# Registration and enhancement of double-sided degraded manuscripts acquired in multispectral modality [*]

Anna Tonazzini Gianfranco Bianco Emanuele Salerno
Istituto di Scienza e Tecnologie dell'Informazione
Consiglio Nazionale delle Ricerche
Via G.Moruzzi, 1 - 56124 Pisa, Italy
firstname.surname@isti.cnr.it

## Abstract

*We propose a system to process multispectral scans of double-sided documents. It can co-register any number of recto and verso channel maps, and reduce the bleed-through/show-through distortions by exploiting blind source separation. From RGB scans, it is also able to recover the original colors, thus improving the readability of a document while maintaining its original appearance. The recto and verso patterns obtained can then be further analyzed. Many approaches to this problem are based on single-channel or multichannel recto-verso scans. In any case, getting rid of the unwanted interferences is a challenging problem. All the methods relying on pixel intensities, such as the one presented here, need a very accurate co-registration, and this is difficult for recto-verso pairs since the relevant information is often very sparse.*

## 1 Introduction

Image processing techniques to improve human or automatic readability of ancient documents are being increasingly employed in the management of libraries and archives. Indeed, this makes easier the access to our documental patrimony by the public and scholars, without altering or further damaging the often precious and fragile originals. Among the most common degradations affecting ancient documents there are the bleed-through and the show-through effects. Bleed-through occurs when some patterns interfere with the main text due to seeping of ink from the opposite side of a page. Show-through is an interference from the opposite side due to the transparency of the paper, and may also appear in modern, well-preserved documents.

Removing these interferences is not trivial, since their intensity is often close to the one of the main text. On the basis of a single-side scan, the techniques attempting to reduce the interference are based on thresholding or segmentation-classification. From a color scan, thresholding can only be used effectively within multiresolution analysis with adaptive binarization, or using directional wavelets [11] [18]. Other proposals include color cluster segmentation via an adaptation of the $k$-means algorithm [9] [5], and a single-scan grayscale classification based on a double binary Markov random field [19]. While these methods certainly perform better than simple thresholding, they do not always succeed in discriminating unambiguously the foreground text from interferences. When show-through or bleed-through are very strong, exploiting both sides of the documents can be more effective than any kind of thresholding [14] [6]. These techniques can rely on segmentation and inpainting [4] [8] [17], adaptive compensation based on a nonlinear data model [14], or nonlinear diffusion [3]. All these approaches, however, pose two kinds of problems. Firstly, all the data maps must be spatially registered before processing. This is not trivial, since the different color channels have often different focus conditions, and, above all, recto and verso normally suffer from topographical differences. Secondly, since the solutions mentioned above remove all the structured background from the foreground text, besides canceling the strokes coming from the reverse side, they can also remove other patterns belonging to the front side (e.g. stamps). This may be undesirable, when these patterns are signs of the document history and authenticity.

A specific research line focuses on applying blind source separation (BSS) algorithms, viewing the foreground and the interference as individual patterns that overlap in the document, and relying on multiple observations for their separation. In [15], we proposed a linear, instantaneous mixture data model, and an independent com-

[*] This work has been supported by the Project "Restarts" (http://www.restarts.org), under Program POR Calabria 2000-2006 - Azione 3.16.d "Ricerca e Sviluppo Tecnologico nelle Imprese Regionali".

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.131

ponent analysis strategy to analyze multispectral single-side scans with several overlapping information layers. In [16], we extended this approach to the reduction of the show-through/bleed-through interference in registered recto-verso grayscale scans. Along this line, in [12] and [10], a convolutive BSS formulation has been proposed, also accounting for a known non-linearity, as done in [14]. Thanks to technological advances, multispectral and hyper-spectral imaging is becoming a tool for the analysis and the documentation of old manuscripts whose text is degraded, overwritten, or partially masked. Archiving additional views along with the standard grayscale or RGB scans has an incalculable value to ensure the document to be preserved for the future and to enrich its documentation, often adding historically relevant information. Any further low-level processing, however, must be based on accurately co-registered sets of images, and this poses registration problems that are much more complicated than the ones arising for grayscale scans. In multichannel processing, human intervention should be reduced as much as possible. Ideally, accurate and fully automatic registration tools are required.

