JOINT BLIND SEPARATION AND RESTORATION OF MIXED DEGRADED IMAGES FOR DOCUMENT ANALYSIS

Anna Tonazzini

Istituto di Scienza e Tecnologie dell'Informazione
Consiglio Nazionale delle Ricerche
via G. Moruzzi 1, I-56124 Pisa, Italy
anna.tonazzini@isti.cnr.it

Ivan Gerace and Francesco Cricco

Dipartimento di Matematica e Informatica
Università degli Studi di Perugia
Via Vanvitelli, 1, I-06123 Perugia, Italy
{gerace, cricco}@dipmat.unipg.it

ABSTRACT

We consider the problem of extracting clean images from noisy mixtures of images degraded by blur operators. This special case of source separation arises, for instance, when analyzing document images showing bleed-through or show-through. We propose to jointly perform demixing and deblurring by augmenting blind source separation with a step of image restoration. Within the ICA approach, i.e. assuming the statistical independence of the sources, we adopt a Bayesian formulation were the priors on the ideal images are given in the form of MRF, and a MAP estimation is employed for the joint recovery of the mixing matrix and the images. We show that taking into account for the blur model and for a proper image model improves the separation process and makes it more robust against noise. Preliminary results on synthetic examples of documents exhibiting bleed-through are provided, considering edge-preserving priors that are suitable to describe text images.

1. INTRODUCTION

This paper deals with the blind separation and reconstruction of images from linear mixtures of degraded versions of the sources. A potential application is the recovery of legible recto and verso text patterns from documents showing bleed-through or show-through. In many ancient documents, bleed-through is caused by seeping of ink from the reverse side, while show-through can appear in the scan of modern documents when the paper is not opaque. So far, this problem has been dealt from the point of view of removing these effects for obtaining a clean foreground text [7][9], and in most methods a preliminary registration of the front and back pages of the document is required. Specific situations in which the interfering texts can be of interests themselves, as where the scan of the verso page is not available, have not been considered. Instead, we adopt the novel approach to formulate the problem as one of separating overlapped independent patterns, and propose to exploit blind source separation techniques [1][5]. In this approach, the overlapping texts and the support are the unknown sources to be recovered, and multispectral views of the document (e.g. the red, green and blue channels) can be considered as the observed data set, constituted of different mixtures with unknown coefficients of the source patterns.

In [11] we showed that, even assuming pure linear mixtures, under suitable conditions (e.g. independent patterns and noiseless images) the problem can be efficiently solved through Independent Component Analysis (ICA) techniques. We employed the FastICA algorithm [4], which is a fully blind and extremely fast procedure, to separate underwritings and overwritings in real ancient documents and in palimpsests. Nevertheless, the physical model underlying bleed-through or show-through is very complicated, since it is actually nonlinear with some unknown parameters [7]. For instance, the bleed-through pattern is very frequently degraded by strong blur, due to the ink diffusion or light spreading through the support. A step towards a mathematical model which is more adhering to the physics of the problem is to augment the linear mixture with a blur model for the individual patterns, and take into account for additive noise in the document scan. In these conditions, pure ICA algorithms do not perform satisfactorily for separation and, obviously, cannot deal with blur in the sources.

In [10] we proposed a Bayesian approach to blind source separation, employing Markov Random Field (MRF) models to describe the image local autocorrelation structure, and showed that this approach is robust against noise in the data. From the wide literature, it is also known that Bayesian approaches and MRF image models can satisfactorily solve the image restoration problem as well. In this paper, we thus propose to integrate demixing and deblurring, by augmenting Bayesian, MRF-based blind source separation with a step of image restoration. Assuming the knowledge of the blur operators, we adopt edge-preserving priors suitable to describe text images, and employ MAP estimation for the joint recovery of the mixing matrix and the images.

0-7803-8554-3/04/$20.00

We show that taking into account for the blur model and adopting proper MRF image models improves the separation process. Furthermore, this formulation is suitable to be extended to the joint estimation of the blur operator [3] as well.

