International Journal of Document Analysis (2006)
DOI 10.1007/s10032-006-0015-z

ORIGINAL PAPER

Anna Tonazzini · Emanuele Salerno · Luigi Bedini

Fast correction of bleed-through distortion in grayscale documents by a blind source separation technique

Received: 13 October 2004 / Revised 16 March 2005 / Accepted: 7 January 2006

Abstract Ancient documents are usually degraded by the presence of strong background artifacts. These are often caused by the so-called bleed-through effect, a pattern that interferes with the main text due to seeping of ink from the reverse side. A similar effect, called show-through and due to the nonperfect opacity of the paper, may appear in scans of even modern, well-preserved documents. These degradations must be removed to improve human or automatic readability. For this purpose, when a color scan of the document is available, we have shown that a simplified linear pattern overlapping model allows us to use very fast blind source separation techniques. This approach, however, cannot be applied to grayscale scans. This is a serious limitation, since many collections in our libraries and archives are now only available as grayscale scans or microfilms. We propose here a new model for bleed-through in grayscale document images, based on the availability of the recto and verso pages, and show that blind source separation can be successfully applied in this case too. Some experiments with real-ancient documents are presented and described.

Keywords Grayscale document restoration · Bleed-through cancellation · Blind source separation · Independent component analysis

1 Introduction

Improving document readability is a common need in libraries and archives. Since the original documents should not be altered physically, readability problems are often treated by applying image-processing techniques on digital document scans. Another need whose importance is rapidly increasing is to include digital documents into searchable databases. For this task, machine-readable versions of the original texts are required. When dealing with very large collections, speed is an issue, and the machine-readable documents should be generated by a minimum of human intervention. This is normally done by optical character recognition systems (OCR), whose performance, however, depends on the quality of the data. Since ancient documents are often severely degraded, image-processing techniques can also be helpful as a preprocessing step before applying OCR. When degraded originals are to be managed, different classes of digital image restoration algorithms can thus become essential to both improve human readability and obtain acceptable OCR performances.

Bleed-through and show-through are very common degradations. Bleed-through occurs when seeping of ink from the back page (normally denoted as *verso*) produces a pattern that interferes with the main text in the front page (normally denoted as *recto*). The show-through distortion consists of a similar interfering pattern due to transparency of paper, and may also appear in modern, well-preserved documents. Removing the bleed-through or show-through patterns from a digital document scan is not trivial, especially from ancient originals, where interferences of this kind are usually very strong. Indeed, dealing with strong bleed-through is practically impossible by any simple thresholding technique, since the intensities of the unwanted background can be very close to the ones of the main text. Thus, adaptive and/or structural approaches have to be adopted. For the general problem of textured background removal, the authors in [1] suggest the investigation of multistage thresholding techniques. In [2], segmentation and grouping techniques, based on the Gestalt cognitive rules, are used to eliminate interfering strokes from skeletonized versions of handwritten documents. In [3], local adaptive filters are applied for homogeneous and textured background removal from handwritten grayscale documents. Other work done on the specific problem of bleed-through/show-through removal has mainly exploited information from both the recto and verso pages [4–6]. Besides requiring a preliminary registration of the two sides, these techniques are usually expensive, as they are based on steps of segmentation, to identify the bleed-through areas,

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

followed by inpainting of estimated pure background areas [7]. In [8, 9], a color scan from a single side is required, but a thresholding technique can only be used in the framework of multiresolution analysis and adaptive binarization.

Recently (e.g. [10]), we proposed to model a document image as the linear combination of the interfering texts, and to separate them by processing multiple "views" of the mixed object. When a color scan of the document is available, three different views can be obtained from the red, green, and blue image channels. Even more views can be available in the cases where the document scans are obtained from narrowband hyperspectral sensors. Since the mixture coefficients are generally not known, the separation of the different patterns can be classified as a blind source separation (BSS) problem. We attempted to solve this problem through independent component analysis (ICA) [11] or other statistical techniques. In practice, this can be viewed as an adaptive representation of the document in a new color space, where the transformed color maps are mutually independent or, at least, uncorrelated. Although our image model is simplified [4], it has already proved to give interesting results in removing bleed-through [12], and extracting partially hidden features, such as paper watermarks [13] and underwritten texts in palimpsests [14].

Unfortunately, color or hyperspectral scans are not always available. Large repositories of digitized documents or microfilms already exist in many archives and libraries, where the images have been captured in grayscale only. For these images, a BSS strategy such as the one described earlier cannot be applied. However, we will show here that the grayscale recto and verso sides of a document affected by bleed-through can still be modeled as a linear superposition of the main texts in the two pages, so that BSS can still be attempted. Furthermore, although the mixing coefficients are unknown, they can reasonably be supposed to give rise to a symmetric mixing matrix, so that the ICA approach is equivalent to a simpler and faster decorrelation of the observed data [15].

The paper is organized as follows. In Sect. 2, we introduce our linear data model. In Sect. 3, we recall the properties of different decorrelation matrices, and, in Sect. 4, we present some experimental results with real printed or manuscript documents. Some final remarks are given in Sect. 5.

