

RESEARCH

Open Access

Restoration of recto–verso colour documents using correlated component analysis

Anna Tonazzini* and Luigi Bedini

Abstract

In this article, we consider the problem of removing see-through interferences from pairs of recto–verso documents acquired either in grayscale or RGB modality. The see-through effect is a typical degradation of historical and archival documents or manuscripts, and is caused by transparency or seeping of ink from the reverse side of the page. We formulate the problem as one of separating two individual texts, overlapped in the recto and verso maps of the colour channels through a linear convolutional mixing operator, where the mixing coefficients are unknown, while the blur kernels are assumed known *a priori* or estimated off-line. We exploit statistical techniques of blind source separation to estimate both the unknown model parameters and the ideal, uncorrupted images of the two document sides. We show that recently proposed correlated component analysis techniques overcome the already satisfactory performance of independent component analysis techniques and colour decorrelation, when the two texts are even sensibly correlated.

Keywords: Document restoration and analysis, See-through interference, Blind source separation, Correlated component analysis

1. Introduction

One of the most common degradations affecting historical and/or archival documents that are written or printed on both sides of the page is see-through, that is an undesired pattern in the background, caused by the text in the reverse side of the page. Such distortion can significantly degrade the readability of the document or make difficult the automatic analysis of its content. See-through is usually distinguished in bleed-through or show-through. Bleed-through is intrinsic to the analogue document, especially in the ancient ones, for effect of the paper thinness or chemical reactions of the ink, e.g. due to humidity. In these situations, the ink in the reverse side might penetrate through the paper fibres, thus emerging in the front side. The digital acquisition of document images through scanners can introduce show-through even in well-preserved documents, for effect of light transmission through the paper, or can worsen the already present degradation.

Several approaches for see-through reduction have been investigated, mainly for grayscale documents, and exploiting the availability of pre-registered scans of both sides (*recto* and *verso*).

In [1], a wavelet technique is applied for iteratively enhancing the foreground strokes and smearing the interfering strokes. In [2], a variational approach, based on nonlinear diffusion and wavelet transforms, has been proposed to model and then remove see-through from either single-sided or double-sided grayscale documents. In [3,4], steps of segmentation to identify the see-through areas are followed by inpainting of estimated pure background areas. Segmentation-classification is the basis also for the methods derived in [5,6].

In [7], the physical model of the show-through in modern scanners is first simplified for deriving a tractable mathematical nonlinear convolutional mixing model. This model is further approximated for decoupling the two recto and verso equations, in order to design an adaptive linear filter that is very effective in correcting a mild show-through.

Recently, the interest in applying blind source separation (BSS) algorithms for solving this problem has increased noticeably. The appearance of the degraded recto and verso scans is first modelled as a parametric superimposition of the uncorrupted recto and verso images, and then a separation algorithm is used to estimate both the mixing parameters and the ideal front and back side images (*sources*). The assumption of a linear

* Correspondence: anna.tonazzini@isti.cnr.it
National Research Council of Italy, Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", Via G. Moruzzi, 1, Pisa, Italy

instantaneous mixing model has led to BSS algorithms such as independent component analysis (ICA) [8,9] or non-negative matrix factorization (NMF) [10]. Some works have also addressed more realistic nonlinear and/or convolutional mixing models, and separation algorithms based on image regularization [11-14].

Specifically, within the linear instantaneous mixing model, in [8], by exploiting reasonable constraints of symmetry for the data model, ICA has been shown to be equivalent to symmetric whitening, and a fast separation algorithm has been proposed for grayscale recto-verso pairs affected by show-through.

NMF has been proposed for minimizing an energy function composed of a data term plus a regularization term compensating for the apparent nonlinearity of the show-through phenomenon, in correspondence of the occlusions between the two texts [10]. Ophir and Malah [11] propose a convolutional BSS formulation, which accounts also for a nonlinearity of the show-through effect, assumed known and derived from [7]. The solution is based on the total variation stabilizer for the ideal images. In [13], a maximum likelihood approach is proposed for two nonlinear mixtures of two texts, where the show-through nonlinearity is approximated as quadratic, and a blur kernel on the interfering pattern is accounted for and estimated.

In spite of the assumption of a linear instantaneous mixing model, ICA has proven to be very cost-effective and versatile for application to different typologies of data and several instances of document restoration and analysis. For example, it can easily be extended to the analysis of multispectral scans of a single-sided document containing multiple information layers [15], or when the data are the RGB recto and verso scans of a colour document [9]. This latter case is particularly interesting if the aim is to produce a restored visible document that, while cleansed of the unwanted interferences, maintains its useful features, e.g. the original colour, as much as possible.

Nevertheless, ICA assumes independence or at least uncorrelation of the individual sources, that is, it forces uncorrelation between the recto and verso ideal images. If uncorrelation can be expected at the high frequencies, this is not realistic at the low frequencies, where we experimentally verified a significant cross-correlation between recto and verso, probably due to the large background areas.

In addition, efficient ICA algorithms have mainly been designed for instantaneous mixtures, whereas convolutional mixtures would be more appropriate to model the blur affecting the text emerging from the back side, which appears smeared for effect of light diffusion through the paper, or ink absorption from the paper fibres. Although ICA solutions to the convolutional mixture case have been proposed (see [16]), the problem has not fully solved, yet.

In this article, we consider the problem of removing see-through interferences from pairs of recto-verso documents, acquired either as grayscale images or RGB images. We still adopt a linear mixing data model, but remove the instantaneous assumption to account for blur on the source images. Hence, our model is linear convolutional. We show that the use of recently proposed correlated component analysis (CCA) techniques [16], based on second-order statistics and working in the Fourier domain, allows both to remove the uncorrelation assumption and to easily manage the convolutional nature of the data model. Our method is based on the joint estimation of the mixing parameters and the source spectra and cross-spectra, performed by alternating minimization of a suitable cost function, with respect to the two sets of variables. Once the estimates are available, the individual sources can be recovered either by a simple inverse filter or, when noise is present, by Wiener filtering, since the estimated spectra can effectively be exploited. At present, we assume that the blur kernels are known *a priori*, for example when the degradation is mainly caused by the scanning process, and the technical characteristics of the equipment used are available, or estimated off-line. In this latter case, the simple selection of a pure see-through area in one of the two sides can efficiently serve to the scope.

