IJDAR (2004) 7: 17–27
Digital Object Identifier (DOI) 10.1007/s10032-004-0121-8

International Journal on Document Analysis and Recognition IJDAR

Independent component analysis for document restoration

Anna Tonazzini, Luigi Bedini, Emanuele Salerno[*]

Istituto di Scienza e Tecnologie dell'Informazione, Area della Ricerca CNR di Pisa, Via G. Moruzzi, 1, 56124 Pisa, Italy

Received: April 15, 2003 / Accepted: December 17, 2003
Published online: 22 April 2004 – © Springer-Verlag 2004

Abstract. We propose a novel approach to restoring digital document images, with the aim of improving text legibility and OCR performance. These are often compromised by the presence of artifacts in the background, derived from many kinds of degradations, such as spots, underwritings, and show-through or bleed-through effects. So far, background removal techniques have been based on local, adaptive filters and morphological–structural operators to cope with frequent low-contrast situations. For the specific problem of bleed-through/show-through, most work has been based on the comparison between the front and back pages. This, however, requires a preliminary registration of the two images. Our approach is based on viewing the problem as one of separating overlapped texts and then reformulating it as a blind source separation problem, approached through independent component analysis techniques. These methods have the advantage that no models are required for the background. In addition, we use the spectral components of the image at different bands, so that there is no need for registration. Examples of bleed-through cancellation and recovery of underwriting from palimpsests are provided.

Keywords: Degraded documents – Blind source separation – Independent component analysis – Palimpsest restoration– Bleed-through cancellation

1 Introduction

Many digital images of documents, either ancient or modern printed texts and manuscripts, are degraded by the presence of strong artifacts in the background. This can either make difficult the readability of the text or, more often, compromise the efficiency of even "intelligent" OCR systems [3]. Background artifacts can derive from many kinds of degradations such as scan optical blur and noise, spots, underwriting, or overwriting. We focus here on the so-called bleed-through or show-through effects, and on palimpsests, that is ancient manuscripts that have been erased and then rewritten.

Bleed-through is intrinsic in many ancient documents because it is caused by seeping of ink from the reverse side, while in modern, double-sided printed documents, show-through appears in the scanned image when the paper is not completely opaque. It is clear that, for an OCR system to work efficiently, these interfering strokes have to be at least significantly attenuated. For palimpsests, the problem is the opposite, and what is desired is to enhance and let "emerge" the traces of the original underwriting. In both cases, the documents can be seen as made of a primary text, to be enhanced, overlapped with a textured background, to be removed.

The use of thresholding techniques to remove the background is often not effective since the intensities of the unwanted background can be very close to those of the foreground text. In these conditions, thresholding either does not remove the background or also eliminates part of the information in the text of interest. Thus, adaptive and/or structural approaches have to be adopted. For instance, in [19] several thresholding techniques are compared for separating text and background in degraded, historical documents, and the result is that neither global nor local thresholding can perform satisfactorily. The authors suggest then the investigation of multistage thresholding techniques. In [10], segmentation and grouping techniques, based on the Gestalt cognitive rules from the visual system behavior, are used to eliminate interfering strokes from skeletonized versions of handwritten documents. In [9] local, adaptive filters, also based on evolutionary algorithms, are applied for homogeneous and textured background removal, still from handwritten gray-level documents.

For the specific problem of bleed-through/show-through, some work has been done, mainly based on the exploitation of information from the front and back

[*] This work has been partially supported by the European Commission project "Isyreadet" (http://www.isyreadet.net), under contract IST-1999-57462)

Correspondence to: Anna Tonazzini (e-mail: anna.tonazzini@isti.cnr.it)

pages. In [23] the physical model of these effects is first simplified for deriving a linear mathematical model, and then an adaptive linear filter is developed that uses scans of both sides of the documents. In [7], the two sides of a gray-level manuscript are compared at each pixel, and, basically, a thresholding is applied. In [24] a wavelet technique is applied for iteratively enhancing the foreground strokes and smearing the interfering strokes. In all these methods, a preliminary registration of the two sides is required. In [22] the front side alone of a color image is processed via a multiscale analysis employing adaptive binarization and edge magnitude thresholding.

We adopt the point of view that extracting the text of interest or, equivalently, removing the interfering pattern can be seen as the problem of separating overlapped texts. Thus, in this paper we propose an approach based on reformulating the problem as one of blind source separation (BSS), where the overlapping texts and the support (paper, parchment, etc.) texture are the unknown sources to be recovered, and multiple acquisitions of the documents in different spectral bands are the observations. The spectral bands considered can be the red, green, and blue channels in which any color image can always be split, and/or nonvisible channels, e.g., in the infrared band.

As highlighted in [23], the physical model underlying both bleed-through and show-through is very complicated in that it is nonlinear with some unknown parameters, and it should also account for the spreading of light or ink in the support, which causes a blurring with unknown characteristics of the show-through component. In [23] suitable transformations and simplifying approximations are adopted to linearize the problem. In this first stage of our study, we also adopted a linear approximation and derived a model in which the observations are seen as linear mixtures of the sources themselves. Unfortunately, however, the coefficients of these mixtures are usually unknown. The problem of separating the overlapped texts thus becomes the highly undetermined problem of jointly estimating the sources and the mixing coefficients. Nevertheless, under particular assumptions, the problem can be efficiently solved through independent component analysis techniques. In particular, we employ the FastICA algorithm [11], which is a fully blind and extremely fast procedure. Furthermore, since our data are the different spectral components of the same image, there is no need for registration. Some useful results can be obtained within the limits imposed by our oversimplified linear model. A few examples are shown from both real and synthetic data images.

The paper is organized as follows. In Sect. 2 the linear BSS problem is formulated in a general context, and the basic principles of independent component analysis and of the FastICA algorithm are stated. Section 3 is devoted to the assessment of the model for images of overlapped texts and then to the formulation of their separation as a BSS problem of a linear, instantaneous mixture of unknown sources and unknown mixing coefficients. A first, artificial example of the performance of the FastICA algorithm is provided as well. In Sect. 4 we provide experimental results of the proposed method for the recovery of partially hidden underwritings from palimpsests. We compare our method with the one applied to separate the overlapped texts in the famous Archimedes Palimpsest [8]. Finally, in Sect. 5 we give the results of the separation of foreground text and bleed-through text from artificial images and from images of ancient documents showing a real bleed-through effect.

2 Blind source separation and independent component analysis

Blind source separation (BSS) became an active research topic in signal processing in the last decade. The most known applications of BSS include the so-called "cocktail party" problem in audio processing, the removal of underlying artifact components of brain activity from EEG records, and the search for hidden factors in parallel financial data series. Only very recently has BSS received attention in image processing and computer vision, e.g., for feature extraction or noise removal from natural images, and source separation in astrophysical microwave maps (see, for example, [13,16]). BSS consists of separating a set of unknown signals from a set of mixtures of them when no full knowledge is available about the mixing operator. The most studied BSS problem refers to a linear data model, where the observations are linear instantaneous mixtures, with unknown coefficients, of the source signals.