2 Formulation of the problem

Let $r(t)$ and $v(t)$, $t = 1, 2, \ldots, T$, be the grayscale images obtained by scanning the recto and verso pages of a document, respectively, where t is a pixel index. Also, suppose that recto and verso have been spatially registered, after a horizontal flip of, say, the verso. We consider $r(t)$ and $v(t)$ as a linear combination of the two images $s_1(t)$ and $s_2(t)$, $t = 1, 2, \ldots, T$, representing the clean main texts in the recto and the verso, respectively. We can write:

$$\begin{aligned} r(t) &= A_{11}s_1(t) + A_{12}s_2(t) \\ v(t) &= A_{21}s_1(t) + A_{22}s_2(t) \end{aligned} \tag{1}$$

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. In general, coefficients A_{ij} are not known. Functions r and v are the data, while s_1 and s_2 are the source patterns to be separated, i.e. estimated from the data with no knowledge of the mixing coefficients. This is normally called a blind source separation (BSS) problem. It is known that one condition under which this type of problems can be solved is that the source functions are mutually independent. In this hypothesis, and if some additional assumptions are verified, 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 orthogonalizing 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 will exploit it to choose the most convenient among various separation techniques available.

The model of Eq. (1) is a particular instance of the more general model we proposed in [10] for the multispectral scans of a single document page. In that case, vectors $\mathbf{x}(t)$ were assumed to have N components, and vectors $\mathbf{s}(t)$ were assumed to have M components. Since we considered document images containing homogeneous texts or drawings, we also assumed that the color of each source is almost uniform. In that case, A_{ij} represented the mean emissivity of the ith source at the jth wavelength channel. The data are thus a collection of T samples from a random N-vector $\mathbf{x}$, which is generated by linearly and instantaneously mixing the components of a random M-vector $\mathbf{s}$ through an $N \times M$ *mixing matrix* A

$$\mathbf{x}(t) = A\mathbf{s}(t) \quad t = 1, 2, \ldots, T \tag{2}$$

In the present case, our model of Eq. (1) is exactly the same of Eq. (2), restricted to the case $M = N = 2$, and where the mixing matrix coefficients are related to ink attenuation indices rather than emissivities, and $\mathbf{x} = (r, v)$.

It is easy to see that this model does not perfectly account for the phenomenon of interfering texts in documents, which derives from complicated processes of ink diffusion and paper absorption. Just to mention one aspect, in the pixels where 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 example, for the phenomenon of show-through, a nonlinear model is derived in [4], but this must be linearized to have a tractable problem. Another simplification we adopt is to neglect both noise and blur. As already said, although the linear model is just a rough approximation of the reality, it has demonstrated to be useful in different applications.

3 The proposed solutions: ICA, PCA and whitening

When no additional assumption is made, problem of Eqs. (1) or (2) is clearly underdetermined, since any nonsingular choice for A can give an estimate of $\mathbf{s}(t)$ that accounts for the evidence $\mathbf{x}(t)$. Even if no specific information is available, statistical assumptions can often be made on the sources. In particular, it can be assumed that the sources are mutually independent. If this assumption is justified, both A and $\mathbf{s}$ can be estimated from $\mathbf{x}$. As mentioned in the section "Introduction," this is the ICA approach [11]. Mutual source independence can be enforced by assuming a factorized form for the joint prior density of $\mathbf{s}$

$$P(\mathbf{s}(t)) = \prod_{i=1}^{N} P_i(s_i(t)) \qquad \forall t \tag{3}$$

The separation problem can be formulated as the maximization of the density in Eq. (3), subject to the constraint $\mathbf{x} = A\mathbf{s}$. This is equivalent to search for a matrix $W = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N)'$ such that, when applied to the data $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$, produces the set of vectors $\mathbf{w}_i'\mathbf{x}$ that are maximally independent, and whose distributions are given by the P_i. By taking the logarithm of Eq. (3), the problem solved by ICA algorithms is then

$$\hat{W} = \arg\max_{W} \sum_{t} \sum_{i} \log P_i(\mathbf{w}_i'\mathbf{x}(t)) + T \log|\det(W)| \tag{4}$$

Matrix $\hat{W}$ is an estimate of A^{-1}, up to arbitrary scale factors and permutations on the columns. Hence, each vector $\hat{\mathbf{s}}_i = \hat{\mathbf{w}}_i'\mathbf{x}$ is one of the original source vectors up to a scale factor.

To make separation possible, a necessary condition besides independence is that all the sources, but at most one, are nongaussian. To enforce nongaussianity, generic super-gaussian or subgaussian distributions can be used as source priors. These have proved to give very good estimates for the mixing matrix and for the sources as well, no matter of the true source distributions, which, on the other hand, are usually unknown [16].

Even if our data model were correct, since there is no apparent physical reason why our sources should be independent, no ICA-based algorithm is assured to achieve separation. However, it is intuitively clear that one can try to maximize the information content in each component of the data vector by decorrelating the observed image channels. In fact, it is directly observable that, while the recto and verso pages are usually highly correlated in presence of bleed-through, the individual main text patterns are, at least, less correlated. Decorrelating the two views thus gives them a new representation, which could make them coincide with the individual source patterns.