In this paper, we propose a procedure for the fully unsupervised registration of any number of multispectral views of double-sided document pages affected by bleed-through or show-through, and extend the approach described in [16] to separate the recto and verso patterns. Our aim is twofold: to produce enhanced versions of all the available scans at the different wavelengths, and a restored visible document that, while cleansed of the unwanted interferences, maintains its useful features as much as possible. An advantage of color or multispectral imaging over grayscale imaging is that, once two separate recto and verso images are obtained, each of them can further be analyzed, e.g., by the methods proposed in [15]. The paper is organized as follows. In Sections 2 and 3, the registration algorithm and the blind restoration technique are presented, respectively. In Section 4, a few experimental results on real manuscripts are shown and discussed. Section 5 concludes the paper and suggests some future prospects.

## 2 Registration of multispectral scans of recto-verso pairs

Image registration is a critical step in appearance-based analysis of multi-image data. This is because appearance-based processing associates the relevant information to image intensities. Thus, when the data come from misaligned images, each pixel must be precisely related to a reference location: one of the data images must be taken as a reference, and all the others must be transformed in order to create a data cube where all the information associated to one pixel is also associated to a fixed spatial location. With recto-verso pairs, all the images from one of the sides must be flipped horizontally before registration. Registering recto-verso pairs is a challenging task, since recto and verso are different not only for image intensity, but also topographically. Moreover, a correspondence between pixels can only be identified in the often small and sparse regions where a pattern present in one side also appears in the opposite side. A complete set of registration parameters can thus be very hard to compute. Because of differences in focusing and rigid translations of some of the channel maps, significant displacements can also be found in different views from the same side. Additional misalignments may be due to accidental causes occurred during the acquisition. Thus, in the case of multispectral recto-verso pairs, many images need to be aligned. This can be a serious hurdle to user-assisted registration, since it may become very cumbersome and prone to subjective errors.

Among the possible automatic registration methods [20], we discarded the feature-based ones, since we found that they are often insufficient to register recto-verso pairs. Rather, we found good results with the area-based methods using the Fourier-Mellin transform and parameter optimization [13] [6] [1]. Applying one of the selected methods, each image to be registered is given a set of transformation coefficients. At present, we are only admitting affine transformations. A final version of our procedure should at least consider, and correct, the geometric distortions due to folding or small sensor misalignments between successive acquisitions. Once all the transformations are computed, the images must be cropped to let them represent the same area and have the same size. As a quality measure for registration of a pair of images, we adopted the normalized mean-square error on the correlation coefficient of the transformed pair, defined as follows:

$$E^2 = 1 - \frac{\max_{u,v} |r_{fg}(u,v)|^2}{\sum_{x,y} |f(x,y)|^2 \cdot \sum_{x,y} |g(x,y)|^2} \qquad (1)$$

where, $f(x,y)$ and $g(x,y)$ are the registered images, and $r_{fg}(u,v)$ is their cross-correlation function. For any fixed pair of images, this quality index can be used to evaluate the relative merits of different registration methods. The Fourier-Mellin and the parameter optimization methods have shown perfectly comparable performances in registering recto-verso pairs. The results are normally suitable to apply the analysis techniques described in the next section. As examples, the errors evaluated to register the images shown in Figure 1 have been 0.2939 for Fourier-Mellin and 0.3095 for parameter optimization. For the images in Figure 2, we found $E^2 = 0.1782$ for Fourier-Mellin and $E^2 = 0.1487$ for parameter optimization.

# 3 Removing interference from registered data

Let us assume that the procedure described in the previous section has provided us with a set of $N$ registered recto-verso pairs. This assumption includes the grayscale case, with $N = 1$, and any color or multispectral data set. In [16], we have shown how in the grayscale case the interfering patterns can be reduced by a second-order blind separation approach. This strategy was motivated by observing that the appearances of the recto and verso sides are normally more correlated than the pure recto and verso patterns. Thus, if we orthogonalize the two data images, the interference of each side in the other is likely to be reduced. Here, we introduce a generalization of that approach, whose output can further be processed and analyzed by low-level approaches. In fact, once the two pure recto and verso patterns have been separated, we still have two independent data cubes. First of all, if these contain the red, green and blue components, then the visible colors of the original document can be reconstructed. Second, if one of or both the output data cubes can be thought of as superpositions of uncorrelated patterns, these can still be separated by, e.g., the techniques we proposed in [15]. Both these possibilities will be exemplified in the next section.