So far, many techniques have been proposed to solve this severely ill-posed inverse problem. Among them, the independent component analysis (ICA) methods are based on the assumption of mutual independence of the sources. Most of these methods were developed in the case of noiseless data and differ from one another in the way they enforce independence. The maximum likelihood (ML) method [5] directly assumes a factorized form for the joint source distribution; in the infomax method [18], entropy is used as a measure of independence; other methods are based on the minimization of contrast functions, related to statistics of order greater than two, still to ensure independence [6]. The strict relationships among the various methods have been investigated as well [1], and in [15,17,20], Bayesian estimation has been proposed as a suitable, unifying framework for BSS within which the other methods can be viewed as special cases. All the above methods have shown good performance in many practical applications. In particular, a very efficient and fast algorithm, the FastICA algorithm, has been proposed in [11]. However, the ICA solution to BSS presents some drawbacks or not yet explored/resolved issues. In fact, the independence requirement can be fulfilled in some practical applications, but in many cases there is clear evidence of correlation among the sources. Furthermore, most ICA algorithms have been developed for noiseless data and do not account for a possible time correlation inside the single sources and/or for different numbers of sources and data signals. Finally, convolutive or nonlinear mixtures could better fit some practical problems. All the above limitations are currently under investigation. In particular, to manage noisy mixtures, the noisy FastICA algorithm [12], and an independent factor analysis method (IFA) [2,16,21] have been developed. Other techniques for the separation of noisy mixtures take advantage of their ability to incorporate into the problem available information about autocorrelation properties of the individual sources [25–27]. Indeed, correlation is an important feature of most real-world signals, especially of images, and, when used as a constraint, it is known to be able to regularize many ill-posed problems. Finally, the possibility of separating cross-correlated sources is an emerging topic as well [4].

Although document images present many of the above features (auto- and cross correlation of the sources, scanner noise, etc.), in this first application of BSS and ICA, we will develop our method in an idealized setting, with the main aim of showing the novelty and the great potentiality of this kind of approach to the processing and analysis of degraded documents.

The data generation model for a noiseless linear and instantaneous BSS problem is given by

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

where $\mathbf{x}(t)$ is the vector of the measurements, $\mathbf{s}(t)$ is the column vector of the unknown sources at location t, and A is the unknown mixing matrix. We assume the same number N of measured and source signals, so that A is an $N \times N$ matrix.

Obviously, solving the system in Eq. 1 with respect to both A and $\mathbf{s} = (\mathbf{s}(1), ..., \mathbf{s}(T))$ would be an undetermined problem, unless more information were exploited. The kind of information used in the ICA approach is the independence of the sources. Assuming the prior distribution for each source is known, the joint prior distribution for $\mathbf{s}$ is thus given by

$$P(\mathbf{s}) = \prod_{i=1}^{N} P_i(\mathbf{s}_i)\,, \tag{2}$$

where $\mathbf{s}_i = (s_i(1), s_i(2), ..., s_i(T))$. The separation problem can be formulated as the maximization of Eq. 2, subject to the constraint $\mathbf{x} = A\mathbf{s}$. This is equivalent to the search for a W, $W = (\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_N)'$, such that, when applied to the data $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N)$, it 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. 2, the problem solved by ICA algorithms is then

$$\hat{W} = \arg\max_{W} \sum_t \sum_i \log P_i(\mathbf{w}_i'\mathbf{x}(t)) + T \log|\det(W)|. \tag{3}$$

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.

It has been shown that, besides independence, to make separation possible, a necessary extra condition for the sources is that they all, but at most one, must be non-Gaussian. To enforce non-Gaussianity, generic super-Gaussian or sub-Gaussian distributions can be used as priors for the sources. These have proven to give very good estimates for the mixing matrix and for the sources as well, regardless of the true source distributions, which, on the other hand, are usually unknown [5].

The FastICA algorithm [14] gives the possibility of choosing among a number of "nonlinearities" to be used in place of the derivatives of the log distributions. It solves Eq. 3 and returns the estimated sources by using a fixed-point iteration scheme [11] that has been found in independent experiments to be 10 to 100 times faster than conventional gradient descent methods.

3 Formulation of the overlapped text separation as a BSS problem

We assume hereafter that a palimpsest image or a document image affected by bleed-through/show-through can be modelled as the superposition of three different sources, or classes, that we will call "background", "overwriting", and "underwriting", respectively. In the BSS formalism, this means that we have three different sources that combine in some way to give the observed image. At the same time, as already mentioned, 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 documents containing text, 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 overwriting, and (r_3, g_3, b_3) for the underwriting.

For the superposition of the three classes we developed an approximated linear mixture model assuming that, at each point of the document, the three reflectance indices of the three classes 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 document can be seen as given by the following equation:

$$\begin{bmatrix} x_r(t) \\ x_g(t) \\ x_b(t) \end{bmatrix} = \begin{bmatrix} r_1 & r_2 & r_3 \\ g_1 & g_2 & g_3 \\ b_1 & b_2 & b_3 \end{bmatrix} \begin{bmatrix} s_1(t) \\ s_2(t) \\ s_3(t) \end{bmatrix}\,, \tag{4}$$

where functions $s_i(t)$, $i = 1, 2, 3$ indicate the "quantity" of background, overwriting, and underwriting, respectively, that concur to form the color at point t. For instance, the reflectance indices of, say, a pure background point t, will be given by

$$\begin{bmatrix} x_r(t) \\ x_g(t) \\ x_b(t) \end{bmatrix} = \begin{bmatrix} r_1 & r_2 & r_3 \\ g_1 & g_2 & g_3 \\ b_1 & b_2 & b_3 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}\,. \tag{5}$$

It can be immediately verified that Eq. 4 is of the same form as Eq. 1, restricted to the 3×3 case, where parameters r_i, g_i, and b_i are the coefficients of the mixing matrix A, and functions $s_i(t)$ are the sources. As stated earlier, in this first application of the method, we assume