To avoid cumbersome notation, and without loss of generality, let us assume to have zero-mean data vectors. We should find a linear transformation $\mathbf{y}(t) = W\mathbf{x}(t)$ such that $\langle y_i y_j \rangle = 0, \forall i, j = 1, \ldots, M, \quad i \neq j$, where W is an $M \times N$ matrix and the notation $\langle \cdot \rangle$ means expectation. In other words, the components of the transformed data vector $\mathbf{y}$ should be orthogonal. It is clear that this operation is not unique, since, given an orthonormal basis of a subspace, any rigid rotation of it still yields an orthonormal basis of the same subspace. It is well known that linear data processing can help to restore color text images. In [17], the authors compare the effect of many fixed linear color transformations on the performance of a recursive segmentation algorithm. They argue that the linear transformation that obtains maximum-variance components is the most effective. They thus derive a fixed transformation that, for a large class of images, approximates the Karhunen–Loeve transformation, which is known to produce maximum-variance orthogonal vectors. This approach is also called principal component analysis (PCA), and one of its purposes is to find the most *useful* among a number of variables [15]. In our case, we can estimate the 2×2 data covariance matrix as

$$R_{\mathbf{xx}} = \langle \mathbf{x}\mathbf{x}^{\mathrm{T}} \rangle \approx \frac{1}{T} \sum_{t=1}^{T} \mathbf{x}(t)\mathbf{x}^{\mathrm{T}}(t) \tag{5}$$

Since the data are normally correlated, matrix $R_{\mathbf{xx}}$ will be nondiagonal. The covariance matrix of vector $\mathbf{y}$ is

$$R_{\mathbf{yy}} = \langle W\mathbf{x}\mathbf{x}^{\mathrm{T}}W^{\mathrm{T}} \rangle = W R_{\mathbf{xx}} W^{\mathrm{T}} \tag{6}$$

To obtain a vector $\mathbf{y}$ whose components are mutually orthogonal, $R_{\mathbf{yy}}$ should be diagonal. Let us perform the eigenvalue decomposition of matrix $R_{\mathbf{xx}}$, and call $V_{\mathbf{x}}$ the matrix of the eigenvectors of $R_{\mathbf{xx}}$, and $\Lambda_{\mathbf{x}}$ the diagonal matrix of its eigenvalues, in increasing order. Now, it is easy to verify that all of the following choices for W yield a diagonal $R_{\mathbf{yy}}$:

$$W_{\mathrm{o}} = V_{\mathbf{x}}^{\mathrm{T}} \tag{7}$$

$$W_{\mathrm{w}} = \Lambda_{\mathbf{x}}^{-1/2} V_{\mathbf{x}}^{\mathrm{T}} \tag{8}$$

$$W_{\mathrm{s}} = V_{\mathbf{x}} \Lambda_{\mathbf{x}}^{-1/2} V_{\mathbf{x}}^{\mathrm{T}} \tag{9}$$

Matrix W_{o} produces a set of vectors $y_i(t)$ that are orthogonal to each other and whose Euclidean norms are equal to the eigenvalues of the data covariance matrix. This is what PCA does [15]. By using matrix W_{w}, we obtain a set of orthogonal vectors of unit norms, i.e. orthogonal vectors located on a spherical surface (*whitening*, or *Mahalanobis transform*). This property still holds true if any whitening matrix is multiplied from the left by an orthogonal matrix. In particular, if we use matrix W_{s} defined in Eq. (9), we have a whitening matrix with the further property of being symmetric. In