The method is very fast, and the experimental results performed show that it significantly outperforms instantaneous ICA, by permitting separation to be achieved also when the individual sources are largely correlated. This is especially true when the patterns that interfere from a side to the other of the page are sensibly blurred.

The remainder of the article is organized as follows. In Section 2, we describe the mathematical model that we assume herein for the see-through phenomenon in a pair of recto-verso colour images, and discuss our previously proposed solution [8,9] that, neglecting the blur on the sources and assuming their statistical independence, makes use of ICA techniques. In Section 3, we describe the mathematical details of the method proposed in this article that, based on a technique of CCA, is able to account for both blur on the source and noise on the data, and allows for relaxing the assumption of full uncorrelation between the sources. Experiments on both synthetic and real documents are discussed in Section 4 and, finally, Section 5 summarizes the main achievements of this research, discusses advantages and limitations, and presents possible future developments.

2. Data model and problem formulation

We consider the case to have at our disposal the RGB scans of the recto and verso sides of a document page affected by see-through, the case of grayscale scans being a mere sub-case. The general model we adopt for our observations is the following:

$$\begin{aligned} x_r^C(t) &= a_{rr}^C h_{rr}^C(t) \oplus s_r^C(t) + a_{rv}^C h_{rv}^C(t) \oplus s_v^C(t) + \mu_r^C(t) \\ x_v^C(t) &= a_{vr}^C h_{vr}^C(t) \oplus s_r^C(t) + a_{vv}^C h_{vv}^C(t) \oplus s_v^C(t) + \mu_v^C(t) \end{aligned} \quad (1)$$

where $x_r^C(t)$ and $x_v^C(t)$ are the recto and (flipped) verso images at pixel t and at channel C , with $C = (R, G, B)$. Analogously, $s_r^C(t)$ and $s_v^C(t)$ are the clean recto and verso text patterns at pixel t , channel C . The positive elements a_{nm}^C , $n = (r, v)$ and $m = (r, v)$, of the so-called mixing matrix A , represent the unknown percentage of ink intensity attenuation of the two texts in the two sides, due to ageing factors, or transparency of the paper or ink seeping from the back to the front page. Analogously, functions h_{nm}^C , $n = (r, v)$ and $m = (r, v)$, of unitary sum, represent blur kernels explaining for the smearing of the ink, due to the same factors, and with symbol \oplus we indicate the convolution operator. Finally, μ_n^C , $n = (r, v)$, is the channel noise.

Restoring the degraded recto-verso images at hand entails solving the system in Equation (1) for ink attenuation indices, blur kernels and sources. Once estimated, the set of sources (s_n^C , $n = r, v$; $C = R, G, B$) can be arranged as (s_r^R, s_r^G, s_r^B) and (s_v^R, s_v^G, s_v^B) to give the restored RGB recto and verso images.

Some physical properties of the see-through phenomenon can be exploited to simplify the problem in Equation (1). In fact, we can observe that the text interfering from the back side is always attenuated, whereas the same text is not, or much less, in the side where it was originally written. Hence, the ink attenuation indices and the blur kernels are expected to be different, for each source, in the two observations, and for the two sources in a same observation. Let us consider that, at least in an idealized setting, the two sides have been written with the same ink, same pressure and at two close moments. Then, it is reasonable to assume that, at each channel, the attenuation of the see-through text in the two sides is the same, i.e. $a_{rv}^C = a_{vr}^C$, as well as the ink smearing, i.e. $h_{rv}^C = h_{vr}^C$. For similar considerations, it is also expected that $a_{rr}^C = a_{vv}^C$ and $h_{rr}^C = h_{vv}^C$. Furthermore, as already mentioned, the ink intensity of the front text in the recto side should be higher than that of the see-through text, i.e. $a_{rr}^C > a_{rv}^C$, with the same relationship holding, reversed, in the verso side. Within these assumptions, system of Equation (1) results to be symmetric.

For the case of grayscale recto-verso pairs, and neglecting both blur and noise, in [8] the restoration problem has successfully been solved assuming statistical independence of the sources, and employing a very fast and fully blind symmetric whitening of the data, which is equivalent to ICA for symmetric mixing matrices. However, the mere application of ICA to the RGB recto-verso case is not suitable and even wrong. Indeed, here the mutual independence of the overall set of sources cannot be assumed, since the different colours of a same text pattern are certainly highly correlated.

A useful observation is that the 6×6 system of equations is separable into three independent 2×2 symmetric problems, which can be then solved separately. In this case, employing ICA to solve each subsystem only entails assuming independence of the recto and verso text at each individual channel, as done for the grayscale case, and no unrealistic uncorrelation is assumed among the various colour maps of the recto (verso) text. This allows reducing the interferences while preserving the original colour of the RGB recto and verso images [9]. However, some residual see-through interference usually remain in the reconstructions. Indeed, if no blur is accounted for, two homologous patterns in the two different sides cannot match exactly; on the other side, this could also indicate that the recto and verso text are actually correlated at each channel.

In the following section, we will show that CCA techniques, recently applied with success to solve different image processing problems [17], permit to account for blur and to fully relax the uncorrelation assumption among the recto and verso texts.

3. Solution through CCA

The essence of the CCA technique is to admit non-zero, although unknown, auto- and cross-correlations for the sources, which must jointly be estimated with the mixing parameters. In our case, this complex joint estimation can take advantage from the explicit exploitation of the symmetries of both the mixing operator and the source covariance matrix. In order to easily enforce suitable constraints that are available on the source spectra the problem is also transformed from the space domain to the Fourier domain.

By further assuming to neglect, at this stage, the attenuation and blurring of the recto (verso) text in the recto (verso) side, at each channel C the model we consider in the space domain is given by following 2×2 system:

$$\begin{aligned} x_r^C(t) &= s_r^C(t) + a^C h^C(t) \oplus s_v^C(t) + \mu_r^C(t) \\ x_v^C(t) &= a^C h^C(t) \oplus s_r^C(t) + s_v^C(t) + \mu_v^C(t) \end{aligned} \quad (2)$$

With this model, our scope reduces to the removal of the see-through interferences only, while the restoration of other degradations undergone by the text in the side where it was originally written are left to possible subsequent processing. However, it is to be noted that, often, archivists and scholars do prefer a restoration intervention limited to the artefact removal, which does not alter the original, aged appearance of the document itself. The proposed method could take into account for additive noise [18], restoring sources with a better SNR. However, since removing additive noise could alter the appearance of the document, as a first approach we decided to neglect noise.

