed optical density of each side is given by the linear combination of the density of the undegraded side and an ink-smeared version of the ideal density of the opposite side. This linear combination is weighted by a pixel-dependent positive parameter, representing the degree of attenuation of the text that shows through, thus making the model non-stationary. Based on simple considerations, the attenuation levels are heuristically estimated from the data, by comparing the densities of corresponding pixels, i.e. the pixels referring to the same physical point in the two sides. The data model is then inverted in a single step, making the algorithm very fast.

The way in which we perform the comparison between pairs of corresponding pixels is motivated by some considerations about the physical phenomenon. Indeed, through experience, we observed that in the majority of the manuscripts examined, due to paper porosity, the seeped ink also diffused through the paper fiber. Hence, in general, the bleed-through pattern is a smeared and lighter version of the opposite text that has generated it. Note that this assumption does not mean that, in a same side, bleed-through is lighter than the foreground text. In fact, in each side the intensity of bleed-through is usually very variable, that is highly non-stationary, and sometimes can



(a) (b)

(c) (d)

Fig. 6. Application of symmetric whitening to the image pair in Figure 1 after their registration: (a) recto; (b) verso; (c) restored recto; (d) restored verso.

be as dark as the foreground text.

Other considerations can be made by reasoning in terms of "quantity" of ink. Indeed, the quantity of ink should be zero in the background, i.e. the unwritten paper, no matter of the color of the paper, maximum in the darker and sharper foreground text, and minimum in the lighter and smoother corresponding bleed-through text. For a measure of the quantity of ink with such properties, we use the concept of optical density, which is related to the intensity as follows:

$$d(t) = D(s(t)) = -log\left(\frac{s(t)}{b}\right) \tag{7}$$

where $s(t)$ is the image intensity at pixel t, and b represents the most frequent (or the average) intensity value of the background area in the image.

Based on the assumptions above, we adopt a linear, non-stationary model in the optical densities, to describe the superposition between background, foreground and bleed-through in the two observed recto and verso images:

$$\begin{aligned} d_r^{obs}(t) &= d_r(t) + q_v(t)D(h_v(t) \otimes s_v(t)) \\ d_v^{obs}(t) &= d_v(t) + q_r(t)D(h_r(t) \otimes s_r(t)) \end{aligned} \tag{8}$$

for each pixel t. In eqs. (8), d^{obs} is the observed optical density, and d is the ideal optical density of the clean image, with the subscripts r and v indicating the recto and verso side, respectively. D is the operator that, applied to intensity, returns the optical density according to eq. (7), and $\otimes$ indicates convolution between the ideal intensity s and a unit volume Point Spread Function (PSF), h, describing the smearing of ink penetrating the paper. At present, we assume stationary PSFs, empirically chosen as Gaussian functions, although a more reliable model for the phenomenon of the ink spreading should consider non-stationary operators. Finally, the space-variant quantities q_r and q_v have the physical meaning of attenuation levels of the density (or ink penetration percentage), from one side to the other.

The proposed algorithm identifies the bleed-through pixels in each side as those whose optical density is lower than that of the corresponding pixels in the opposite side, i.e., of the foreground that has generated the bleed-through. Thus, on the basis of eqs. (8), at each pixel we first compute the following ratios:

$$\begin{aligned} q_r(t) &= \frac{d_v^{obs}(t)}{D(h_r(t) \otimes s_r^{obs}(t)) + \epsilon} \\ q_v(t) &= \frac{d_r^{obs}(t)}{D(h_v(t) \otimes s_v^{obs}(t)) + \epsilon} \end{aligned} \tag{9}$$

Then, for all pixels, we maintain the smallest between the two computed attenuation levels, and set to zero the other. This allows us to correctly discriminate the two instances of foreground in one side and bleed-through in the other, so that, all pixels where $q_r > 0$ are classified as bleed-through in the verso side, whereas those where $q_v > 0$ are classified as bleed-through in the recto side.

However, it is evident that with the criterion above we can obtain wrong positive attenuation levels, in one of the two sides, in correspondence of some background pixels and some occlusion pixels, i.e. where the two foreground texts superimpose to each other. This happens because in the cases background–background and foreground–foreground the two densities are almost the same, around zero in the first case and around the maximum density in the other, with small oscillations that make the value of the ratio unpredictable. To correct this possible overestimation of the bleed-through pixels, we set to zero the attenuation level when the densities d_r^{obs} and d_v^{obs} are both low (or high, respectively) and close to each other.