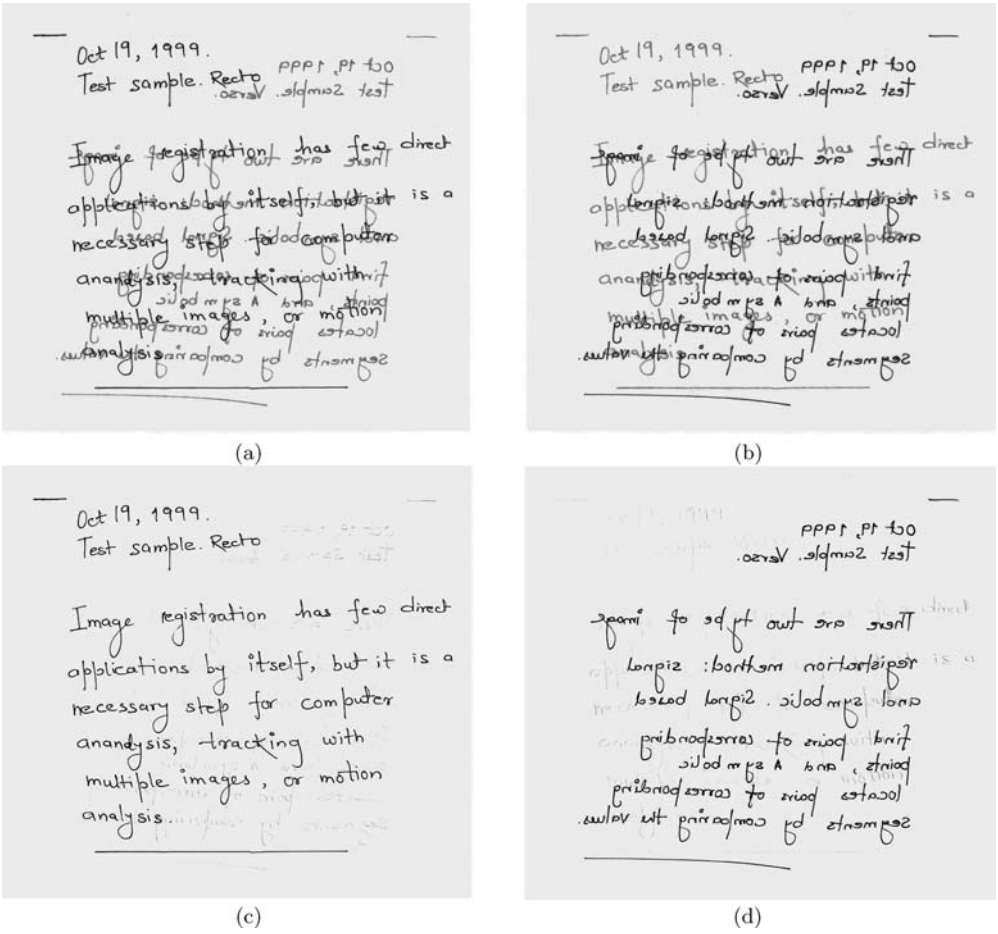


Fig. 1 Text separation with symmetric orthogonalization. **a** Recto of a real–fake manuscript showing bleed-through (from http://www.site.uottawa.ca/~edubois/documents). **b** Flipped verso of the same manuscript. **c** First symmetric orthogonalization output. **d** Second symmetric orthogonalization output

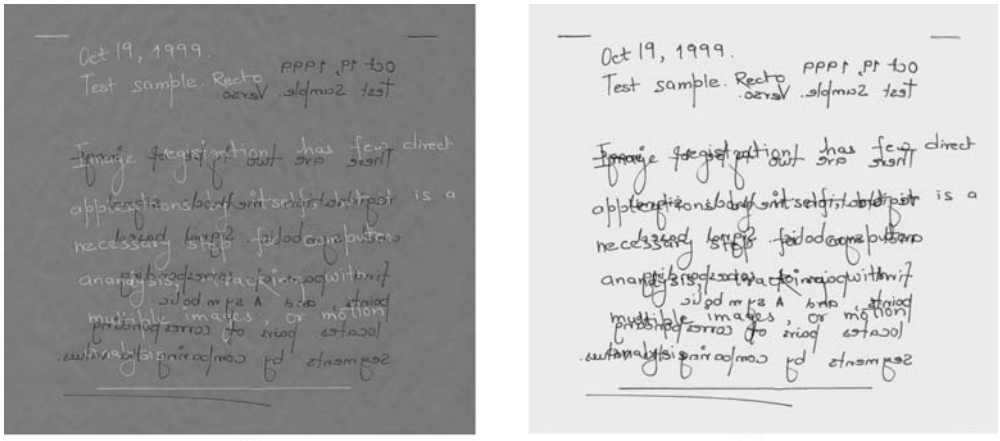
[15], it is observed that application of matrix W_s is equivalent to ICA when matrix A is symmetric. This is the property that we intend to exploit here, due to the mentioned features of the mixing matrix in the model of Eq. (1) assumed for our application. In this way, we can derive a separation technique that, while performing similarly to ICA, is much faster and simpler. In more general situations, ICA applies a further rotation to the output vectors, based on higher-order statistics.

However, the assumption of a symmetric diagonal-dominant matrix, with $A_{11} = A_{22}$, has not been explicitly introduced in the solutions, and can be used to check the results *a posteriori*. In other words, if our assumptions are reasonable, our estimate of the mixing matrix should approximately be symmetric, diagonal dominant, and with equal diagonal elements. In the 2×2 case considered here, for example, besides W_s, W_o is also symmetric. Thus, we can at first consider them as our candidate demixing matrices, since they verify at least one of our assumptions. Nevertheless, by using all the other assumptions to check the consistence of the solutions, we will see that only W_s is able to produce satisfactory separations, in accordance with the previously mentioned theoretical result that symmetric whitening is equivalent to ICA for symmetric mixing matrices.

4 Experimental results

Our experimental work has consisted in applying matrices W_o and W_s to typical images of documents degraded by bleed-through or show-through distortions, and in comparing the results with the ones produced by ICA. Our aim was to obtain clean texts in the whitened vectors. Of course, the results are different for different whitening matrices. We show here some examples from our experimentation.

The first example (Fig. 1) describes the processing of recto and verso pages of a real–fake manuscript affected by a strong bleed-through. This is a double-sided real manuscript, manually built so to obtain a natural bleed-through. We compared the results of the FastICA algorithm [18], the PCA, and the symmetric whitening, all applied to the two



(a) (b)

Fig. 2 Application of PCA to the recto and verso of the real–fake manuscript in Fig. 1. **a** First PCA output. **b** Second PCA output

views, and found that practically the same result is obtained by ICA and symmetric orthogonalization (SO). The estimated mixing matrices are

$$\hat{A}_{\mathrm{SO}} = \begin{bmatrix} 1.000 & 0.420 \\ 0.420 & 0.989 \end{bmatrix} \quad \hat{A}_{\mathrm{ICA}} = \begin{bmatrix} 1.000 & 0.496 \\ 0.368 & 1.043 \end{bmatrix}$$

Note that both matrices have been rescaled so as to have their first entries equal to 1, and both are diagonal dominant with diagonal elements nearly equal.

