Blind Source Separation Techniques for Detecting Hidden Texts and Textures in Document Images

Anna Tonazzini, Emanuele Salerno, Matteo Mochi, and Luigi Bedini *

Istituto di Scienza e Tecnologie dell'Informazione - CNR
Via G. Moruzzi, 1, I-56124 PISA, Italy
anna.tonazzini@isti.cnr.it

Abstract. Blind Source Separation techniques, based both on Independent Component Analysis and on second order statistics, are presented and compared for extracting partially hidden texts and textures in document images. Barely perceivable features may occur, for instance, in ancient documents previously erased and then re-written (palimpsests), or for transparency or seeping of ink from the reverse side, or from watermarks in the paper. Detecting these features can be of great importance to scholars and historians. In our approach, the document is modeled as the superposition of a number of source patterns, and a simplified linear mixture model is introduced for describing the relationship between these sources and multispectral views of the document itself. The problem of detecting the patterns that are barely perceivable in the visible color image is thus formulated as the one of separating the various patterns in the mixtures. Some examples from an extensive experimentation with real ancient documents are shown and commented.

1 Introduction

Revealing the whole contents of ancient documents is an important aid to scholars that are interested in dating the documents or establishing their origin, or reading older and historically relevant writings they may contain. However, interesting document features are often hidden or barely detectable in the original color document. Multispectral acquisitions in the non-visible range, such as the ultraviolet or the near infrared, constitute a valid help in this respect. For instance, a method to reveal paper watermarks is to record an infrared image of the paper using transmitted illumination. Nevertheless, the watermark detected with this method is usually very faint and overlapped to the contents of the paper surface. To make the watermark pattern, or whatever feature of interest, more readable and free from interferences due to overlapped patterns, an intuitive strategy is to process, for instance by arithmetic operations, multiple "views" of the document. In the case where a color scan is available, three different views can be obtained from the red, green, and blue image channels. When

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

A. Campilho, M. Kamel (Eds.): ICIAR 2004, LNCS 3212, pp. 241–248, 2004.

available, scans at non-visible wavelengths can be used alone or in conjunction with the visible ones. By processing the different color components, it is possible to extract some of the overlapped patterns, and, sometimes, even to achieve a complete separation of all them. Indeed, since all these color components contain the patterns in different "percentage", simple difference operations between the colors, after suitable regulation of the levels, can "cancel" one pattern and enhance the other. For the case of watermarks, another infrared image taken using only the reflected illumination can be used for this purpose [1]. On the other hand, some authors claim that subtracting the Green from the Red is able to reveal hidden characters in charred documents [9]. These are however empirical, document-dependent, strategies. We are looking, instead, for automatic, mathematically based, techniques that are able to enhance or even to extract the hidden features of interest from documents of any kind, without the need for adaptations to the specific problem at hand.

Our approach to this problem is to model all the document views as linear combinations of a number of independent patterns. The solution consists then in trying to invert this transformation. The overlapped patterns are usually the main foreground text, the background pattern, i.e. an image of the paper (or parchment, or whatever) support, which can contain different interfering features, such as stains, watermarks, etc., and one or more extra texts or drawings, due to previously written and then erased texts (palimpsests), seeping of ink from the reverse side (bleed-through), transparency from other pages (show-through), and other phenomena. Although our linear image model roughly simplifies the physical nature of overlapped patterns in documents [8], it has already proved to give interesting results. Indeed, this model has been proposed in [4] to extract the hidden texts from color images of palimpsests, assuming to evaluate by visual inspection the mixture coefficients. Nevertheless, in general, the mixture coefficients are not known, and the separation problem becomes one of blind source separation (BSS). It has been shown that an effective solution to BSS can be found if the source patterns are mutually independent. The independence assumption gives rise to separation techniques based on independent component analysis, or ICA [6]. Although the linear data model is somewhat simplified, and the independence assumption is not always justified, we already proposed ICA techniques for document image processing [10], and obtained good results with real manuscript documents.

In this paper we compare the performance of ICA techniques with simpler methods that only try to decorrelate the observed data. As is known, this requirement is weaker than independence, and, in principle, no source separation can be obtained by only constraining second-order statistics, at least if no additional requirement is satisfied. However, our present aim is the enhancement of the overlapped patterns, especially of those that are hidden or barely detectable, and we experimentally found that this can be achieved in most cases even by simple decorrelation. On the other hand, while the color components of an image are usually spatially correlated, the individual classes or patterns that compose the image are at least less correlated. Thus, decorrelating the color components gives a different representation where the now uncorrelated components of the image could coincide with the single classes. Furthermore, the second-order approach

is always less expensive than ICA algorithms, and due to the poor modeling or to the lack of independence of the patterns, the results from decorrelation can also be better than the ones from ICA.