In case of RGB images, this model, as the previous, instantaneous one, holds independently for each pair of the three channels, and, as before, the algorithm can be separately applied to each pair of recto-verso color channels to obtain the restored RGB sides.

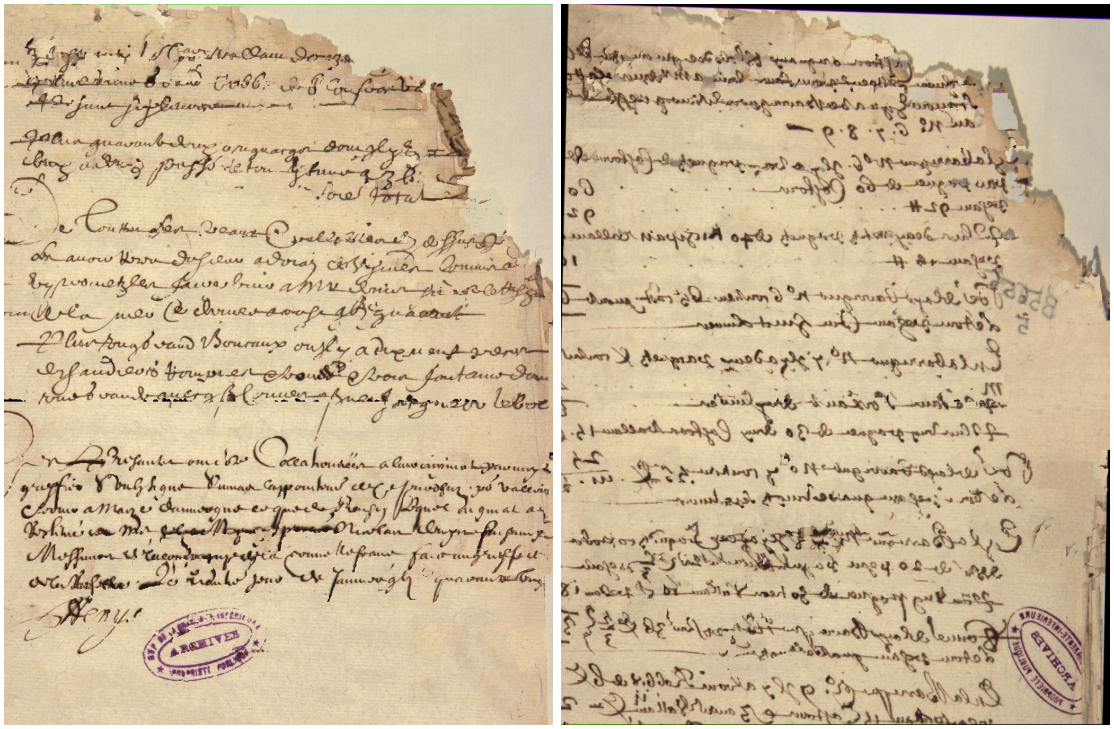
Also this algorithm is able to remove only the unwanted interferences while preserving other patterns, such as stamps or pencil annotations, that are peculiar of each side, as well as the original colors of foreground and background.

Nevertheless, as the optical density is defined with respect to a unique, average value of the background, in the practice the algorithm substitutes the identified bleed-through pixels with values around the used average background value. In cases where the background is textured and non-uniform, this makes some unpleasant imprints of the bleed-through pattern to be still visible. Thus, to obtain a virtual restored image which is faithful to the original as much as possible, we recently proposed an improvement to this algorithm. In this improved version, the identified bleed-through pattern is inpainted in continuity with its surrounding, by using techniques of image inpainting based on sparse representation [32], [33].

In these techniques, a sparse representation dictionary is learned from the set of the overlapping patches in the manuscript that do not contain bleed-through pixels. Then, the estimated bleed-through map is locally filled in with a combination of atoms of the dictionary, whose coefficients give the best sparse representation of the non-bleed-through surrounding area. More specifically, in each patch of the image, the bleed-through pixel values are estimated using the sparse coefficients that describe the uncorrupted pixels of the patch itself and of similar neighbouring patches. This assists in the estimation of the missing pixels and guarantees a smooth transition by exploiting the local similarity typical of natural images. In this way, the bleed-through pattern is filled in with the background texture closest to it, and the occlusion text areas wrongly identified as bleed-through are brought back to the foreground texture observable in their surrounding.