It is also interesting to examine the results produced by PCA, which gives the following estimated mixing matrix:

$$\hat{A}_{\mathrm{PCA}} = \begin{bmatrix} -0.7024 & 0.7118 \\ 0.7118 & 0.7024 \end{bmatrix}$$

Note that this matrix is not diagonal dominant, and that the separation result is unsatisfactory (Fig. 2). For example, the second output, i.e. the principal component corresponding to the highest eigenvalue, is given by a combination of the two input mixtures with coefficients of the same sign.

In Fig. 3, we report another example where we processed the recto and verso scans of a real document. Again, FastICA and symmetric orthogonalization perform similarly, giving a pretty good separation of the two main texts.

So far, we only evaluated the performance of our method qualitatively. Indeed, standard quantitative measures would require ground truth data to be available, and this is not feasible in actual practice, since ground truth can only be given in synthetic experiments. However, there is another strategy to evaluate, sometimes quantitatively, the performance of an algorithm. This is based on the improvements that can be achieved by coupling the algorithm under investigation to some subsequent task. Of course, this would be a measure of the performance of the latter task when the quality of the input data is modified, but this would also be a specific measure of usefulness for the algorithm under investigation. In our case, we assessed the usefulness of our method on image analysis procedures such as OCR. Since we only have a commercial OCR software available, we carried out our evaluation on printed texts. We considered recto and verso pairs captured from a printed book and affected by a strong show-through. We then evaluated the OCR performance before and after the application of our restoration method. After analyzing several pages, we found it difficult to quantify the improvement in recognition rate permitted by our procedure, since the strong interference prevented any character recognition almost everywhere when no restoration was applied. In most cases, the automatic OCR was not able to distinguish between text strings and grayscale images and, sometimes, show-through strokes were interpreted as punctuation marks or other symbols. Thus, we were only able to compare the OCR performance in the areas where a significant part of the text was correctly recognized even with no restoration. The results of one of the experiments performed are shown in Figs. 4 and 5. In Fig. 4, we show the original and processed versions of a recto–verso pair. The show-through has been removed quite effectively, but significant residuals can be observed in Fig. 4c and d. This depends, again, on the simplified model we use. In this case, it is the assumption of space invariance of the mixing matrix that is to be questioned, since the transparency of the paper is not actually uniform. Another critical issue is the necessary registration between recto and flipped verso pages. We noted that the quality of the results strongly relies on a very accurate registration. To be able to compare the results of OCR, we selected a part of the text where most characters were recognized even without restoration. In Fig. 5, we show the OCR outputs from this area, taken from about the middle of the recto page. The erroneous or spurious characters are highlighted by bounding boxes, while the wrongly merged or split characters are underlined. The advantage of applying our strategy is apparent, since OCR on the processed image only produced five errors, while OCR on the original scan produced dozens of misinterpetations and also much spurious text due to the presence of the interfering material.