As a consequence, in the Fourier domain model of Equation (2) becomes

$$\begin{aligned} X_r^C(\omega) &= S_r^C(\omega) + a^C H^C(\omega) S_v^C(\omega) \\ X_v^C(\omega) &= a^C H^C(\omega) S_r^C(\omega) + S_v^C(\omega) \end{aligned} \quad (3)$$

In matrix form, the complete system to be solved is given by

$$\begin{bmatrix} X_r^R(\omega) \\ X_v^R(\omega) \\ X_r^G(\omega) \\ X_v^G(\omega) \\ X_r^B(\omega) \\ X_v^B(\omega) \end{bmatrix} = \begin{bmatrix} 1 & a^R H^R(\omega) & 0 & 0 & 0 & 0 \\ a^R H^R(\omega) & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & a^G H^G(\omega) & 0 & 0 \\ 0 & 0 & a^G H^G(\omega) & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & a^B H^B(\omega) \\ 0 & 0 & 0 & 0 & a^B H^B(\omega) & 1 \end{bmatrix} \begin{bmatrix} S_r^R(\omega) \\ S_v^R(\omega) \\ S_r^G(\omega) \\ S_v^G(\omega) \\ S_r^B(\omega) \\ S_v^B(\omega) \end{bmatrix} \quad (4)$$

The three problems to be separately solved are thus

$$\begin{aligned} X_C(\omega) &= \begin{bmatrix} X_r^C(\omega) \\ X_v^C(\omega) \end{bmatrix} \\ &= \begin{bmatrix} 1 & a^C H^C(\omega) \\ a^C H^C(\omega) & 1 \end{bmatrix} \begin{bmatrix} S_r^C(\omega) \\ S_v^C(\omega) \end{bmatrix} \\ &= a^C H_C(\omega) S_C(\omega) C = R, G, B \end{aligned} \quad (5)$$

where $H_C(\omega)$ is the 2×2 matrix defined as follows

$$H_C(\omega) = \begin{bmatrix} \frac{1}{a^C} & H^C(\omega) \\ H^C(\omega) & \frac{1}{a^C} \end{bmatrix}$$

and $S_C(\omega)^T = [S_r^C(\omega), S_v^C(\omega)]$.

Without losing generality, in the following we assume the blur kernels h^C to be circularly symmetric. In a first approach, we assume them Gaussian with known variance. This assumption greatly simplifies the estimation process. It is also to be noted that, in a large part of documents, the off-line estimation of the blur kernel does not present significant difficulty.

Let us partition the Fourier domain into l^{\max} annular bins of sufficiently small width $\Delta\omega$. Let Z_l be the l th annular bin with internal radius $(l - 1) \Delta\omega$ and external radius $l\Delta\omega$. The circular spectra of the data and the sources are defined as follows:

$$S_{X_C}(l) = \frac{1}{N_l} \sum_{\omega \in Z_l} X_C(\omega) X_C(\omega)^* \quad (6)$$

and

$$S_{S_C}(l) = \frac{1}{N_l} \sum_{\omega \in Z_l} S_C(\omega) S_C(\omega)^* \quad (7)$$

respectively, where N_l is the number of Fourier modes contained in Z_l and $*$ indicates the conjugate transpose. It is easy to verify that, if data and sources are real, the circular spectra are real as well. In view of Equation (5), Equation (6) can be rewritten as follows:

$$S_{X_C}(l) = \frac{1}{N_l} \sum_{\omega \in Z_l} a^C H_C(\omega) S_C(\omega) S_C(\omega)^* a^C H_C^*(\omega) \quad (8)$$

From the circular symmetry, it is $H_C^*(\omega) = H_C^T(\omega)$.

Letting $B_C(\omega) = a^C H_C(\omega)$, and provided that the width of the annular bins in the Fourier domain are chosen sufficiently small to assume the blur kernels as constant within each Z_l , it is

$$\begin{aligned} S_{X_C}(l) &\approx B_C(l) \left[\frac{1}{N_l} \sum_{\omega \in Z_l} S_C(\omega) S_C(\omega)^* \right] B_C^T(l) \\ &= B_C(l) S_{S_C}(l) B_C^T(l) \end{aligned} \quad (9)$$

From several simulations we found that, for the specific problem at hand, by using either Equation (8) or (9) almost identical spectra can be obtained. Thus, Equation (9) establishes the existing relationship, $\forall l$, between the circular cross-spectra of the data and the sources. Note that, $\forall l$, $S_{X_C}(l)$ and $S_{S_C}(l)$ are 2×2 real and symmetric matrices. Hence, the number of independent equations is 3.

It is well known that, $\forall l$, Equation (9) can be transformed in the following way:

$$d(l) = [B_C(l) \otimes B_C(l)] g(l) \quad (10)$$

where \otimes indicates the Kronecker product, $d(l)$ and $g(l)$ are the lexicographic forms of $S_{X_C}(l)$ and $S_{S_C}(l)$, respectively.

Calling $D_C(l)$ the 4×4 matrix $B_C(l) \otimes B_C(l)$, let us introduce now the following cost function:

$$E = \sum_{l=1}^{l^{\max}} d(l) - D_C(l) g(l)^2 + \lambda \Phi[g(1), \dots, g(l^{\max})] \quad (11)$$

with λ being a positive regularization parameter and Φ a certain function expressing regularization constraints on the circular cross-spectra of the sources. Since we assume herein to know the blur kernels, cost function E

has a^C and $g(l)$, $l = 1, \dots, l_{\max}$ as unknowns. The solution of the problem can then be computed as follows:

$$\begin{aligned} & (\hat{a}^C, \hat{g}(l), \quad l = 1, \dots, l_{\max}) \\ & = \arg \min_{a^C, g(l), l=1, \dots, l_{\max}} E \end{aligned} \quad (12)$$