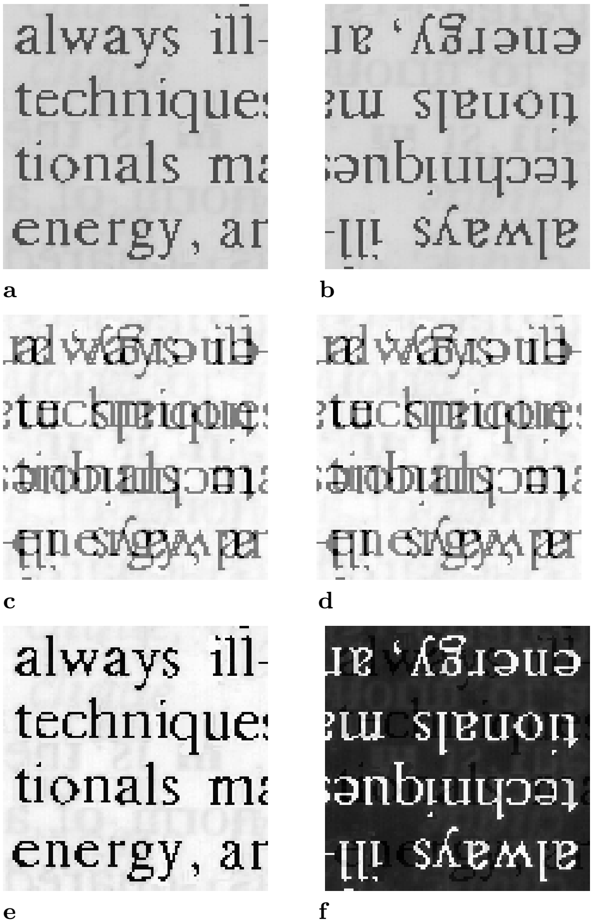


Fig. 1. Synthetic noiseless example. **a** First original image. **b** Second original image. **c** First mixture. **d** Second mixture. **e** First FastICA output. **f** Second FastICA output

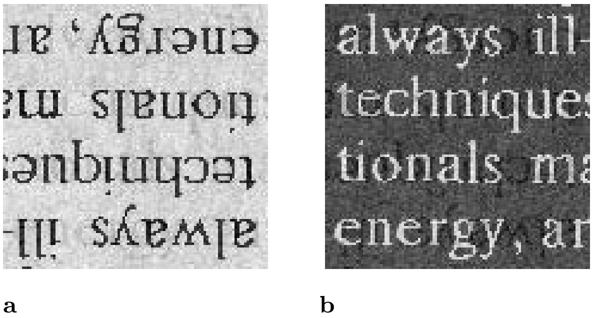


Fig. 2. Synthetic noisy example (40 dB SNR). **a** First FastICA output. **b** Second FastICA output

that noise or blur in the documents can be neglected, i.e., that the data model is a noiseless instantaneous mixture. In these conditions, the ICA principle described in the previous section can be applied to our data model, and both matrix A and the functions $s_i(t)$ can be estimated by using the FastICA algorithm.

If the data are actually affected by a moderate amount of noise, FastICA is still able to give a good estimate of A. In this case, separation can still be achieved, but the output noise-to-signal ratio will be amplified by the condition number of matrix A. If the data are affected by significant noise, the estimate of the mixing matrix provided by FastICA will be biased and separation can no longer be obtained. A step toward the solution to this problem for Gaussian noise with known covariance matrix is the noisy FastICA algorithm [12].

For a correct application of the 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. However, when the mixing matrix is ill-conditioned or singular, extra observations taken in nonvisible spectral bands, used in conjunction or in place of the visible channels, could be a remedy. A pointer to quasisingularity could be the spectrum of the data covariance $C_{\mathbf{x}}$. Indeed, if the sources are mutually independent, without loss of generality we can assume that their covariance is the identity matrix. Hence, in the noiseless case, $C_{\mathbf{x}}$ is

$$C_{\mathbf{x}} = < \mathbf{x}\mathbf{x}' > = AA' \,, \tag{6}$$

where <> means expectation. Therefore, the condition number of the data covariance matrix coincides with the squared condition number of matrix A.

Below we give a first example of the performance of the FastICA algorithm for separating two overlapped images. The example is completely artificial, though we used as a basis the scan of a real gray-level document. We considered this image as the first source (Fig. 1a) and a rotated version of it as the second source (Fig. 1b). To generate the observations, we linearly mixed the two available sources with the following, randomly generated, 2×2 matrix:

$$A = \begin{bmatrix} 0.6992 \; 0.7275 \\ 0.4784 \; 0.5548 \end{bmatrix} .$$

Note that these observations are not intended to simulate the color components of an image containing overlapped texts, but they represent just the linear superposition of two images, without reference to any specific application. The two mixtures are shown in Figs. 1c and d, respectively. Applying FastICA to these observations, we obtained the images shown in Figs. 1e and f, and the 2×2 estimated matrix:

$$\hat{A} = \hat{W}^{-1} = \begin{bmatrix} 0.7986 & -0.7152 \\ 0.5542 & -0.5459 \end{bmatrix} .$$

To check the quality of this estimate, we consider the matrix $\hat{W}A$. In the case of a perfect separation, this should be a permutation and scaling matrix, so that only one element per row and per column should be nonzero. Note that this element represents the scale factor affecting the related reconstructed source. In the practical case of a nonperfect separation, each row and column will only have a dominant element. The root-mean-square value (RMS) of the nondominant elements, each divided by the dominant element of its column, will be used as a quality index for the estimate. In our case we have

$$\hat{W}A = \begin{bmatrix} 0.9987 & 0.0088 \\ 0.1376 & -1.0073 \end{bmatrix} ,$$

and $RMS = 0.0976$.

The estimated sources, to be compared with the originals of Figs. 1a and b, are shown in Figs. 1e and f. Note that, due to the negative scale factor in the second column of the estimated matrix, the second estimated source is the negative image of the original source.

In order to show the effect of a small amount of noise on the data, we executed the same procedure after adding a Gaussian noise with SNR = 40 dB to the above mixtures. In this case we obtained a satisfactory, though worse, estimate of matrix A. The quality index computed for this matrix is

$$\hat{W}A = \begin{bmatrix} 0.1454 & 0.9553 \\ -0.9673 & 0.2192 \end{bmatrix},$$

and $RMS = 0.1940$. Consequently, we obtained the separation of the overlapped texts, as shown in Figs. 2a and b, but, as expected, the reconstructed sources show a considerable noise amplification (the condition number of the original mixing matrix was about 37).

As mentioned above, this example was totally artificial. When applied to real document images the situation is more complicated and we have to cope with a number of problems. First, the linear mixture model of Eq. 4 does not describe correctly the actual image features. Indeed, in all the regions where overwriting is superimposed on other image components, the resulting color is not likely to be a linear mixture of the colors of the individual components. In other words, where overwriting is present, it masks the background rather than combining linearly. Thus, a necessary step toward the resolution of our problem would be the derivation of a suitable nonlinear model. The possibility of adopting an ICA strategy to separate nonlinearly mixed sources is currently under study; see, for example, [13]. Another important problem is the possible cross correlation of the sources, which makes the basic ICA assumption not satisfied. For example, this can happen when a large overlap area exists between overwriting and underwriting strokes. In fact, this happens in the above example, where the cross correlation between the two sources was about 13%, and this could be an additional reason for the poor performance of FastICA against noise in that case. However, extensions of ICA to correlated sources are now being studied [4].

Despite these difficulties, we obtained good experimental results from our linear, noiseless model. Some of them are shown in the following sections.