To build a model for our data, we only consider two distinct patterns, one in the recto-side and one in the verso-side of our page. Thus, we admit that the appearances of the two sides are given, for each observation channel, by $2 \times 2$ linear mixtures of these patterns:

$$\begin{aligned} r^k(t) &= A^k_{11} s^k_1(t) + A^k_{12} s^k_2(t), \quad t = 1, 2, \ldots, T \\ v^k(t) &= A^k_{21} s^k_1(t) + A^k_{22} s^k_2(t), \quad k = 1, 2, \ldots, N \end{aligned} \tag{2}$$

where $r^k(t)$ and $v^k(t)$ are the recto and verso appearances, respectively, at the $k$-th channel, $s^k_1(t)$ and $s^k_2(t)$ are the reflectance maps at the $k$-th channel (or $sources$), associated to the recto and verso patterns, respectively, $t$ is a pixel index, and $A^k_{ij}$ are unknown mixing coefficients. Physically, $A^k_{12}/A^k_{11}$ and $A^k_{21}/A^k_{22}$ represent the verso-to-recto and recto-to-verso attenuations, respectively. These depend on the properties of the transmission medium and other factors, such as ink fading. By grouping all the equations in (2), we get a $2N \times 2N$ block-diagonal linear system with unknown matrix:

$$\begin{bmatrix} r^1(t) \\ v^1(t) \\ \ldots \\ r^N(t) \\ v^N(t) \end{bmatrix} = \begin{bmatrix} A^1 & 0 & \ldots & 0 \\ 0 & A^2 & \ldots & \\ \ldots & \ldots & \ldots & \\ 0 & & & A^N \end{bmatrix} \cdot \begin{bmatrix} s^1_1(t) \\ s^1_2(t) \\ \ldots \\ s^N_1(t) \\ s^N_2(t) \end{bmatrix} \tag{3}$$

where the structure of the diagonal blocks is clear from Eq. (2). If we denote the data vector by $\mathbf{x}(t)$, the source vector by $\mathbf{s}(t)$, and the system matrix (or $mixing\ matrix$) by $\mathbf{A}$, then Eq. (3) can be written in vector form:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) \tag{4}$$

The fact that the mixing matrix must be estimated along with the sources motivates the adjective $blind$ used for this kind of separation problems.

Reasonable properties of the sources are that, for any possible pair of channels $(k, h)$, the recto and verso components $s^k_1(t)$ and $s^h_2(t)$ are almost uncorrelated, whereas the recto sources $s^k_1(t)$ and $s^h_1(t)$ are strongly correlated since they are just different frequency components of the same reflectance function. The same is true for any pair of verso sources, $s^k_2(t)$ and $s^h_2(t)$. In other words, the source covariance matrix has the form

$$R_{\mathbf{s}} = \begin{bmatrix} I & \Sigma_{12} & \Sigma_{13} & \ldots & \Sigma_{1N} \\ \Sigma_{12} & I & \Sigma_{23} & \ldots & \\ \Sigma_{13} & \Sigma_{23} & I & \ldots & \\ & & & \ldots & \\ & & & & I \end{bmatrix} \tag{5}$$

Having assumed, without loss of generality, that all the sources have unit variance, matrix $I$ is the $2 \times 2$ identity, and

$$\Sigma_{kh} = \begin{bmatrix} \sigma^{kh}_1 & 0 \\ 0 & \sigma^{kh}_2 \end{bmatrix} \tag{6}$$