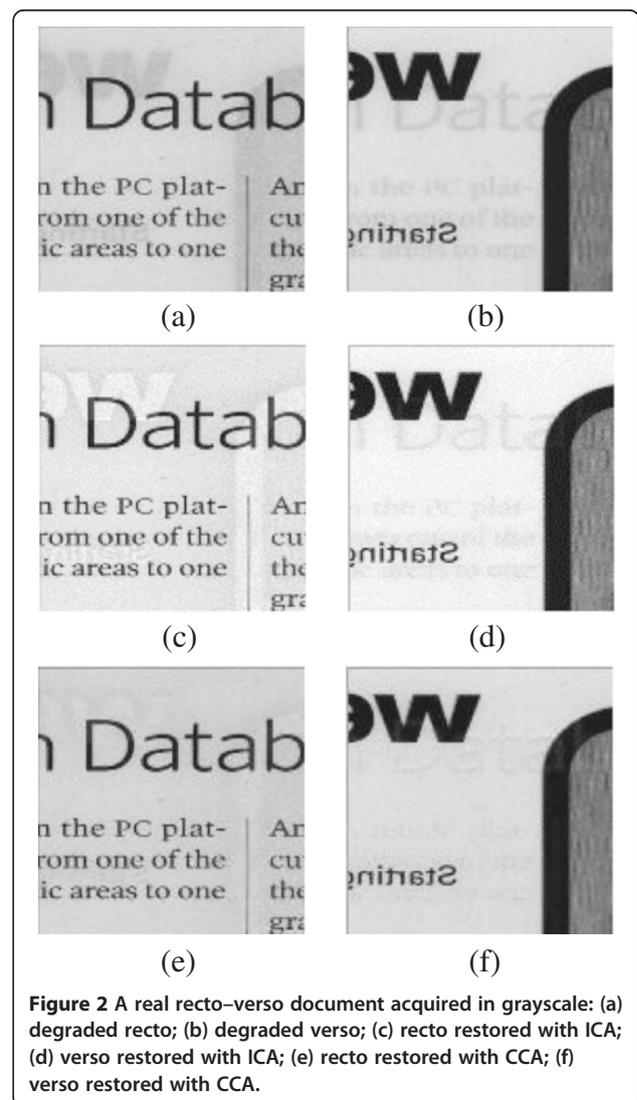
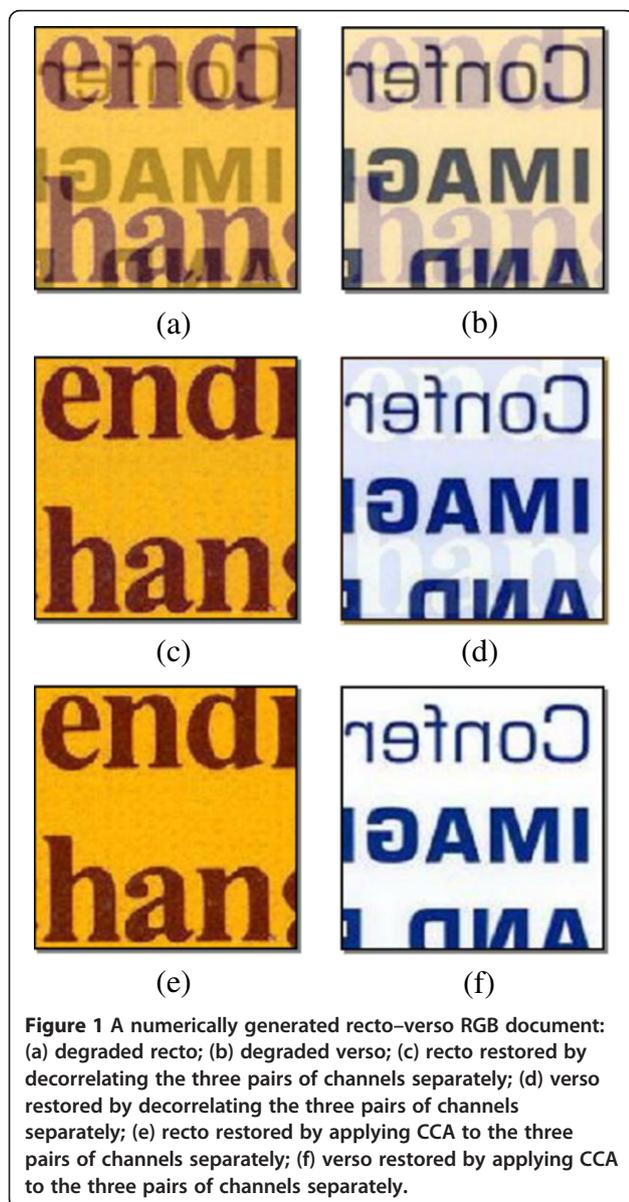
which can be simplified by considering that

$$\min_{a^C, g(l), l=1, \dots, l_{\max}} E = \min_{a^C} \min_{g(l), l=1, \dots, l_{\max}} E \quad (13)$$

and then performing alternate minimizations with respect to the mixing parameters and the spectra. By choosing suitable regularization functions, the minimizer with respect to $g(l)$, $l = 1, \dots, l_{\max}$, can be computed in analytical form with a very low computational load,

while the minimizer with respect to a^C can be computed iteratively or employing stochastic algorithms. Functions Φ are chosen in such a way to enforce a global regularization constraint on the cross-spectra. Global smoothness and minimum energy are the constraints most frequently used for this purpose. We implemented both, and verified substantially equivalent performances, being the minimum energy slightly simpler. Thus, the results presented in this article have been obtained by enforcing a constraint of minimum energy on the cross-spectra. The minimization with respect to the model parameters requires, in general, the use of stochastic algorithms, of the type of simulated annealing. In our case, since we need to estimate a single parameter, we employed a faster iterative technique.