4 Recovery of underwriting from palimpsests

For the case of palimpsests, the information of interest is usually the old text that has been erased before rewriting a new one. In this situation, underwriting often disappears in the visible range, but very frequently some traces of it appear in the the infrared band. For this reason, one or more infrared channels can be used in addition to or in substitution of the visible channels in order to permit the separation of the three classes.

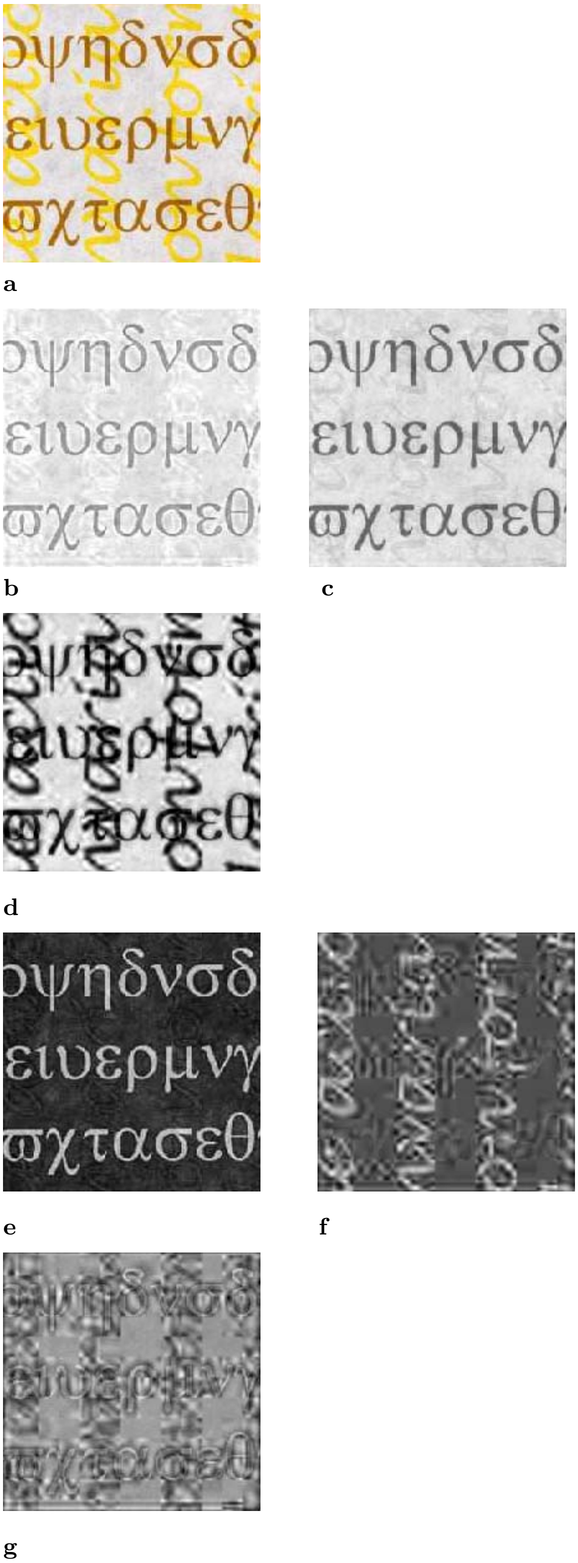
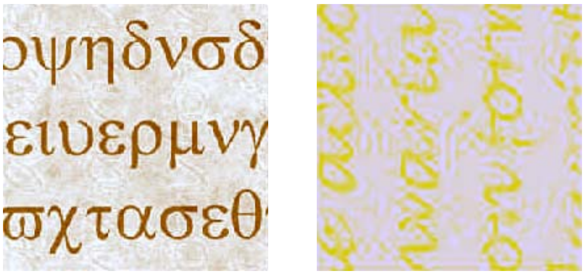


Fig. 3. Synthetic example of a palimpsest. **a** Color image. **b** Red channel. **c** Green channel. **d** Blue channel. **e** First FastICA output. **f** Second FastICA output. **g** Third FastICA output

In the following discussion we will give a first example of the application of our method for recovering



a

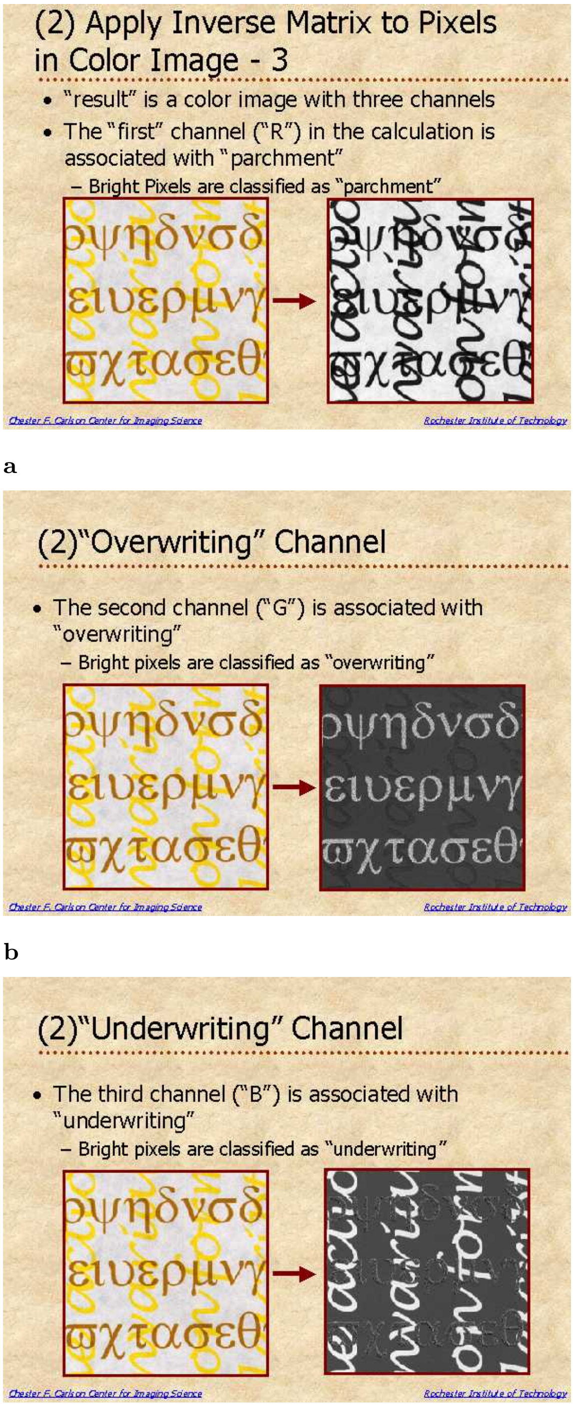
b

Fig. 4. Recovered color images for the separations in Fig. 2. **a** First FastICA output. **b** second FastICA output

the underwriting from an artificial palimpsest proposed as an exercise on the Web site of Roger L. Easton of the Rochester Institute of Technology [8]. Easton uses this exercise to explain the technique that he and his group adopted for the restoration of the Archimedes Palimpsest. He proposed two methods: a simpler one that is able to recover only the underwriting and a more sophisticated one that is able to recover all the three sources (overwriting, underwriting, and background). In this second method, for the color at each point of the image, they adopt a linear mixture model like the one shown in Eq. 4 but assume to know in advance the mixing coefficients. These are estimated by manually reading the values of the red, green, and blue components in a number of points of each class and then computing an average value for each color and each class. The classes are reconstructed by inverting the 3×3 matrix so obtained. Thus, their method is basically the same as ours, with the only, but relevant, difference that their method is not blind, i.e., it requires the a priori knowledge of the mixing matrix, while our method is fully blind, i.e., it estimates the mixing matrix jointly with the sources. It has to be noted that the manual estimation of the mixing matrix is a cumbersome task, which is possible only when there are areas in the image where the three classes are well separated and distinguishable from each other.