2. BAYESIAN FORMULATION OF JOINT DEMIXING AND DEBLURRING

According to the BSS and the image degradation formalism, the data generation model we consider is given by:

$$x_i(t) = \sum_{j=1}^{N} A_{ij}(H_j \mathbf{s}_j^T)(t) + n_i(t), \qquad t = 1, 2, \ldots, T \quad i = 1, 2, \ldots, N \quad (1)$$

where $\mathbf{x}_i$, $\mathbf{s}_i$ and $\mathbf{n}_i$ are the i-th measured, source and noise signals, respectively, in row vector form. The same number N of measurements and unknown sources has been assumed so that A, the unknown mixing matrix, is $N \times N$. Quantity $H_i \mathbf{s}_i^T$ is the degraded version of source $\mathbf{s}_i$, where the blur matrix H_i, assumed known, performs convolution between a source image and a blur mask as a matrix-vector product. Thus, the data images are noisy linear instantaneous mixtures of the blurred sources. We adopt the notation where $\mathbf{x}(t)$ is the column vector of the N measurements at pixel t, while $\mathbf{x}$ is the $N \times T$ matrix of all the measurements. (This notation extends also to the other variables. In particular, $\overline{\mathbf{Hs}}$ indicates the $N \times T$ matrix of the degraded sources). Considering a white and Gaussian noise with zero mean, the logarithm of the likelihood $P(\mathbf{x}|\mathbf{s}, A)$ is given by:

$$-\frac{1}{2} \sum_t \left(A\overline{\mathbf{Hs}}(t) - \mathbf{x}(t)\right)^T \Sigma_t^{-1} \left(A\overline{\mathbf{Hs}}(t) - \mathbf{x}(t)\right) \quad (2)$$

where Σ is the covariance matrix of the noise, assumed, in general, to be location-dependent. In a fully Bayesian approach, both A and $\mathbf{s}$ are assumed as independent unknowns, and are assigned with prior distributions $P(A)$ and $P(\mathbf{s})$, respectively. Then, A and $\mathbf{s}$ can be simultaneously estimated by maximizing the posterior distribution $P(\mathbf{s}, A|\mathbf{x})$. Hence, our problem can be stated as the following MAP estimation problem:

$$\begin{aligned} (\hat{\mathbf{s}}, \hat{A}) &= arg \max_{\mathbf{s},A} P(\mathbf{s}, A, |\mathbf{x}) = \\ &= arg \max_{\mathbf{s},A} P(\mathbf{x}|\mathbf{s}, A) P(\mathbf{s}) P(A) \end{aligned} \quad (3)$$

In the ICA approach A is assumed to have a uniform prior and the sources are assumed only to be mutually independent, that is $P(\mathbf{s})$ is given in the form of a product of the marginals $P_i(\mathbf{s}_i)$. In this paper, we assume specific Gibbs distributions, i.e. Markov Random Field models, for the marginals, in such a way to describe regularity properties, under the form of local spatial correlation (smoothness), for the individual sources.

A popular strategy to reduce to very complex joint MAP estimations consists in iteratively alternating steps of estimation with respect to the different sets of variables. In our case, this results in steps of estimation of A and $\mathbf{s}$, respectively:

$$A^{(k)} = arg \max_{A} P(\mathbf{x}|\mathbf{s}^{(k-1)}, A) P(A) \quad (4)$$

$$\mathbf{s}^{(k)} = arg \max_{\mathbf{s}} P(\mathbf{x}|\mathbf{s}, A^{(k)}) P(\mathbf{s}) \quad (5)$$

For general forms of the prior adopted for A, problem of eq. 4 can be non-concave, and algorithms for non-convex optimization must be adopted, such as simulated annealing (SA). In this case, however, owing to the small number of variables, even SA is reasonably cheap. Moreover, SA can be suitable when $P(A)$ enforces constraints on A that cannot be expressed in analytical form (e.g. bounds on the admissible values). The algorithm for solving problem of eq. 5 mainly depends on the form adopted for the $P_i(\mathbf{s}_i)$. However, when appropriate, the various $P_i(\mathbf{s}_i)$ can be chosen in such a way to ensure the concavity of the function to be optimized, so that gradient ascent can be used. Otherwise, still relatively cheap algorithms for non-convex optimization can be used, such as Graduated Non-Convexity (GNC) [2].