By analysing several documents affected by see-through, we found that at the high frequencies the cross-spectra go quickly to zero, which means that recto and verso are uncorrelated beyond a certain frequency l_1 that can be



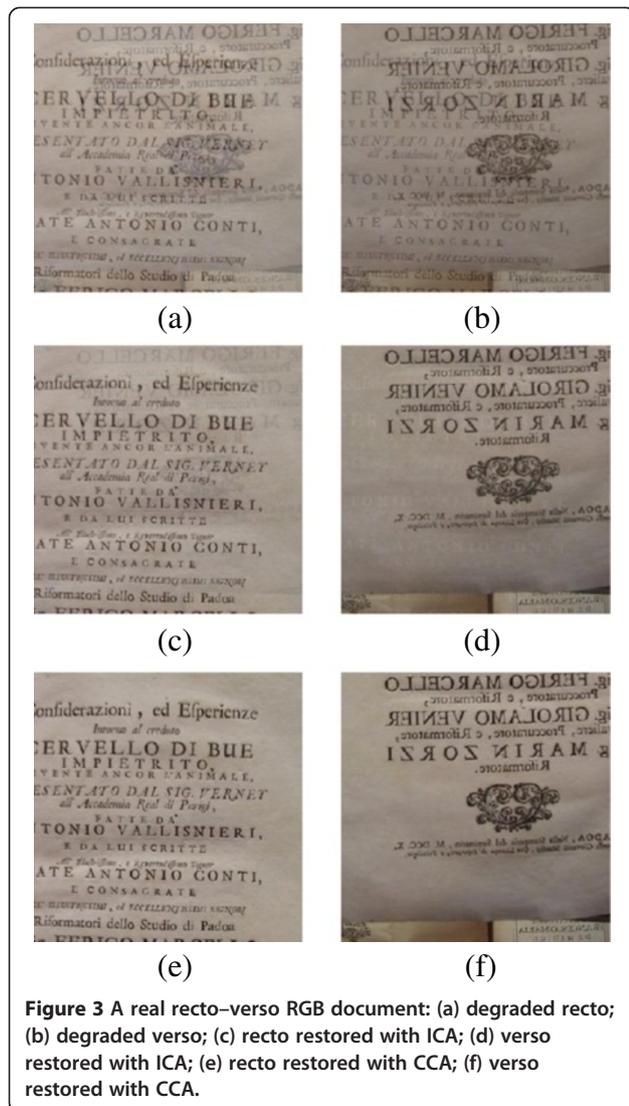


Figure 3 A real recto-verso RGB document: (a) degraded recto; (b) degraded verso; (c) recto restored with ICA; (d) verso restored with ICA; (e) recto restored with CCA; (f) verso restored with CCA.

determined experimentally. Specifically, we found that, in most cases, by taking $\Delta\omega = 1.5$, it is $l_1 = 10$. Hence, the solution to problem of Equation (12) can be found by enforcing the further constraint of null cross-spectra for $l > 10$.

Once the estimates are available, the individual sources can be recovered, at each Fourier mode, by inverse filtering. When noise is present, Wiener filtering is employed instead, since the estimated spectra can effectively be exploited. In both cases, the method is very fast. In particular, its complexity is comparable to that of the FastICA algorithm [19], and much lower of that of methods based on nonlinear data models, such as the one proposed in [11].

4. Experimental results

In this section, we will show the performance of the CCA technique described above compared with that of the ICA technique, this latter implemented either through symmetric whitening or through the FastICA algorithm [19].

In a first set of experiments, described in Section 4.1, we processed a variety of both grayscale and RGB recto-verso pairs, either affected by show-through or bleed-through, including some among the ones that are mostly tested in the literature on the subject.

During the revision process of this article, we became aware of the existence of a recently published online database of high-resolution grayscale images of ancient documents affected by bleed-through [20]. This database has been created by the project Irish Script on Screen (ISOS) of the School of Celtic Studies, Dublin Institute for Advanced Studies, in conjunction with the SIGMEDIA group of the Department of Electrical and Electronic Engineering at Trinity College Dublin.

Hence, in Section 4.2, we analyse the results of applying our techniques to those images.

4.1. Test images: miscellaneous

A first synthetic experiment, on a recto-verso pair built numerically, is shown in Figure 1. This example has the aim to quantitatively analyse the performance of our method, for the general case of RGB scans. However, to let ICA working at its best, neither blur nor noise has been added to the data. Figure 1a,b shows the recto and verso images, respectively, obtained with a linear instantaneous

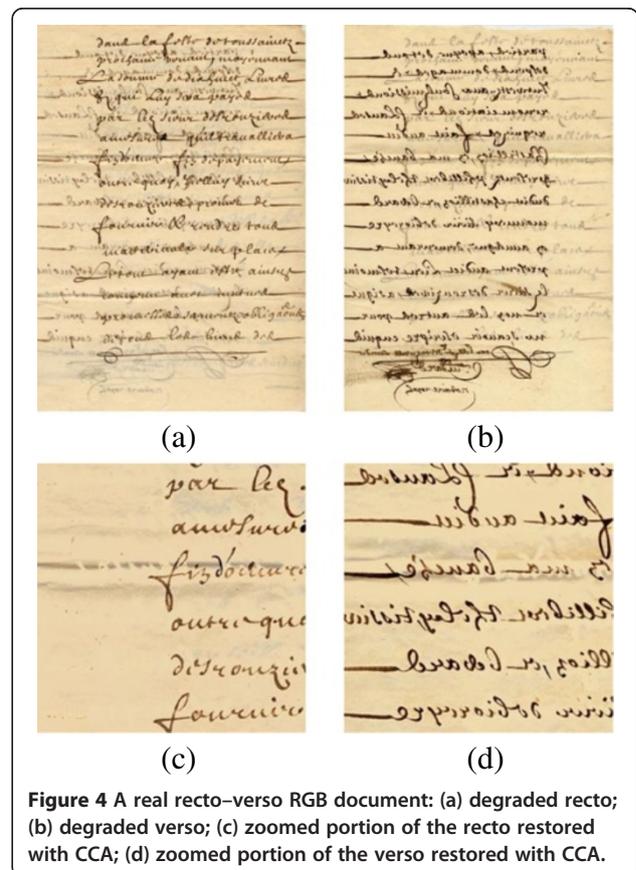


Figure 4 A real recto-verso RGB document: (a) degraded recto; (b) degraded verso; (c) zoomed portion of the recto restored with CCA; (d) zoomed portion of the verso restored with CCA.