$\sigma^{kh}_1$ and $\sigma^{kh}_2$ being the covariances of the $k$-th and $h$-th channels of the recto and verso sources, respectively. The idea behind solving system (3) by decorrelating the recto and verso appearances boils down to find a linear transformation $W_s$ on $\mathbf{x}$ such that the covariance matrix of the transformed vector has the same structure as matrix (5). With $N = 1$, the problem becomes to diagonalize the $2 \times 2$ covariance matrix $R_{\mathbf{x}}$ of the recto and verso grayscale appearances. In a general $2N \times 2N$ case, however, trying to diagonalize $R_{\mathbf{x}}$ would be a mistake, since, as is clear from Eqs. (5) and (6), this would force zero correlations between strongly correlated pairs. One way to solve this problem is to observe, from (3), that the mixing model is separable, and thus it can be solved channel by channel, as done in the grayscale case, by diagonalizing separately the $N$ channel covariance matrices of size $2 \times 2$, estimated as

$$R^k_{\mathbf{x}} = < (\mathbf{x}^k)(\mathbf{x}^k)^* > \approx \frac{1}{T} \sum_{t=1}^{T} [\mathbf{x}^k(t)][\mathbf{x}^k(t)]^* \tag{7}$$

where the superscript $k$ denotes the channel and the asterisk means transposition. To choose among all the possible diagonalizing matrices, we assume that all the diagonal blocks of matrix $\mathbf{A}$ are symmetric [16]. Indeed, if, as is reasonable, we assume $A^k_{11} = A^k_{22}$ and $A^k_{12}/A^k_{11} = A^k_{21}/A^k_{22}$, then we will also have $A^k_{12} = A^k_{21}$ for each $k$. In this case [2], applying a symmetric diagonalizing matrix can be proved to be

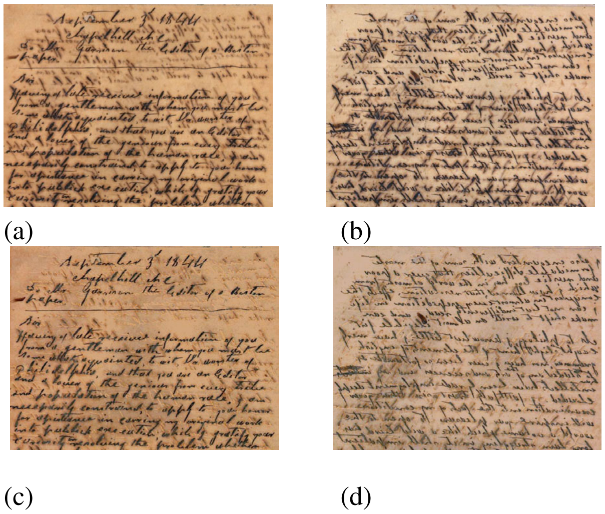
(a) (b) (c) (d)

**Figure 1. Restoration of a real recto-verso RGB pair: (a) recto; (b) flipped verso; (c) restored recto; (d) flipped restored verso.**

equivalent do perform an independent component analysis separation. Such a matrix can be derived from the following formula:

$$W_s^k = (V_{\mathbf{x}}^k)(\Lambda_{\mathbf{x}}^k)^{-\frac{1}{2}}(V_{\mathbf{x}}^k)^* \qquad (8)$$

where $V_{\mathbf{x}}^k$ is the matrix of the eigenvectors of $R_{\mathbf{x}}^k$ and $\Lambda_{\mathbf{x}}^k$ is the related eigenvalue matrix. By premultiplying all the available recto-verso pairs by matrices $W_s^k$, we get a new set of images, whose covariance matrix has the same diagonal blocks as matrix $R_{\mathbf{s}}$. As shown in the next section, the results obtained by this approach are often satisfactory, and the output images are less affected by interferences than the original scans. By doing so, however, we completely disregard the values in the offdiagonal blocks (6), which are left completely free, whereas all the crosscovariances of recto and verso images should be zero. At present, we are trying to take this information into account by also relying on the structure of the mixing matrix. Ideally, the values of the cross-covariances $\sigma_1^{kh}$ and $\sigma_2^{kh}$ should also be estimated.