In Fig. 3a we show the original color image, while the three channels, the observations, are provided in Figs. 3b, c and d, respectively. Given these observations alone, FastICA is able to reconstruct the sources shown in Figs. 3e, f and g, respectively.

Note that, since the elements of the recovered mixing matrix are the estimated mean reflectance indices $\hat{r}_i, \hat{g}_i, \hat{b}_i$ for the fundamental colors of each class, it is possible to reconstruct the color of the separated images. More specifically, considering the generic estimated source $\hat{\mathbf{s}}_i$, its three color components will be given by $\hat{r}_i\hat{\mathbf{s}}_i$, $\hat{g}_i\hat{\mathbf{s}}_i$, and $\hat{b}_i\hat{\mathbf{s}}_i$, respectively. Figure 4 shows the result of this procedure applied to the maps of Figs. 3e and f. In these color images, the background does not correspond to the "background class" but must be intended as a residual, due to the other sources, of a nonperfect separation. Apart from a permutation in the outputs of the FastICA algorithm, our results are similar to those provided in Easton's Web site and shown in Figs. 5a, b, and c. A higher contrast and a cleaner appearance of our estimated sources can be achieved by simple postprocessing, such as thresholding.



(2) Apply Inverse Matrix to Pixels in Color Image - 3

- "result" is a color image with three channels
- The "first" channel ("R") in the calculation is associated with "parchment"
  - Bright Pixels are classified as "parchment"

Chester F. Carlson Center for Imaging Science Rochester Institute of Technology

a

(2)"Overwriting" Channel

- The second channel ("G") is associated with "overwriting"
  - Bright pixels are classified as "overwriting"

Chester F. Carlson Center for Imaging Science Rochester Institute of Technology

b

(2)"Underwriting" Channel

- The third channel ("B") is associated with "underwriting"
  - Bright pixels are classified as "underwriting"

Chester F. Carlson Center for Imaging Science Rochester Institute of Technology

c

Fig. 5. Treatment of a synthetic example of a palimpsest (from: "Text recovery from the Archimedes Palimpsest", www.cis.rit.edu/people/faculty/easton/k-12/exercise/index.htm). **a** Estimated background. **b** Estimated overwriting. **c** Estimated underwriting

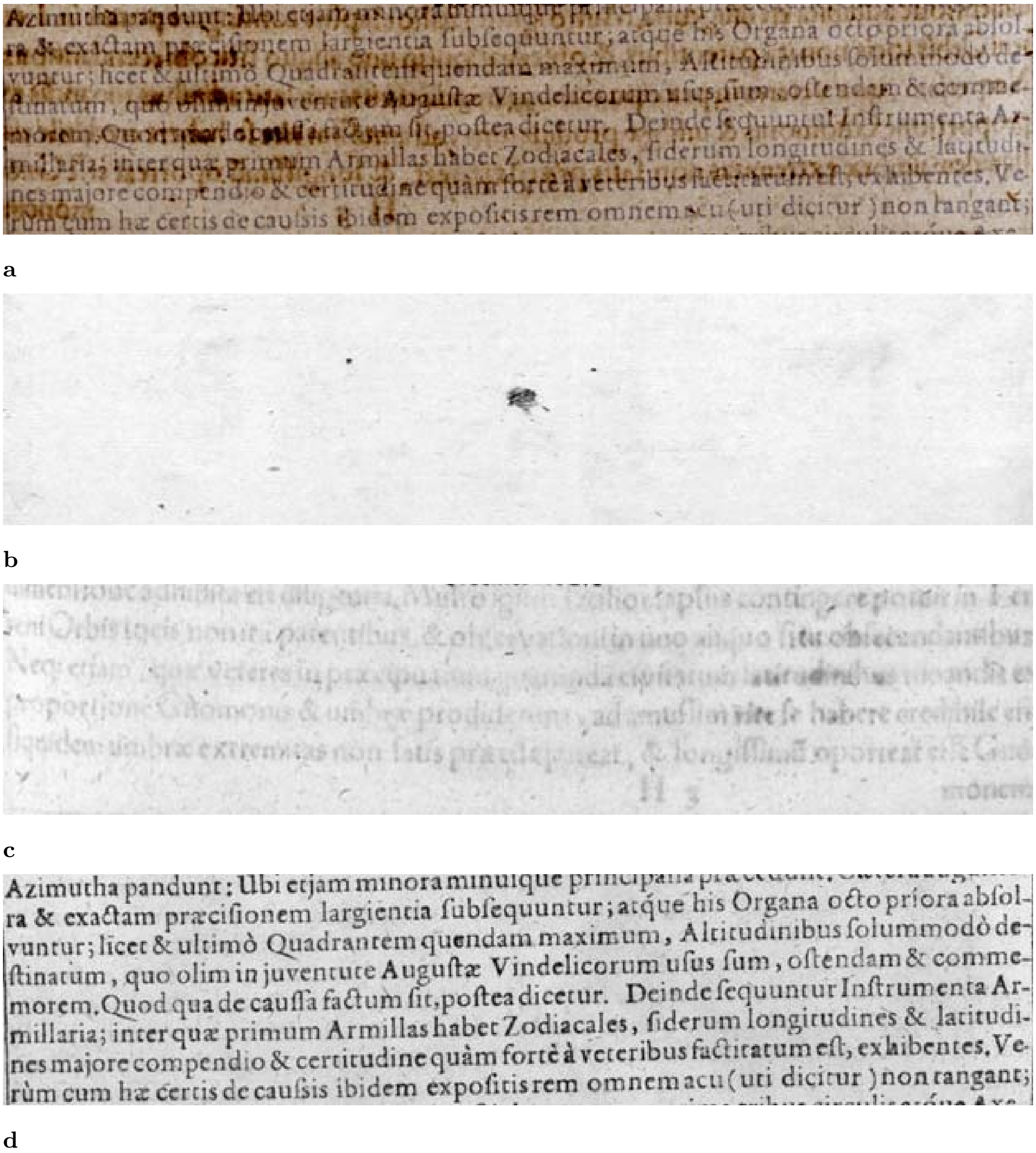


Fig. 6. Synthetic image. **a** Color image. **b** First FastICA output. **c** Second FastICA output (mirrored and inverted). **d** Third FastICA output

5 Separation of bleed-through and show-through

In this section, we provide the results obtained with our ICA approach for the separation of foreground text and bleed-through/show-through from a synthetic color image generated according to the model of Eq. 4 and from real images of ancient documents.