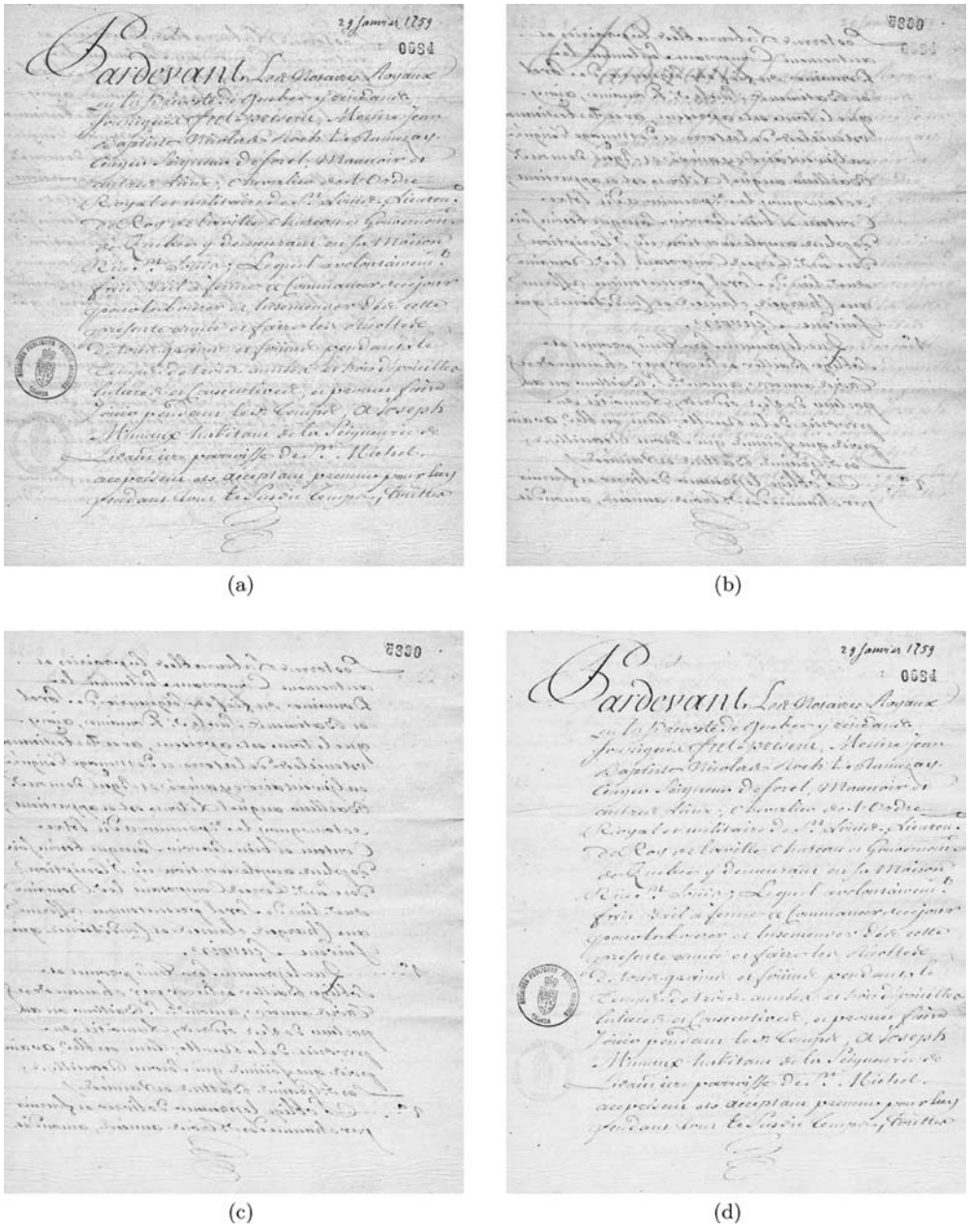


Fig. 3 Text separation by symmetric orthogonalization. **a** Recto of a real ancient document showing bleed-through (from http://www.site.uottawa.ca/~edubois/documents). **b** Flipped verso of the same document. **c** First symmetric orthogonalization output. **d** Second symmetric orthogonalization output

5 Conclusions

We have derived an approximated but physically sound linear model for describing the overlapping of texts in the grayscale recto and verso scans of documents affected by bleed-through or show-through distortion. In particular, this model satisfies some physical intuitions, such as the possible symmetry and the diagonal dominance of the mixing matrix that describes the relative ink attenuation in the two images. Based on the blind source separation theory, we developed a simple and fast technique which is able to separate the two mixed texts, thus producing clean versions of the recto and verso pages. This technique has its theoretical justification in the independent component analysis approach to BSS, and in the fact that symmetric data sphericization allows the sources to be separated when the mixing matrix is symmetric.

Unlike other methods proposed so far, our technique requires a single, very fast, processing step, with no need for segmentation or inpainting. Its usefulness is mainly related