As mentioned, when the pages are captured by RGB sensors, the original document colors can then be reconstructed by recomposing the color channels. This fact has been exploited to produce restored visible documents that, while cleansed of the unwanted interferences, maintain their original appearances as much as possible.

## 4 Experimental results

In this section, we show some examples from our extensive experimentation on real degraded documents. Figures 1 and 2 show the RGB reconstructions of two RGB recto-verso scans affected by a strong bleed-through. In both cases, a significant attenuation of the bleed-through has been obtained, and the original color has been pretty well recovered. As suggested in Section 3, the RGB recto and verso images thus obtained can further be analyzed to extract possible extra uncorrelated patterns. In Figure 3, we show the result we obtained through symmetric decorrelation applied to the RGB image in Fig. 2 (c). As is apparent, the stamp pattern and the main manuscript text have been separated very well. Of course, the procedure exemplified here follows a low-level strategy, only based on the pixel spectral signatures. We are just suggesting that the tehniques we proposed in [15] can be useful to process the restored recto and verso data cubes, but it is apparent that any other procedure can be useful, including the ones based on geometric, syntactic or semantic features. Some of the patterns present in the document could thus be enhanced or highlighted through pattern recognition approaches. Going further in this investigation, however, is out of our present scope.

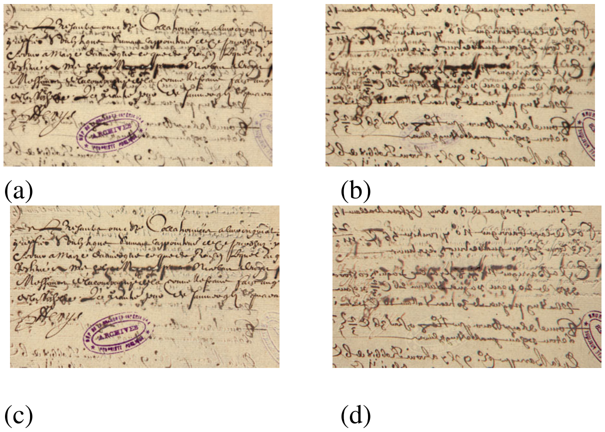
(a) (b) (c) (d)

**Figure 2. Restoration of a real RGB recto-verso pair: (a) recto; (b) flipped verso; (c) restored recto; (d) flipped restored verso.**

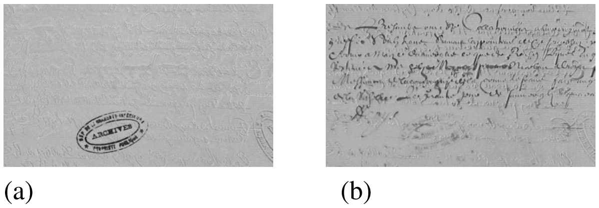
(a) (b)

**Figure 3. Symmetric orthogonalization of the restored RGB recto in Figure 2 (c): (a) the stamp; (b) the cleansed main text.**

# 5 Concluding remarks

We propose a procedure for the fully unsupervised restoration of double-sided documents affected by bleed-through or show-through distortions, and acquired in color or multispectral modality. This is based on the preliminary registration of all the available views, followed by a statistical analysis of the registered scans. The overall procedure could constitute a fast, reliable and effective system to be routinely used in libraries and archives for the enhancement of multispectral scans of degraded documents. The method is flexible for use in various contexts of document analysis, such as the extraction of hidden or masked patterns.

For the near future, we are planning to complete our investigation about the possible exploitation of the overall structure assumed for the source covariance matrix, and the estimation of the diagonal elements in submatrices (6). The basic problem with this kind of approaches to interference reduction is the adequacy of the data models adopted. Indeed, bleed-through and show-through are very complex phenomena and simple linear instantaneous models are not sufficient to account for all the relevant features of the degraded documents. Nevertheless, our simple approach proves to be often useful to mitigate the interferences. Before trying to introduce nonlinearities in the data model, we are now developing a linear convolutional model that should be able to account for ink spreading in bleed-through and for the blurring of the show-through pattern due to light diffusion effects within the paper.