Figure 6 shows the results of the experiment performed on the synthetic color image (Fig. 6a). This image has been obtained by artificially mixing a gray-level text image, a gray-level nonuniform background, and a gray-level bleed-through, by using the following matrix:

$$A = \begin{bmatrix} 0.7200 \; 0.3600 \; 0.4500 \\ 0.7000 \; 0.3500 \; 0.6000 \\ 0.5200 \; 0.5200 \; 0.7800 \end{bmatrix}$$

The first column ideally represents the reflectance indices of the front side text in the RGB channels. The second column represents the reflectance indices of the background. Finally, the third column represents the reflectance indices of the bleed-through.

The output images are shown in Figs. 6b–d. The image in Fig. 6c has been mirrored as it reproduces the bleed-through pattern. In addition, its grayscale has been inverted to permit a better readability. The recovered matrix was in this case

$$\hat{A} = \begin{bmatrix} 1.4224 \; -21.0528 \; 24.7754 \\ 1.5021 \; -28.0178 \; 24.0673 \\ 2.2020 \; -36.2938 \; 17.8483 \end{bmatrix},$$

which, as often happens, differs from the actual one for scale factors and column permutations. Matrix $\hat{W}A$ is

$$\hat{W}A = \begin{bmatrix} -0.0004 & 0.3089 & 0.0013 \\ -0.0001 & 0.0049 & -0.0215 \\ 0.0290 & 0.0009 & -0.0002 \end{bmatrix} .$$

As mentioned above, for a correct separation, this matrix should have a dominant element in each row and each column. The positions of the dominant elements are determined by the permutations that occurred among the source images; each dominant value is the scale factor that affects the related source. In our case, the first row has the dominant value as the second element. This means that the first FastICA output corresponds to the second input source, i.e., the background. Analogously, the second output corresponds to the bleed-through, and the third output to the foreground text. The value of RMS is in this case 0.0262, indicating an average 2.6% contamination on each output image due to residuals from the other sources.

We highlight again that, while in the example of the palimpsest we were able to recover the underwriting only where it was not masked by the overwriting, in this case it has been possible to recover the bleed-through even where it appears as occluded by the foreground text. Indeed, at the occlusion points in the palimpsest the colors of the two classes do not mix, and the overwriting completely covers the underwriting. The case of Fig. 6 is so successful because the data have been exactly generated following Eq. 4. Nevertheless, we found that bleed-through recovery has been achieved in a number of real situations, as we show in the example below. However, when the visible channels do not allow us to recover the bleed-through, this could be obtained by using additional channels in nonvisible bands. Mathematically, as stated earlier, this is related to a possible improvement of the conditioning of the mixing matrix. Thus, if the linear model is at least approximately justified, the bleed-through image can be recovered in the same way as the underwriting image in palimpsests. This could be of interest, for example, with scans of modern, well-preserved documents, where show-through may appear because of the type of paper used and a scan of the reverse page is not available. However, we must point out that by far the most interesting and promising application of our technique is the removal of background artifacts, with subsequent extraction of a clean primary text. This, as was already highlighted, is of paramount importance to significantly improve OCR tasks. In fact, in all our experiments, even in those unfortunate cases where it was not possible to extract the bleed-through, we always obtained a clean image of the foreground text as one of our outputs.

We now show a real case where the bleed-through has been effectively recovered. Figure 7a shows the original color image of an ancient document, while Figs. 7b and c show two of the three outputs obtained by using the red, green, and blue channels. As anticipated, in this case we obtained all the three classes that form the image. The first output image (not shown) is related to the background class, apart from residuals of the other two classes, due to an imperfect separation. Similarly, the second output image corresponds to bleed-through, while the third represents the extracted foreground text. In this case, the recovered bleed-through is completely unreadable since it is highly blurred. Some standard postprocessing of the image in Fig. 7b, such as median filter and histogram stretching, allowed us to obtain the better quality image of Fig. 8. However, we think that, to significantly improve the readability of the bleed-through image, the operator associated with ink spreading has to be suitably formalized and included in the image model. The separation could then be improved by more sophisticated ICA algorithms, exploiting, for example, the smoothness constraint implied by the image model. Once the blurred components are recovered, standard or ad hoc deblurring techniques can be used to improve legibility. A last example shows an

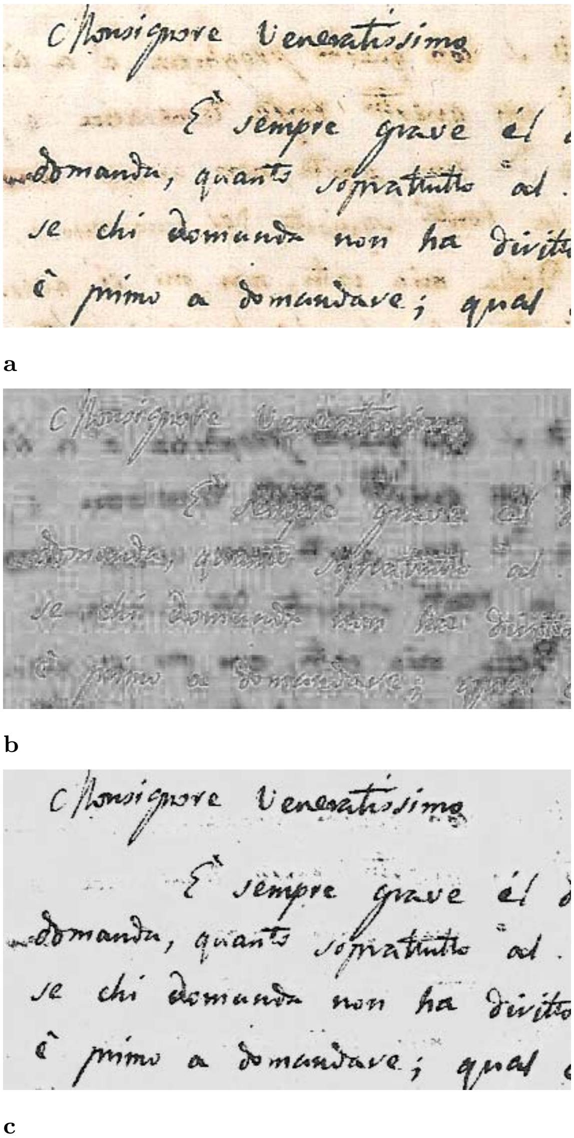


Fig. 7. Real document showing bleed-through. **a** Original color image. **b** Second FastICA output. **c** Third FastICA output