2 Formulation of the Problem

Let us assume that each pixel (of index t in a total of T) of a multispectral scan of a document has a vector value $\mathbf{x}(t)$ of N components. Similarly, let us assume to have M superimposed sources represented, at each pixel t, by the vector $\mathbf{s}(t)$. Since we consider images of documents containing homogeneous texts or drawings, we can also reasonably assume that the color of each source is almost uniform, i.e., we will have mean reflectance indices A_{ij} for the i-th source at the j-th wavelenght. Thus, we will have a collection of T samples from a random N-vector $\mathbf{x}$, which is generated by linearly and istantaneously mixing the components of a random M-vector $\mathbf{s}$ through an $N \times M$ mixing matrix A:

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

where the source functions $s_i(t)$, $i = 1, 2, ..., M$ denote the "quantity" of the M patterns that concur to form the color at point t. Estimating $\mathbf{s}(t)$ and A from knowledge of $\mathbf{x}(t)$ is called a problem of blind source separation (BSS). In this application, we assume that noise and blur can be neglected. When only the visible color scan is available, vector $\mathbf{x}(t)$ has dimension $N = 3$ (it is composed by the red, green, and blue channels). However, most documents can be seen as the superposition of only three ($M = 3$) different sources, or classes, that we will call "background", "main text" and "interfering texture". In general, by using multispectral/hyperspectral sensors, the "color" vector can assume a dimension greater than 3. Likewise, we can also have $M > 3$ if additional patterns are present in the original document. In this paper, we only consider the case $M = N$, that is, same number of sources as of observations, although in principle there is no difference with the general case.

It is easy to see that this model does not perfectly account for the phenomenon of interfering texts in documents, which derives from complicated chemical 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 color is not the vector sum of the colors of the two components, but it is likely to be some nonlinear combination of them. In [8], a nonlinear model is derived even for the simpler phenomenon of show-through. However, although the linear model is only a rough approximation, it has demonstrated its usefulness in different applications, as already mentioned above [4] [10].

3 The Proposed Solutions: ICA, PCA, and Whitening

When no additional assumption is made, problem (1) 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 introduction, this is the ICA approach [6]. If the prior distribution for each source is known, independence is equivalent to assume a factorized form for the joint prior distribution of $\mathbf{s}$:

$$P(\mathbf{s}(t)) = \prod_{i=1}^{N} P_i(s_i(t)) \qquad \forall t \tag{2}$$

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, $W = (\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_N)^T$, such that, when applied to the data $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N)$, produces the set of vectors $\mathbf{w}_i^T\mathbf{x}$ 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:

$$\hat{W} = arg \max_{W} \sum_{t} \sum_{i} logP_i(\mathbf{w}_i^T\mathbf{x}(t)) + Tlog|det(W)| \tag{3}$$

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

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, no matter of the true source distributions, which, on the other hand, are usually unknown [2].

Although we already obtained some promising result by this approach [10], there is no apparent physical reason why our original sources should be mutually independent, so, even if the data model (1) was correct, the ICA principle is not assured to be able to separate the different classes. 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. To avoid cumbersome notation, and without loss of generality, let us assume to have zero-mean data vectors. We thus seek for a linear transformation $\mathbf{y}(t) = W\mathbf{x}(t)$ such that $< y_i y_j >= 0$, $\forall i, j = 1, \ldots, M, \;\; i \neq j$, where W is generally an $M \times N$ matrix and the notation $< \cdot >$ means expectation. In other words, the components of the transformed data vector $\mathbf{y}$ are 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, although the linear model is not fully justified. In [7], 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 give orthogonal output vectors, one of which has maximum variance. 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 [3]. Our data covariance matrix is the $N \times N$ matrix:

$$R_{\mathbf{xx}} =< \mathbf{x}\mathbf{x}^T >\approx \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}(t)\mathbf{x}^T(t) \tag{4}$$

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

$$R_{\mathbf{yy}} =< W\mathbf{x}\mathbf{x}^T W^T >= WR_{\mathbf{xx}}W^T \tag{5}$$