Another problem to be addressed is to find a procedure to evaluate the quality of the results. In our experience, finding an absolute measure to evaluate the algorithm performance heavily depends on the final goals and, consequently, on the quantities chosen for evaluation. For document virtual restoration, of course, a perfect removal of bleed-through is the ultimate goal. Wherever this cannot be reached, however, a measure of effectiveness can be defined on the basis of the improvement achievable in some subsequent task. For example, in [16] we measured the effectiveness of our grayscale-based bleed-through reduction technique through the improvements induced in a subsequent OCR. Especially for heavy bleed-through, the OCR error rate was highly reduced, even in the presence of a significant residual interference. We are now planning to extend that study to our multichannel results, but this is by no means an easy task. To just mention one of the difficulties, the results will depend on the OCR software chosen, which, in turn, must be reasonably effective on all the test documents. Thus, a sensible evaluation criterion should be defined on the cascaded acquisition-registration-enhancement-OCR procedures rather than on interference suppression alone.

# References

[1] G. Bianco, A. Tonazzini, and E. Salerno "Assessing automatic registration methods applied to digital analysis of historical documents", in *Abstracts SIMAI 2008*, Rome, Italy, 15-19 September 2008, p. 24.

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

[3] M. Cheriet, and R. F. Moghaddam, "Degradation modeling and enhancement of low quality documents", in *Proc. WOSPA 2008*, 2008.

[4] P. Dano "Joint Restoration and Compression of Document Images with Bleed-through Distortion". Master Thesis, Ottawa-Carleton Institute for Electrical and Computer Engineering, School of Information Technology and Engineering, University of Ottawa , June 2003.

[5] F. Drida, F. Le Bourgeois, and H. Emptoz, "Restoring ink bleed-through degraded document images using a recursive unsupervised classification technique", in *Proc. 7th Workshop on Document Analysis Systems*, 2006, pp. 38–49.

[6] 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.

[7] Google, Book Search Dataset, Version v edition, 2007.

[8] K. Knox, "Show-Through Correction for Two-Sided Documents." United States Patent 5,832,137, Nov. 1998.

[9] Y. Leydier, F. Le Bourgeois and H. Emptoz "Serialized unsupervised classifier for adaptive color image segmentation: application to digitized ancient manuscripts", in *Proc. Int. Conf. on Pattern Recognition*, 2004, pp. 494–497.

[10] F. Merrikh-Bayat, M. Babaie-Zadeh, and C. Jutten, "A nonlinear blind source separation solution for removing the show-through effect in the scanned documents", in *Proc. EUSIPCO 2008*.

[11] H. Nishida, and T. Suzuki, "A Multiscale Approach to Restoring Scanned Color Document Images with Show-Through Effects", in *Proc. ICDAR 2003*.

[12] 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.

[13] B. S. Reddy, and B. N. Chatterji, "An FFT-based technique for translation, rotation and scale-invariant image registration", *IEEE Trans. on Image Processing*, Vol. 5, No. 8, pp. 1266–1271, August 1996.

[14] G. Sharma, "Show-through cancellation in scans of duplex printed documents", *IEEE Trans. on Image Processing*, Vol. 10, pp. 736–754, 2001.

[15] A. Tonazzini, L. Bedini, and E. Salerno, "Independent Component Analysis for document restoration", *IJDAR*, Vol. 7, pp. 17–27, 2004.

[16] A. Tonazzini, E. Salerno, and L. Bedini, "Fast correction of bleed-through distortion in grayscale documents by a Blind Source Separation technique", *IJDAR*, Vol. 10, pp. 17–25, June 2007.

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

[18] Q. Wang, T. Xia, L. Li, and C. L. Tan, "Document Image Enhancement Using Directional Wavelet", in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, Vol. 2, pp. 534–539, 2003.

[19] C. Wolf, "Document ink bleed-through removal with two hidden Markov random fields and a single observation field", Laboratoire d'Informatique en Images et Systémes d'Information, INSA de Lyon, France, Tech. Rep. RR-LIRIS-2006-019, November 2006.

[20] B. Zitová, and J. Flusser, "Image registration methods: A survey", *Image and Vision Computing*, Vol. 21, pp. 997–1000, 2003.