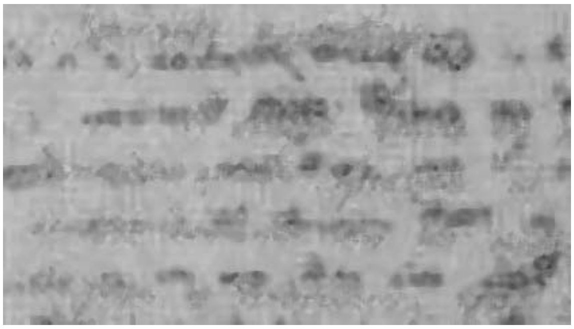


Fig. 8. Bleed-through source after postprocessing

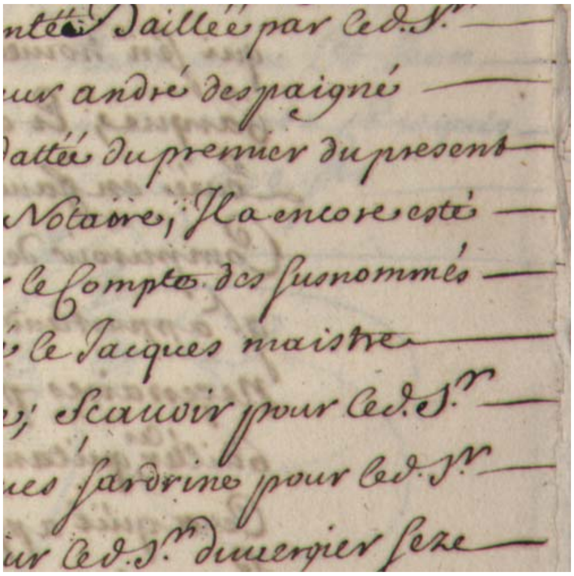


Fig. 9. Portion of a color scan of an ancient manuscript

instance where bleed-through recovery was unsuccessful, but the advantages of using additional channels are apparent. Figure 9 shows a portion of a color scan of an ancient manuscript. Using the RGB components of this image the FastICA outputs were the ones shown in Fig. 10. Note that the overwriting and the bleed-through are not correctly separated (Figs. 10a and b). However, as shown in Fig. 10c, an unexpected pattern has been extracted. This might be due to transparency from an underlying page and was barely detectable in the color image. This brought us to assume the present document as being made of at least four overlapped components and to use two additional scans available in the near-infrared and ultraviolet bands. By applying FastICA to the five channels thus obtained, we still were unable to recover an isolated bleed-through pattern, but the foreground text extracted is now much cleaner than before (Fig. 11). The remaining outputs are similar to the ones already shown; the above-mentioned unexpected component was extracted also in this case and is almost free of interferences from the other components.

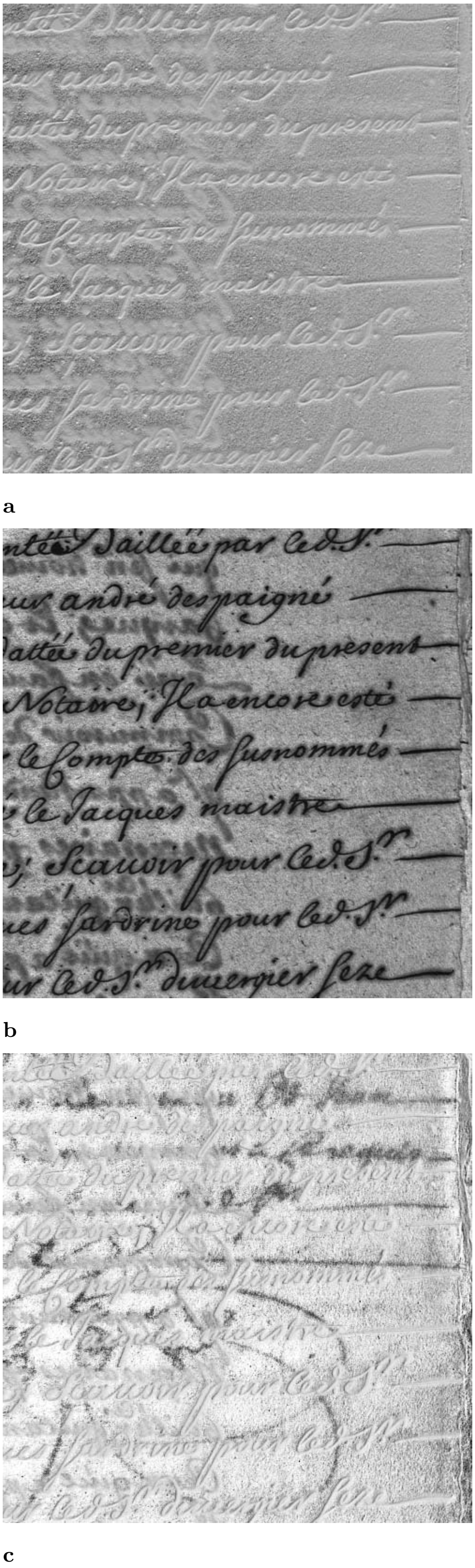
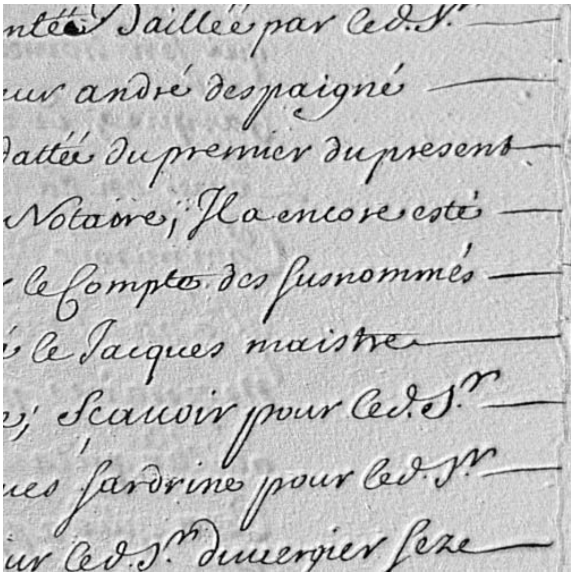


Fig. 10. Applying FastICA to the RGB components of the document in Fig. 9. **a** First FastICA output. **b** Second FastICA output. **c** Third FastICA output



ntée Baillée par Ced. Sr. —
eur andré despaigné —
datté du premier du present —
Notaire; Il a encore esté —
le Compte des Susnommés —
de Jacques maistre —
y; Scavoir pour Ced. Sr. —
ues Jardine pour Ced. Sr. —
eur Ced. Sr. duuerpier Leze —

Fig. 11. Recovered overwriting by using the additional near-infrared and ultraviolet channels of the document in Fig. 9

6 Conclusions

We presented preliminary results of the application of independent component analysis techniques to the problem of the separation of overlapped texts in documents showing bleed-through or show-through and in palimpsests. Our approach is the first result of a study on the possibility of formulating the problem as a particular kind of blind source separation, which is a well-established discipline in signal processing, but still at the initial stage in the field of image processing and especially of document analysis. Our method involves a linear, noiseless data model where each color channel of the input image is a mixture of all the patterns to be extracted. Although we adopted this model, which is oversimplified, our first results are promising.