mixing model, by using $a^R = 0.8$, $a^G = 0.4$ and $a^B = 0.2$, respectively. Figure 1c,d shows the solution that can be obtained when the three pairs of recto and verso images at the three R , G and B channels are decorrelated individually and separately via symmetric whitening. This strategy only enforces uncorrelation of the recto text and the verso text at each channel. However, note the slight residual interference in the second output, clearly indicating that the recto text and the verso text are actually correlated. The imperfect separation affects also the reconstructed colour. Finally, Figure 1e,f shows the perfect reconstruction that can be obtained by employing the CCA strategy described in the previous section. The mixing parameters estimated with $l_1 = 10$ were $a^R = 0.82$, $a^G = 0.43$ and $a^B = 0.21$. The estimated cross-spectra highlighted a significant correlation among recto and verso text patterns. In the space domain, the normalized estimated covariance matrix was

$$E \left\{ \begin{bmatrix} s_r^R \\ s_r^G \\ s_r^B \\ s_v^R \\ s_v^G \\ s_v^B \end{bmatrix} \begin{bmatrix} s_r^R & s_r^G & s_r^B & s_v^R & s_v^G & s_v^B \end{bmatrix} \right\} = \begin{bmatrix} 1.0000 & 0.9619 & 0.0600 & \mathbf{0.1643} & \mathbf{0.2056} & \mathbf{0.3086} \\ 0.9619 & 1.0000 & 0.2207 & \mathbf{0.1009} & \mathbf{0.1291} & \mathbf{0.1922} \\ 0.0600 & 0.2207 & 1.0000 & \mathbf{0.0370} & \mathbf{0.0245} & \mathbf{-0.0122} \\ \mathbf{0.1643} & \mathbf{0.1009} & \mathbf{0.0370} & 1.0000 & 0.9949 & 0.9440 \\ \mathbf{0.2056} & \mathbf{0.1291} & \mathbf{0.0245} & 0.9949 & 1.0000 & 0.9652 \\ \mathbf{0.3086} & \mathbf{0.1922} & \mathbf{-0.0122} & 0.9440 & 0.9652 & 1.0000 \end{bmatrix}$$

Figure 2 shows the results of the CCA method, in the case of a real grayscale document. From the original scans shown in Figure 2a,b, it is apparent that the show-through patterns are blurred. Since our method, at present, does not foresee the estimation of the blur kernels, we must provide an estimate obtained off-line. A simple, heuristic way to estimate the blur of a show-through pattern, when the recto and verso scans are available, is described in [14], and is based on reversing the principle of image deconvolution. In brief, rather than estimating the original image knowing the degraded image and the point spread function (PSF) as usually done, we estimate the PSF knowing the degraded and the original images. With this technique, for the images in Figure 2a,b we computed two slightly different PSFs, one for the recto and the other for the verso, that were then roughly approximated with a same Gaussian of standard deviation 1. It is to be said that, whereas the inclusion of a blur kernel is often essential for the method to be successful, we experimentally found that the value of this kernel is not really critical. Hence, by using the

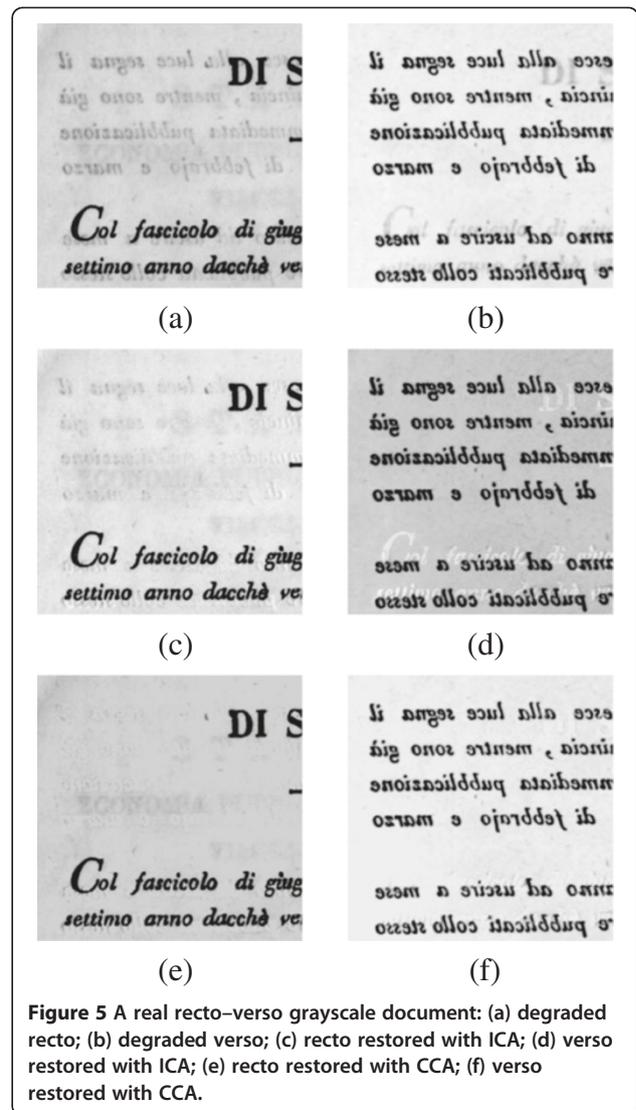


Figure 5 A real recto-verso grayscale document: (a) degraded recto; (b) degraded verso; (c) recto restored with ICA; (d) verso restored with ICA; (e) recto restored with CCA; (f) verso restored with CCA.

above specified blur kernel, and limiting the non-zero annular bins to 10, we obtained the restored images shown in Figure 2e,f, whereas Figure 2c,d shows the results of FastICA. It is apparent that the same behaviour of CCA with respect to ICA, already highlighted in the synthetic experiment, can clearly be appreciated also in this case. Specifically, the residual interferences left by ICA are significantly reduced in the CCA solution, especially those of the recto side. Note that this result is comparable with that presented for the same image in [11], where a nonlinear model is solved through total variation regularization. Nevertheless, the algorithm proposed therein is much more computationally demanding than CCA.

Figures 3 and 4 show two of the several experiments performed on recto and verso pairs of real documents acquired in RGB.

In the first example of Figure 3, we deemed to neglect the blur effect, and compared the performance of CCA with that of FastICA. In particular, Figure 3a,b shows the original recto and verso scans, Figure 3c,d shows the results obtained by FastICA and Figure 3e,f shows the results of CCA. Again, CCA outperforms ICA by producing almost perfectly cleansed reconstructions.