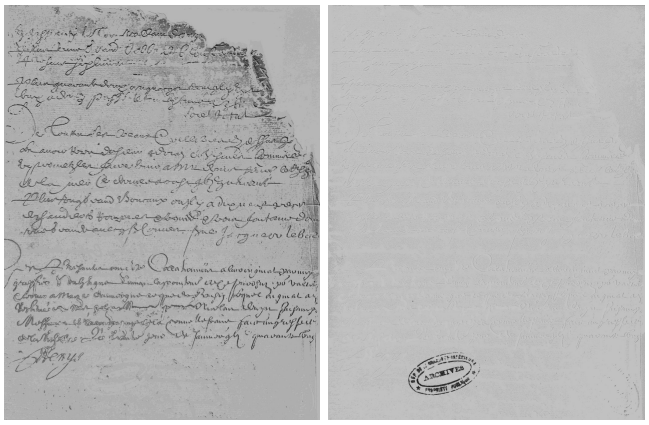
Figure 7 shows the results of this second algorithm, to be compared with those of the BSS based algorithm in Figure 6. The much better removal of bleed-through by this algorithm can be clearly appreciated. Observe that the restored image of Figure 7(a) contains now only two layers of information, the foreground text and the stamp, which exhibit different spectral behaviours. To this image we can then apply BSS tecniques, e.g. ICA, to attempt to separate the two layers. The excellent result is shown in Figure 8. This experiment shows how preliminary virtual restoration of a degraded manuscript can facilitate the task of document layout analysis.

Figure 9(b) shows the result of the binarization of the restored recto with the Sauvola algorithm [34], to be compared



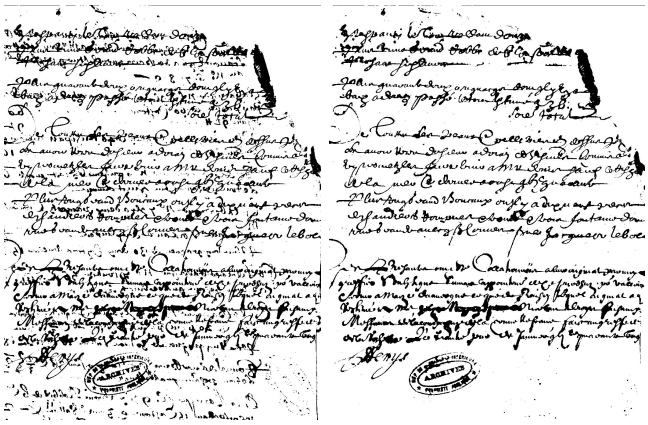
(a) (b)

Fig. 7. Application of pixel-wise bleed-through identification and sparse image inpainting: (a) restored recto; (b) restored verso.



(a) (b)

Fig. 8. Application of ICA to the restored recto of Fig. 7(a) : (a) extracted foreground text; (b) extracted stamp.



(a) (b)

Fig. 9. Application of the Sauvola binarization before and after virtual restoration: (a) binarized original recto; (b) binarization of the recto restored with the second proposed method.

with Figure 9(a), showing the binarization with the same algorithm of the original degraded recto.

In [31] we provided qualitative and quantitative evaluation of the bleed-through removal method described above, with reference to the popular dataset of recto-verso manuscripts available at website https://www.isos.dias.ie/. In [33], with reference to the same dataset, comparisons with the methods in [13] and [4] have been provided as well.

V. Conclusions

We described techniques for the virtual restoration of ancient manuscripts degraded by the bleed-through distortion. We considered a manuscript image as the superposition of different layers of information, which can be disentangled by exploiting their spectral diversity in available multimodal observations, such as multispectral (e.g. RGB) acquisitions and recto-verso acquisitions. In this way, the interfering bleed-through pattern can be removed, leaving the useful informative layers unaltered. The algorithms we developed within such an approach provide restored manuscripts that maintain their original appearance, and satisfy two major requirements: i) to allow easier reading and interpretation by scholars; ii) to facilitate subsequent tasks of automatic analysis. We also

described a simple and effective algorithm for the preliminary registration of the available multimodal acquisitions. All the proposed algorithms are very fast and suitable to be used routinely in libraries and archives.

Our future research plans in this domain regard the possibility of exploiting the non-visible acquisition bands often available (e.g. Infrared and Ultraviolet) as sources of additional information to be used in place of the verso side.