On the one hand, for the extraction of hidden text, the advantages of exploiting multiple channels over the analysis of single-channel images have been confirmed. Within the approaches that operate some linear combination of the input channels, our method has the additional advantage that the related coefficients do not need to be known in advance.

On the other hand, the possibility of rejecting patterns interfering with the main text will certainly be useful for improving legibility. This could mean both legibility by a human reader and by an OCR system.

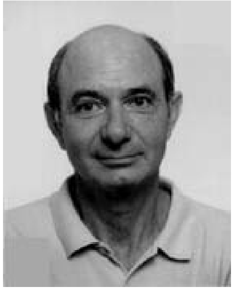
We plan to continue our study in several directions. As shown, FastICA gives some promising results but fails in many cases. This can depend on the presence of noise, on the presence of nonnegligible cross correlation between the sources, and on the nonlinearity of the mixture. All three of these possibilities rely on specific faults in the assumed data model. As far as noise is concerned, we plan to test the performance of robust, though still linear, ICA algorithms. The presence of cross correlation between different sources means that the basic ICA assumptions are not satisfied. In noiseless cases, FastICA shows some robustness to cross correlation; not so in noisy cases. We are now developing linear separation techniques that take cross correlation into account (dependent component analysis). These can be totally blind, as happens with the technique described here, or assume some prior knowledge on the structure of the mixing matrix. Moreover, the structure (e.g., autocorrelation) of the individual signals can also be taken into account to make separation more reliable. To assume a more realistic mathematical model of the data, we will also need to investigate nonlinear as well as noisy mixing models. This will imply a deep analysis of the physical degradation process. Once this model is available, a suitable separation algorithm will still need to be found. This is in general a complicated problem that is being studied very actively by several authors [13].

References

1. Amari S, Cichocki A (1998) Adaptive blind signal processing – neural network approaches. Proc IEEE 86:2026–2048
2. Attias H (1999) Independent factor analysis. Neural Comput 11:803–851
3. Avi-Itzhak HI, Diep TA, Garland H (1995) High accuracy optical character recognition using neural networks with centroid dithering. IEEE Trans Patt Anal Mach Intell 17:218–224
4. Barros AK (2000) The independence assumption: dependent component analysis. In: Girolami M (ed) Advances in independent component analysis, chap 4. Springer, Berlin Heidelberg New York, pp 63–71
5. Bell AJ, Sejnowski TJ (1995) An information maximization approach to blind separation and blind deconvolution. Neural Comput 7:1129–1159
6. Cardoso JF (1999) High-order contrasts for independent component analysis. Neural Comput 11:157–192
7. Dubois E, Pathak A (2001) Reduction of bleed-through in scanned manuscript documents. In: Proceedings of the IS&T conference on image processing, image quality, image capture systems, Montreal, 22–25 April 2001, pp 177–180
8. Easton RL (2001) Text recovery from the Archimedes Palimpsest. `http://www.cis.rit.edu/people/faculty/easton/k-12/exercise/index.htm`
9. Franke K, Köppen M (2001) A computer-based system to support forensic studies on handwritten documents. Int J Doc Anal Recog 3:218–231
10. Govindaraju V, Srihari N (1991) Separating handwritten text from overlapping nontextual contours. In: Proceedings of the international workshop on frontiers in handwriting recognition, Chateau de Bonas, France, September 1991, pp 111–119
11. Hyvärinen A (1999a) Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans Neural Netw 10:626–634
12. Hyvärinen A (1999b) Gaussian moments for noisy independent component analysis. IEEE Signal Process Lett 6:145–147
13. Hyvärinen A, Karhunen J, Oja E (2001) Independent component analysis. Wiley, New York
14. Hyvärinen A et al (2003) The FastICA package for MATLAB. `www.cis.hut.fi/projects/ica/fastica/`

15. Knuth K (1998) Bayesian source separation and localization. Proc of the SPIE: Bayesian inference for inverse problems, vol 3459, San Diego, July 1998, pp 147–158
16. Kuruoglu E, Bedini L, Paratore MT, Salerno E, Tonazzini A (2003) Source separation in astrophysical maps using independent factor analysis. Neural Netw 16(3–4):479–491
17. Lee SE, Press SJ (1998) Robustness of Bayesian factor analysis estimates. Commun Statist Theory Meth 27(8):1871–1893
18. Lee T, Lewicki M, Sejnowski T (1999) Independent component analysis using an extended infomax algorithm for mixed sub-gaussian and super-gaussian sources. Neural Comput 11:409–433
19. Leedham G, Varma S, Patankar A, Govindaraju V (2002) 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, 6–8 August 2002, pp 244–249
20. Mohammad-Djafari A (2001) A Bayesian approach to source separation. AIP Conference proceedings 567:221–244
21. Moulines E, Cardoso JF, Gassiat E (1997) Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models. In: Proceedings of the ICASSP, Munich, Germany, 21–24 April 1997, pp 3617–3620
22. Nishida H, Suzuki T (2002) Correction show-through effects in document images by multiscale analysis. In: Proceediongs of the 16th conference on pattern recognition, Quebec City, Canada, 11–15 August 2002, pp 65–68
23. Sharma G (2001) Show-through cancellation in scans of duplex printed documents. IEEE Trans Image Process 10(5):736–754
24. Tan CL, Cao R, Shen P (2002) Restoration of archival documents using a wavelet technique. IEEE Trans Patt Anal Mach Intell 24(10):1399–1404
25. Tonazzini A, Bedini L, Kuruoglu EE, Salerno E (2001) Blind separation of time-correlated sources from noisy data. Technical Report TR-42-2001 IEI-CNR, Pisa, Italy
26. Tonazzini A, Bedini L, Kuruoglu EE, Salerno E (2003) Blind separation of auto-correlated images from noisy mixtures using MRF models. In: Proceedings of the 4th international symposium on independent component analysis and blind source separation, Nara, Japan, 1–4 April 2003, pp 675–680
27. Tong L, Liu RW, Soon VC, Huang Y-F (1991) Indeterminacy and identifiability of blind identification. IEEE Trans Circuits Sys 38:499–509

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 has cooperated in special programs for basic and applied research on image processing and computer vision and is coauthor 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.



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 noninvasive diagnostic techniques. At present, his research interests are in the field of digital signal processing, image reconstruction, and neural networks applied to image processing. He is coauthor 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.

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 various responsibilities in research programs in nondestructive testing, robotics, numerical models for image reconstruction and computer vision, and neural network techniques in astrophysical imagery. At present, he is the 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.