In the second example of Figure 4, we included a blur in the form of a Gaussian of standard deviation 2, to account for the sensible smearing of the see-through pattern, and obtained the results of Figure 4c,d (an enlarged portion is shown for a better qualitative evaluation).

Finally, a last experiment on a real grayscale recto-verso pair is shown in Figure 5, where the results of CCA are compared with those obtained with FastICA or, equivalently, by data decorrelation through symmetric whitening.

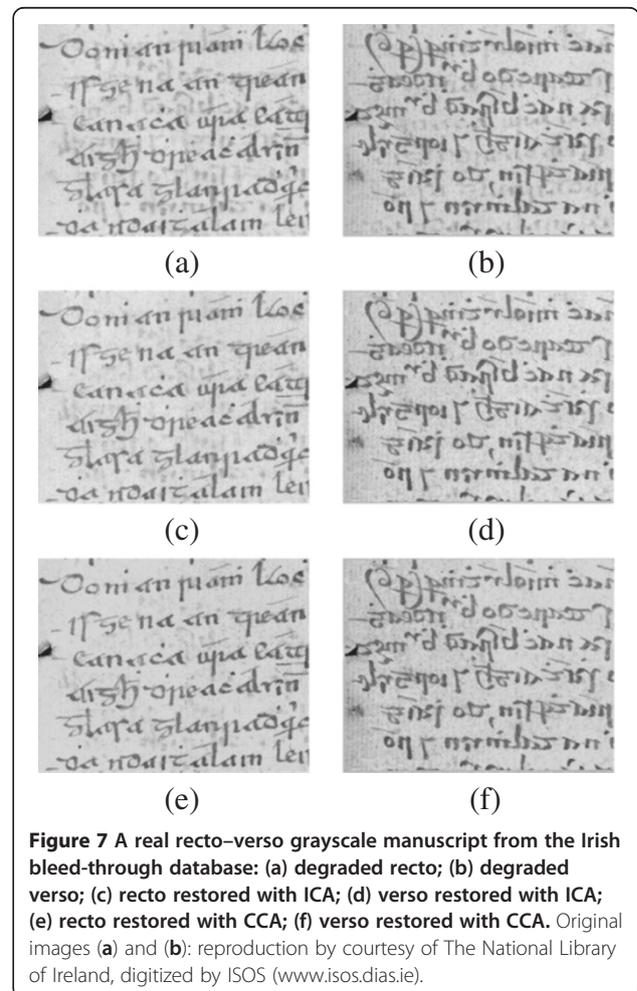
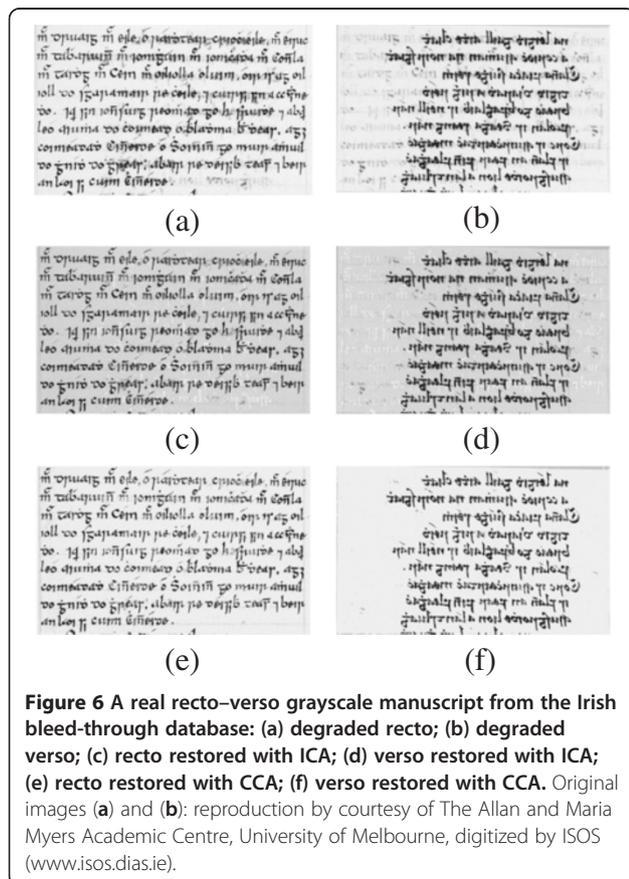
4.2. Test images: the Irish bleed-through database

The bleed-through database at the website [20] comprises 25 registered recto-verso sample grayscale image pairs, taken from larger high-resolution manuscript images, with varied degrees of bleed-through. In addition, for each image a binary ground-truth mask of the foreground text is provided. Although these ground truth

images are synthetic, i.e. manually created, they can be useful for a quantitative analysis of the results. Furthermore, in our case, they can also be used for estimating the cross-correlation of the clean, ideal recto and verso foreground texts.

We have experimented both ICA and CCA on a large subset of the whole database, and have found that the images can roughly be subdivided into two categories: those images where CCA and ICA perform similarly, and those images where CCA is definitely superior to ICA. As one might expect, the images whose corresponding ground-truths exhibit a low correlation fall in the first category, whereas when the cross-correlation of the ground-truths is significant, CCA outperforms ICA. In the following, we report two examples that are representative of this general behaviour of the two algorithms.

Figure 6 shows a manuscript belonging to the Allan and Maria Myers Academic Centre, University of Melbourne (Figure 6a,b), processed with symmetric whitening (Figure 6c,d), and with CCA (Figure 6e,f), respectively. The CCA results have been obtained by neglecting blur, and limiting the non-zero annular bins to 4. The



apparent superiority of CCA with respect to ICA can be explained by the fact that the correlation index between the recto and verso ground-truth images is rather high, specifically 0.1291. To quantitatively evaluate the quality of the reconstructions, we applied an automatic Otsu binarization [21] to the images of Figure 6e,f, and then counted the number of pixels that are different from their homologous in the corresponding ground-truth images. For the recto side, we obtained a percentage of 1.95, whereas for the verso side the percentage of wrong pixels was 1.62. Of course, these measurements of quality can only be considered as indicative, in that, as already said, the ground-truth images were built manually, so that they are not fully trustable.

Figure 7 shows instead an example of images from the same database where CCA performs qualitatively similarly to ICA. A reason for that could be the low correlation index between the recto and verso clean foreground texts that, as estimated from the ground-truth masks, results to be -0.0336 .