ACKNOWLEDGMENT

This work has been partially supported by the European Research Consortium for Informatics and Mathematics (ERCIM), within the Alain Bensoussan Fellowship Programme.

REFERENCES

[1] C. Brockmann, M. Friedrich, O. Hahn, B. Neumann, and I. Rabin, Eds., *Natural Sciences and Technology in Manuscript Studies*, ser. Manuscript Cultures. Hamburg: University of Hamburg, 2014, vol. 7.

[2] D. Fadoua, F. L. Bourgeois, and H. Emptoz, "Restoring ink bleed-through degraded document images using a recursive unsupervised classification technique," *Document Analysis Systems VII, Lecture Notes in Computer Science, vol. 3872. Springer*, pp. 27–38, 2006.

[3] C. Wolf, "Document ink bleed-through removal with two hidden markov random fields and a single observation field," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 431–447, 2010.

[4] B. Sun, S. Li, X. P. Zhang, and J. Sun, "Blind bleed-through removal for scanned historical document image with conditional random fields," *IEEE Trans. Image Process.*, pp. 5702–5712, 2016.

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

[6] G. A. Hanasusanto, Z. Wu, and M. S. Brown, "Ink-bleed reduction using functional minimization," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2010, pp. 825–832.

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

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

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

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

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

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

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

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

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

[16] J. Wang, M. S. Brown, and C. L. Tan, "Accurate alignment of double-sided manuscripts for bleed-through removal," in *Proc. 8-th IAPR Workshop on Document Analysis Systems*, 2008, pp. 69–75.

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

[18] V. Rabeux, N. Journet, and J. P. Domenger, "Document recto-verso registration using a dynamic time warping algorithm," in *Proc. Int. Conf. on Document Analysis and Recognition (ICDAR)*, 2011, pp. 1230–1234.

[19] B. Li, W. Wang, and H. Ye, "Multi-sensor image registration based on algebraic projective invariants," *Optics express*, vol. 21, pp. 9824–9838, 2013.

[20] A. Myronenko and S. Xubo, "Intensity-based image registration by minimizing residual complexity," *IEEE Transactions on Medical Imaging*, vol. 29, p. 18821891, 2010.

[21] J. Wang and C. L. Tan, "Non-rigid registration and restoration of double-sided historical manuscripts," in *Proc. Int. Conf. on Document Analysis and Recognition (ICDAR)*, 2011, p. 13741378.

[22] R. Rowley-Brooke, F. Piti, and A. Kokaram, "Nonrigid recto-verso registration using page outline structure and content preserving warps," in *Proc. 2nd International Workshop on Historical Document Imaging and Processing, HIP 2013*, 2013, p. 813.

[23] P. Savino and A. Tonazzini, "Digital restoration of ancient color manuscripts from geometrically misaligned recto-verso pairs," *Journal of Cultural Heritage*, vol. 19, pp. 511–521, 2016.

[24] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing*. New York: Wiley, 2002.

[25] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: Wiley, 2001.

[26] A. Tonazzini, E. Salerno, M. Mochi, and L. Bedini, "Blind source separation techniques for detecting hidden texts and textures in document images," in *Proc. International Conference on Image Analysis and Recognition ICIAR 2004*, 2004, pp. 241–248.

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

[28] E. Salerno, A. Tonazzini, and L. Bedini, "Digital image analysis to enhance underwritten text in the archimedes palimpsest," *Int. Journal on Document Analysis and Recognition*, vol. 9, pp. 79–87, April 2007.

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

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

[31] A. Tonazzini, P. Savino, and E. Salerno, "A non-stationary density model to separate overlapped texts in degraded documents," *Signal, Image and Video Processing*, vol. 9, pp. 155–164, 2015.

[32] T. Ogawa and M. Haseyama, "Image inpainting based on sparse representations with a perceptual metric," *EURASIP J. Adv. Signal Process.*, vol. 179, pp. 1200–1212, 2013.

[33] M. Hanif, A. Tonazzini, P. Savino, and E. Salerno, "Non-local sparse image inpaintig for document bleed-through removal," *Journal of Imaging*, vol. 4, p. 68, 2018.

[34] J. Sauvola and M. Pietikäinen, "Adaptive document image binarization," *Pattern Recognition*, vol. 33, p. 225236, 2000.