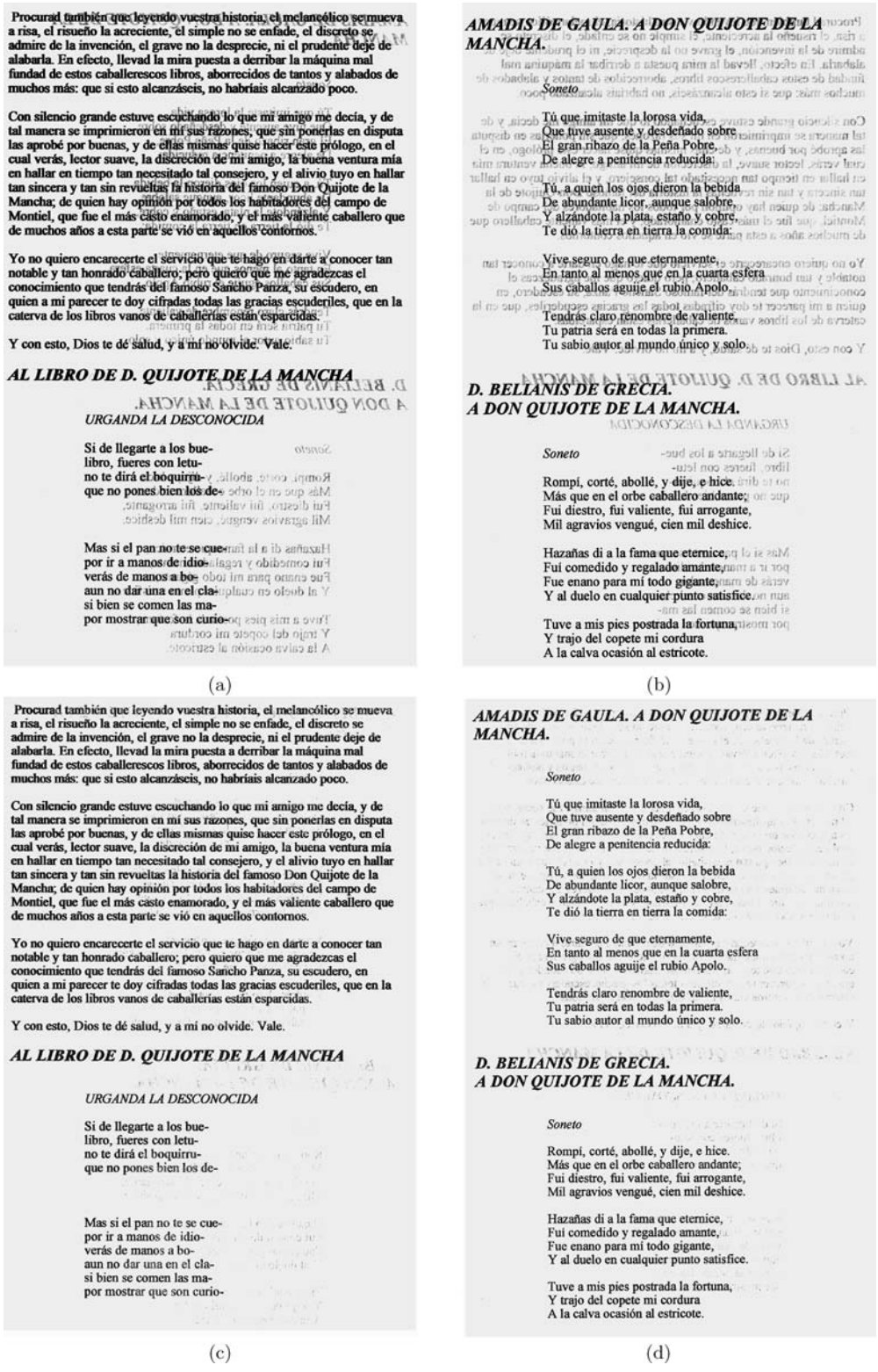


Fig. 4 Text separation with symmetric orthogonalization. **a** Recto of a printed document affected by show-through; **b** Verso of the same document. **c** First symmetric orthogonalization output. **d** Second symmetric orthogonalization output (flipped)

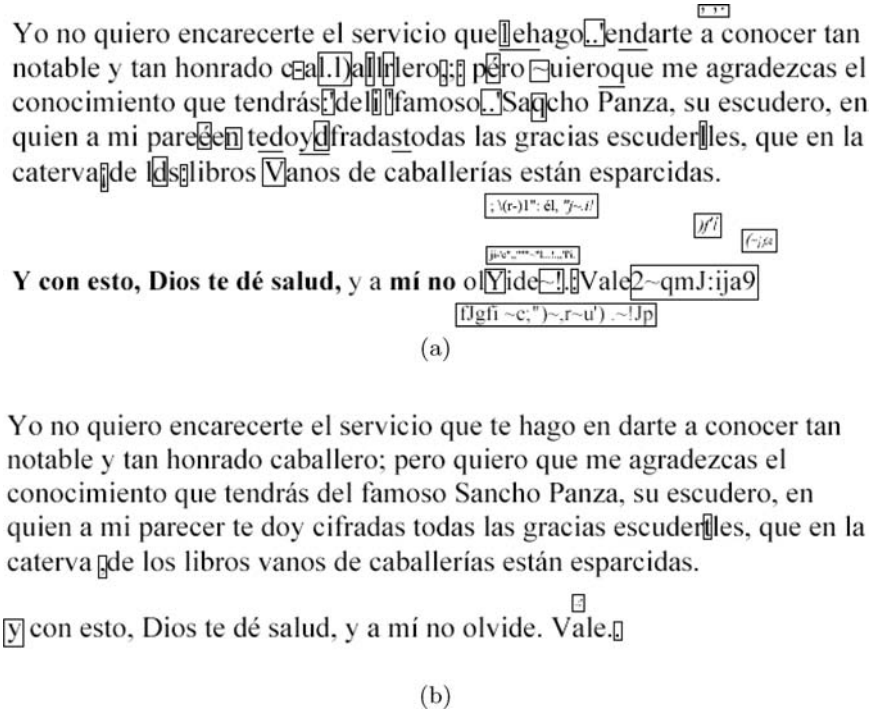


Fig. 5 Results of a commercial OCR on the unprocessed **a** and processed **b** versions of the same portion from Fig. 4a and c

to the possibility of quickly processing large databases of grayscale microfilms or digital scans, already available in many libraries and archives. Indeed, while new color or multispectral acquisitions could not be permitted, owing to the poor conservation status of the originals, grayscale recto and verso scans are often either available or readily obtained from microfilm archives.

Removing bleed-through or show-through is obviously of paramount importance for improving both human and automatic readability. In this respect, we showed some successful examples from an experimentation with real-ancient documents.

The criticality of our method is at present related to the oversimplified model assumed. In particular, linearity, instantaneousness, space invariance, lack of flexibility with respect to nonperfect registration often constitute serious limitations on the applicability of our approach.

Our research programs for the near future regard the development of more accurate numerical models for the phenomenon of pattern overlapping in documents, and the derivation of new BSS algorithms specifically designed for such models.

Acknowledgements This work has been supported by the European Commission Project "Isyreadet" (http://www.isyreadet.net), under contract IST-1999-57462, and by the European Project Network of Excellence MUSCLE - FP6-507752 (Multimedia Understanding through Semantics, Computation and Learning).