To obtain orthogonal $\mathbf{y}$, $R_{\mathbf{yy}}$ 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 decreasing order. Now, it is easy to verify that all of the following choices for W yield a diagonal $R_{\mathbf{yy}}$:

$$W_o = V_{\mathbf{x}}^T \tag{6}$$

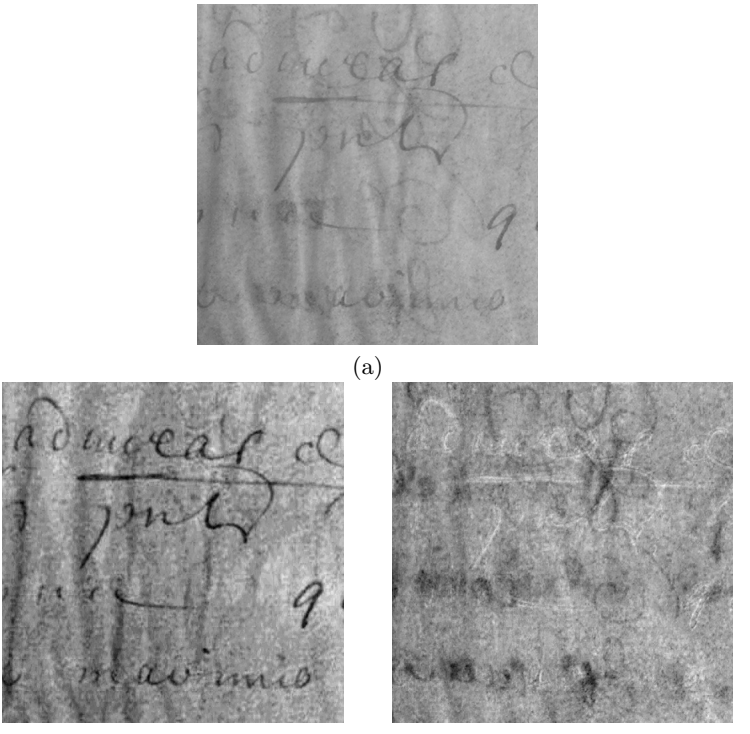
$$W_w = \Lambda_{\mathbf{x}}^{-\frac{1}{2}} V_{\mathbf{x}}^T \tag{7}$$

$$W_s = V_{\mathbf{x}}\Lambda_{\mathbf{x}}^{-\frac{1}{2}} V_{\mathbf{x}}^T \tag{8}$$

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 [3]. 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 (8), we have a whitening matrix with the further property of being symmetric. In [3], it is observed that application of matrix W_s is equivalent to ICA when matrix A is symmetric. In general, ICA applies a further rotation to the output vectors, based on higher-order statistics.

4 Experimental Results and Concluding Remarks

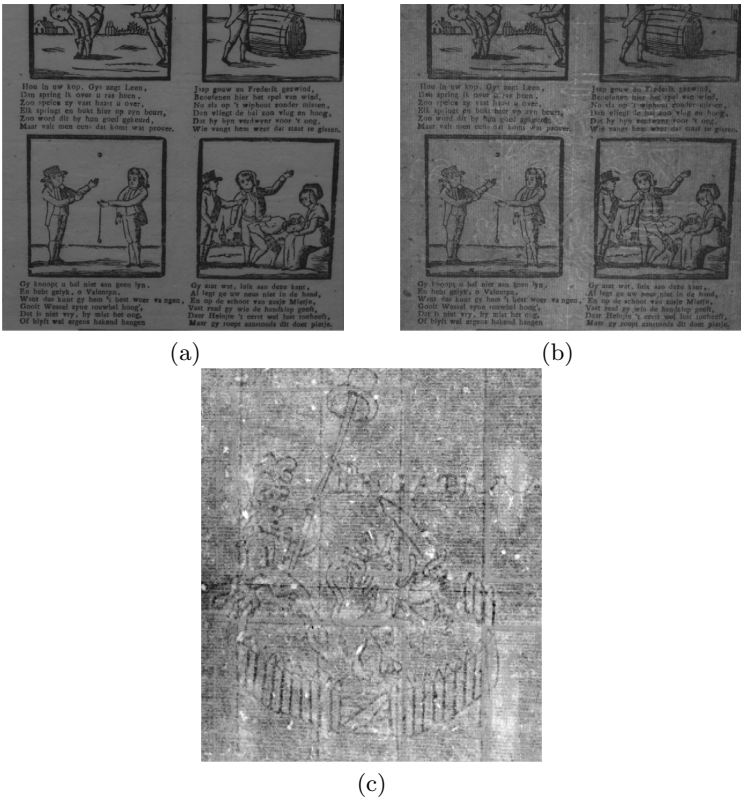
Our experimental work has consisted in applying the above matrices to typical images of ancient documents, with the aim at emphasizing the document hidden features in the whitened vectors. For each test image, the results are of course different for different whitening matrices. However, it is interesting to note that the symmetric whitening matrix often performs better than ICA, and, in some cases, it can also achieve a separation of the different components, which is the final aim of BSS. Here, we show some examples from our extensive experimentation. The first example (Figure 1) describes the processing of an ancient manuscript which presents three overlapped patterns: a main text, an underwriting barely visible in the original image, and a noisy background with