5. Conclusions

We have shown that a technique of CCA significantly outperforms ICA when applied to the restoration of RGB recto-verso pairs of historical documents. This can be achieved without affecting the typical computational efficiency and the unsupervised nature of BSS techniques, which make them suitable also for routinely application to large datasets of archival documents.

Differently from ICA, with CCA separation can be achieved also when the individual sources are largely correlated. This is especially true when the patterns that interfere from a side to the other of the page are sensibly blurred, for effect of light or ink spreading through the support. Although the method can easily account for these blur kernels on the sources, at present we considered them known, or we estimate them off-line. We are currently studying a strategy to jointly estimate the blur kernels along with all the other parameters.

We should point out that our method is based on a linear, although convolutional, mixing model. This is undoubtedly a limitation, in that the see-through effect is likely to be nonlinear. In fact, some recent works (see, e.g. [14]) have shown that, using a nonlinear convolutional model, excellent results can be obtained, although at the price of a higher computational cost.

However, many issues still remain open, along the difficult way to find a comprehensive model that is able to describe all multiple and varied causes behind the see-through phenomenon in ancient documents. In our opinion, the two most critical open issues are correlation of the sources and non-stationarity of the degradation. Neither models nor methods are presently available to simultaneously address both problems. This article aims

to give a contribution, supported by promising results, towards the solution of the source correlation problem in the linear convolutional case. A next step could be to include the treatment of source correlation within a nonlinear convolutional data model.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This study was supported by the European funds, through the program POR Calabria FESR 2007–2013 - PIA Regione Calabria, project ITACA (Innovative Tools for cultural heritage ArChiving and restorAtion).

Received: 25 September 2012 Accepted: 25 February 2013

Published: 24 March 2013

References

1. CL Tan, R Cao, P Shen, Restoration of archival documents using a wavelet technique. *IEEE Trans. Pattern Anal.* **24**, 1399–1404 (2002)
2. RF Moghaddam, M Cheriet, Low quality document image modeling and enhancement. *Int. J. Doc. Anal. Recognit.* **11**, 183–201 (2009)
3. E Dubois, A Pathak, Reduction of bleed-through in scanned manuscript documents, in *Proceedings of the Image Processing, Image Quality, Image Capture Systems Conference (PICS)*, vol. 4 (Montreal, Canada, 2001), pp. 177–180
4. P Dano, *Joint restoration and compression of document images with bleed-through distortion* (Ottawa-Carleton Institute for Electrical and Computer Engineering, School of Information Technology and Engineering, University of Ottawa, 2003). Dissertation
5. K Knox, *Show-through correction for two-sided documents*, U.S. Patent 5,832,137, 1998
6. Q Wang, CL Tan, Matching of double-sided document images to remove interference, in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)* (Hawaii, USA 1, 1084–1089 (2001)
7. G Sharma, Show-through cancellation in scans of duplex printed documents. *IEEE Trans. Image Process.* **10**, 736–754 (2001)
8. A Tonazzini, E Salerno, L Bedini, Fast correction of bleed-through distortion in greyscale documents by a blind source separation technique. *Int. J. Doc. Anal. Recognit.* **10**, 17–25 (2007)
9. A Tonazzini, G Bianco, E Salerno, Registration and enhancement of double-sided degraded manuscripts acquired in multispectral modality, in *Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR)* (Barcelona, Spain, 2009), pp. 546–550
10. F Merrikh-Bayat, M Babaie-Zadeh, C Jutten, Using non-negative matrix factorization for removing show-through, in *Proceedings of the International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, vol. LNCS 6365, ed. by (St. Malo, France, 2010), pp. 482–489
11. B Ophir, D Malah, Show-through cancellation in scanned images using blind source separation techniques, in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, vol. 3, ed. by (San Antonio, Texas, USA, 2007), pp. 233–236
12. A Tonazzini, I Gerace, F Martinelli, Multichannel blind separation and deconvolution of images for document analysis. *IEEE Trans. Image Process.* **19**, 912–925 (2010)
13. F Merrikh-Bayat, M Babaie-Zadeh, C Jutten, Linear-quadratic blind source separating structure for removing show-through in scanned documents. *Int. J. Doc. Anal. Recognit.* **14**, 319–333 (2011)
14. E Salerno, F Martinelli, A Tonazzini, Nonlinear model identification and seethrough cancellation from recto-verso data. *Int. J. Doc. Anal. Recognit.* published online 17 March 2012
15. A Tonazzini, L Bedini, E Salerno, Independent component analysis for document restoration. *Int. J. Doc. Anal. Recognit.* **7**, 17–27 (2004)
16. P Comon, C Jutten, *Handbook of Blind Source Separation*, 1st edn. (Academic Press, New York, 2010)
17. S Ricciardi, A Bonaldi, P Natoli, G Polenta, C Baccigalupi, E Salerno, K Kayabol, L Bedini, G De Zotti, Correlated component analysis for diffuse component separation with error estimation on simulated Planck polarization data. *Mon. Not. R. Astron. Soc.* **406**, 1644–1658 (2010)
18. L Bedini, E Salerno, *Fourier-domain implementation of correlated component analysis, with error estimation* (Internal Report ISTI-CNR, 2008)

19. H Gävert, J Hurri, J Särelä, A Hyvärinen, *The FastICA package for MATLAB*, 2005. <http://research.ics.aalto.fi/ica/fastica/>
20. Irish Script On Screen, *Sigmedia, Bleed-Through Database*, 2012. <http://www.isos.dias.ie/master.html?http://www.isos.dias.ie/libraries/Sigmedia/english/index.html?ref=>
21. N Otsu, A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man. Cybern.* **9**(1), 62–66 (1979)

doi:10.1186/1687-6180-2013-58

Cite this article as: Tonazzini and Bedini: Restoration of recto-verso colour documents using correlated component analysis. *EURASIP Journal on Advances in Signal Processing* 2013 **2013**:58.

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- ▶ Convenient online submission
- ▶ Rigorous peer review
- ▶ Immediate publication on acceptance
- ▶ Open access: articles freely available online
- ▶ High visibility within the field
- ▶ Retaining the copyright to your article

Submit your next manuscript at ▶ springeropen.com