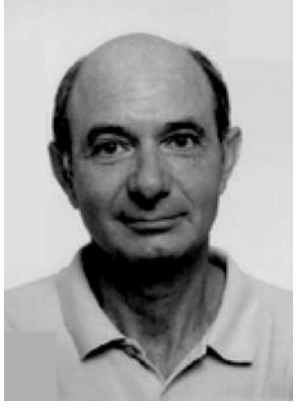
References

1. Leedham, G., Varma, S., Patankar, A., Govindaraju, V.: Separating text and background in degraded document images—a comparison of global thresholding techniques for multi-stage thresholding. In: Proceedings of the 8th International Workshop on Frontiers in Handwriting Recognition, Niagara on the Lake, Canada, pp. 244–249 (2002)
2. Govindaraju, V., Srihari, N.: Separating handwritten text from overlapping nontextual contours. In: Proceedings of the International Workshop on Frontiers in Handwriting Recognition, Chateau de Bonas, France, pp. 111–119 (1991)
3. Franke, K., Köppen, M.: A computer-based system to support forensic studies on handwritten documents. IJDAR **3**, 218–231 (2001)
4. Sharma, G.: Show-through cancellation in scans of duplex printed documents. IEEE Trans. Image Process. **10**(5), 736–754 (2001)
5. Dubois, E., Pathak, A.: Reduction of bleed-through in scanned manuscript documents. In: Proceedings of the IS&T Image Processing, Image Quality, Image Capture Systems Conference, Montreal, Canada, pp. 177–180 (2001)
6. Tan, C.L., Cao, R., Peiyi, S.: Restoration of archival documents using a wavelet technique. IEEE Trans. Pattern Anal. Machine Intell. **24**, 1399–1404 (2002)
7. Dano, P.: 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 (2003)
8. Nishida, H., Suzuki, T.: Correcting of show-through effects on document images by multiscale analysis. In: Proceedings of the 16th Conference on Pattern Recognition, Quebec City, Canada, pp. 65–68 (2002)

9. Nishida, H., Suzuki, T.: A multiscale approach to restoring scanned color document images with show-through effects. In: Proceedings of the 7th International Conference on Document Analysis and Recognition (ICDAR 2003) (2003)
10. Tonazzini, A., Bedini, L., Salerno, E.: Independent component analysis for document restoration. IJDAR **7**(1), 17–27 (2004)
11. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. Wiley, New York (2001)
12. Tonazzini, A., Salerno, E., Mochi, M., Bedini, L.: Bleed-through removal from degraded documents using a color decorrelation method. In: Document Analysis Systems VI, LNCS 3163, pp. 229–240. Springer, Berlin Heidelberg New York (2004)
13. Tonazzini, A., Salerno, E., Mochi, M., Bedini, L.: Blind source separation techniques for detecting hidden texts and textures in document images. In: Image Analysis and Recognition, LNCS 3212, Part II, pp. 241–248. Springer, Berlin Heidelberg New York (2004)
14. Salerno, E., Tonazzini, A., Bedini, L.: Digital image analysis to enhance underwritten text in the Archimedes palimpsest. IJDAR (submitted)
15. Cichocki, A., Amari, S.-I.: Adaptive Blind Signal and Image Processing. Wiley, New York (2002)
16. Bell, A.J., Sejnowski, T.J.: An information maximization approach to blind separation and blind deconvolution. Neural Comput. **7**, 1129–1159 (1995)
17. Ohta, Y., Kanade, T., Sakai, T.: Color information for region segmentation. Comput. Graph. Vis. Image Process. **13**, 222–241 (1980)
18. Hyvärinen, A., Oja, E.: Independent component analysis: algorithms and applications. Neural Netw. **13**, 411–430 (2000)

Anna Tonazzini graduated cum laude in Mathematics from the University of Pisa, Italy, in 1981. In 1984 she joined the Istituto di Scienza e Tecnologie dell'Informazione of the Italian National Research Council (CNR) in Pisa, where she is currently a researcher at the Signals and Images Laboratory. She cooperated in special programs for basic and applied research on image processing and computer vision, and is co-author of over 60 scientific papers. Her present interest is on inverse problems theory, image restoration and reconstruction, document analysis and recognition, independent component analysis, neural networks and learning.

Emanuele Salerno graduated in Electronic Engineering from the University of Pisa, Italy, in 1985. In September 1987 he joined the Italian National Research Council (CNR) at the Department of Signal and Image Processing, Information Processing Institute (now Institute of Information Science and Technologies, ISTI, Signals and Images Laboratory), Pisa, Italy, where he has been working in applied inverse problems, image reconstruction and restoration, microwave nondestructive evaluation, and blind signal separation. He has been assuming different responsibilities in research programs in nondesctructive testing, robotics, numerical models for image reconstruction and computer vision, neural networks techniques in astrophysical imagery. At present, he is local scientific responsible in the framework of the European Space Agency's "Planck Surveyor Satellite" mission, and takes part in the European CRAFT project "ISyReADeT", for document image restoration.



Luigi Bedini graduated cum laude in Electronic Engineering from the University of Pisa, Italy, in 1968. Since 1970 he has been a Researcher of the Italian National Research Council, Istituto di Scienza e Tecnologie dell'Informazione, Pisa, Italy. His interests have been in modelling, identification, and parameter estimation of biological systems applied to non-invasive diagnostic techniques. At present, his research interest is in the field of digital signal processing, image reconstruction and neural networks applied to image processing. He is co-author of more than 80 scientific papers. From 1971 to 1989, he was Associate Professor of System Theory at the Computer Science Department, University of Pisa, Italy.