(a) (b) (c)

Fig. 1. Full separation with symmetric orthogonalization: (a) grayscale representation of the color scan of an ancient manuscript containing a partially hidden text; (b) first symmetric orthogonalization output from the RGB components of the color image; (c) second symmetric orthogonalization output from the same data set.

significant paper folds. We compared the results of the FastICA algorithm [5] [10], the PCA, and the symmetric whitening, all applied to the RGB channels, and found that full separation and enhancement of the three classes is obtained by the symmetric orthogonalization only. ICA failure might depend, in this case, on the data model inaccuracy and/or the lack of mutual independence of the classes. In Figure 2, we report another example where a paper watermark pattern is detected and extracted. In this case, we assume the document as constituted of two only classes: the foreground pattern, with drawings and text, and the background pattern with the watermark, so that two only views are needed. We used two infrared acquisitions, the first taken under front illumination, the second taken with illumination from the back. In this case a good extraction is achieved by using all the three methods proposed. However, the best one is obtained with FastICA. Finally, Figure 3 shows a last example of extraction of a faint underlying pattern, using the RGB components. In this case, all the three proposed methods performed similarly.



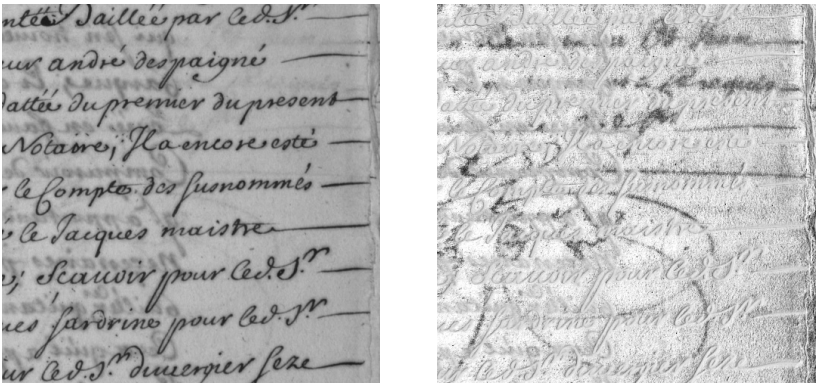
(a) (b)

(c)

Fig. 2. Watermark detection: (a) infrared front view; (b) back illumination infrared view; (c) one FastICA output.

These experiments confirmed our initial intuition about the validity of BSS techniques for enhancing and separating the various features that appear as overlapped in many ancient documents. No conclusions can be instead drawn about the superiority of one method over the others for all documents. We can only say that, when the main goal is to enhance partially hidden features, at least one of the three methods proposed always succeeded in reaching the scope in all our experiments. The advantages of these techniques are that they are quite simple and fast, and do not require reverse side scans or image registration. Our research programs for the near future regard the development of more accurate numerical models for the phenomenon of pattern overlapping in documents.

Acknowledgements. We would like to thank the Isyreadet partners for providing the original document images. Composition of the Isyreadet consortium: TEA SAS (Catanzaro, Italy), Art Innovation (Oldenzaal, The Netherlands), Art Conservation (Vlaardingen, The Netherlands), Transmedia (Swansea, UK), Ate-



(a) (b)

Fig. 3. Detection of an underlying pattern: (a) grayscale version of the original color document; (b) underlying pattern detected by symmetric orthogonalization.

lier Quillet (Loix, France), Acciss Bretagne (Plouzane, France), ENST (Brest, France), CNR-ISTI (Pisa, Italy), CNR-IPCF (Pisa, Italy).

References

1. http://www.art-innovation.nl/
2. Bell AJ, Sejnowski TJ : Neural Computation (1995) 7:1129–1159
3. Cichocki, A., Amari, S.-I.: Adaptive Blind Signal and Image Processing (2002) Wiley, New York.
4. Easton, R.L.: http://www.cis.rit.edu/people/faculty/easton/k-12/index.htm
5. Hyvärinen, A., Oja, E.: Neural Networks (2000) 13:411–430.
6. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis (2001) John Wiley, New York.
7. Ohta, Y., Kanade, T., Sakai, T.: Computer Graphics, Vision, and Image Processing (1980) 13:222–241.
8. Sharma, G.: IEEE Trans. Image Processing (2001) 10:736–754.
9. R. Swift: http://www.cis.rit.edu/research/thesis/bs/2001/swift/thesis.html.
10. Tonazzini, A., Bedini, L., Salerno, E.: Int. J. Document Analysis and Recognition (2004) in press.