3. THE MRF IMAGE MODEL

MRF models are very popular especially in connection to inverse problems of image processing, such as restoration, denoising, segmentation, optical flow estimation, etc. Through MRF models, it is indeed possible to describe local properties of the images, such as edges, in order to make space-variant the regularizing smoothness constraint. Furthermore, the local nature of these models allows us to devise distributed and parallel algorithms. Let us consider then the distribution of the $i-th$ source $s_i(t)$ in our problem. According to the MRF formalism, it must have the following Gibbs form:

$$P_i(\mathbf{s}_i) = \frac{1}{Z_i} \exp\left\{-U_i(\mathbf{s}_i)\right\} \quad (6)$$

where Z_i is the normalizing constant and $U_i(\mathbf{s}_i)$ is the prior energy in the form of a sum of potential functions over the set of cliques of interacting locations. The number of different cliques and their shape are related to the order of the neighborhood system adopted for the MRF, i.e. to the extent of correlation among the pixels, while the functional form of the potentials determines the correlation strength. At present, we limit our choice to the set of cliques constituted of two adjacent pixels t and r, in the 2D grid of the image. We then define $U_i(\mathbf{s}_i)$ as:

$$U_i(\mathbf{s}_i) = \lambda_i \sum_t \sum_{r \in S_t} g_i\left(s_i(t) - s_i(r)\right) \quad (7)$$

where λ_i is the positive regularization parameter, S_t is the first order neighborhood for location t, and g_i is a stabilizer function that describes the regularity in the $i-th$ source, by penalizing high gradient values. This regularity is an essential, physically plausible, constraint to prevent instability of the reconstructions when the data are noisy. Nevertheless, the source signals can present some steep fronts (e.g. the character boundaries) which must be preserved. We thus refer to stabilizers for edge-preserving image recovery [6], which have been proved to be very efficient in deblurring techniques. A stabilizer is edge-preserving when it increases slower than quadratically. Among those possessing this property, we considered the convex one suggested by Shulman and Herve [8]:

$$g(z) = \begin{cases} z^2 & \text{if } |z| \leq \delta \\ 2\delta|z| - \delta^2 & \text{if } |z| > \delta \end{cases} \tag{8}$$

where parameter δ is the threshold for the intensity gradient above which the stabilizer becomes linear and therefore adaptive to discontinuities.

In our study, we referred to the most general assessment of BSS problems, i.e. no prior information is assumed for the mixing matrix, so that $P(A)$ is the uniform distribution. This, together with the choice of convex source models, and the assumption of white and Gaussian noise, makes the posterior distribution concave in both variables, so that the alternating maximization scheme of eqs. (4)- (5) can be implemented without stochastic relaxation. In particular, a gradient ascent can be used to solve eq. (5), while eq. (4) results in the following analytic updating formula for A:

$$A = \mathbf{x}\overline{\mathbf{Hs}}^T \left[\overline{\mathbf{Hs}}\,\overline{\mathbf{Hs}}^T\right]^{-1} \tag{9}$$

It is worth noting that this alternating maximization procedure is suitable to be augmented with steps of estimation of the MRF model parameters and of the blur operators.

4. SIMULATION RESULTS

The performance of the method described in this paper was tested on synthetic images that simulate multispectral views of documents exhibiting the bleed-through or show-through effect. A first experiment was however aimed at showing the ability of the method in simultaneously separating and restoring generic images, even in presence of significant blur and noise. We thus considered two simple 64×64 piecewise constant images, both degraded with a uniform 3×3 blur mask, mixed with the following, randomly generated, mixing matrix:

$$A = \begin{bmatrix} 0.7035 & 0.2985 \\ 0.3107 & 0.4096 \end{bmatrix}$$

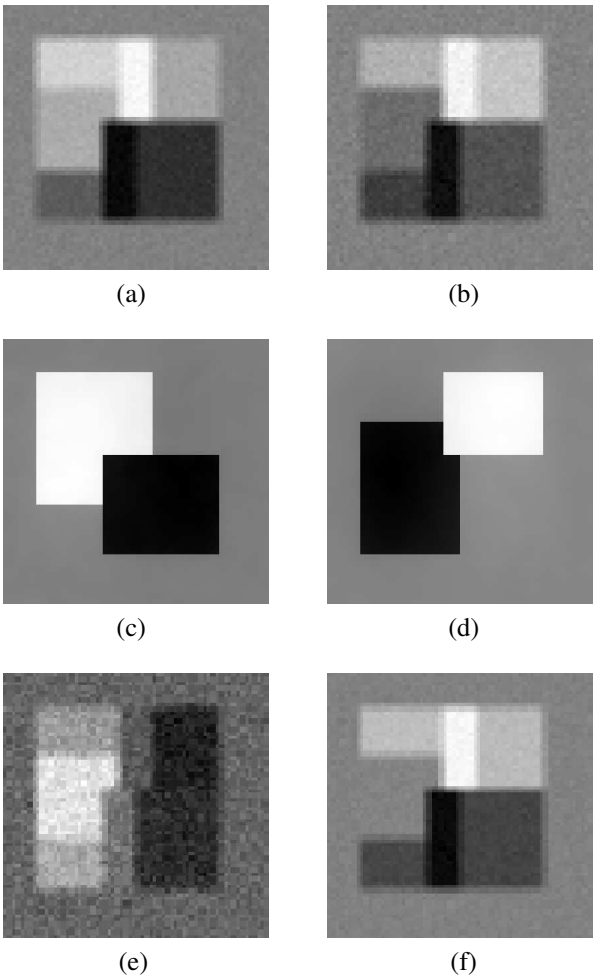

Fig. 1. Synthetic piecewise constant images: (a) first mixture; (b) second mixture; (c) first demixed deblurred image; (d) second demixed deblurred image; (e) first FastICA output; (f) second FastICA output.

and then added with a space-invariant white Gaussian noise (SNR=26 dB). Employing the stabilizer of eq. 8 with parameters $\lambda = 1$ and $\delta = 2$ for both images, we obtained the results shown in Figure 1, that includes the FastICA outputs for comparison. The better performance of our method in separating is apparent, and in fact the estimated mixing matrices (after column rescaling) were respectively:

$$\hat{A}_{MRF} = \begin{bmatrix} 0.7035 & 0.2985 \\ 0.3186 & 0.3968 \end{bmatrix} \quad \hat{A}_{FI} = \begin{bmatrix} 0.7035 & 0.2985 \\ -0.4135 & 0.2731 \end{bmatrix}$$

for our method and for FastICA, respectively. With several randomly selected mixing matrices and several noise realizations, we always obtained similar results.

For the experiments with synthetic images that simulate documents with bleed-through or show-through, we considered the overlapping of two blurred texts, and therefore used

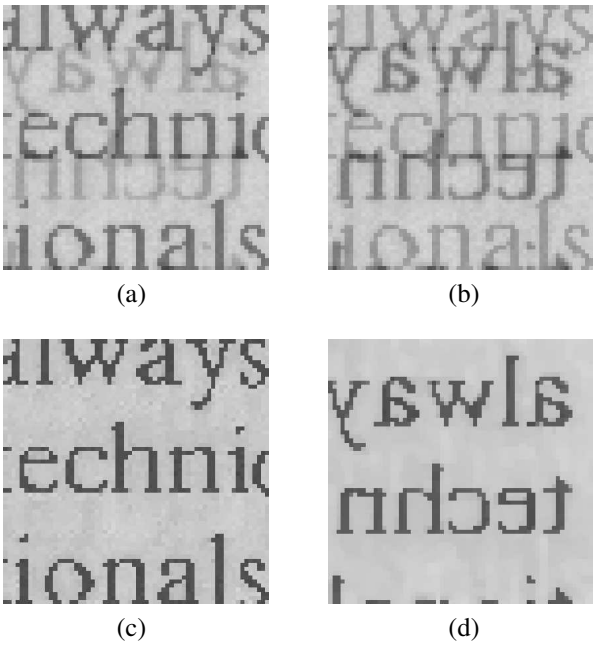

Fig. 2. Synthetic images simulating overlapped blurred texts: (a) first mixture; (b) second mixture; (c) first demixed deblurred image; (d) second demixed deblurred image.

two views as data set. Since the two ideal text patterns have similar characteristics, we assumed the same stabilizer of eq. 8 for each of them, but adopted different parameters, to account for the different amount of degradation. Indeed, the first image, ideally corresponding to the foreground text, was blurred with the following h_1 blur mask, while the second one, corresponding to the bleed-through pattern, was blurred with the heavier h_2 mask:

$$h_1 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 10 & 1 \\ 0 & 1 & 0 \end{bmatrix} \qquad h_2 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 3 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

Figure 2 shows the results obtained in this synthetic but realistic case, using $\lambda = 0.3$ and $\delta = 2.1$ for the first image and $\lambda = 0.3$ and $\delta = 1.9$ for the second image, respectively, when the noise in the data is of 26dB. Experiments on real documents are in progress.

5. CONCLUSIONS

We proposed a Bayesian formulation of ICA techniques for the joint blind source separation and restoration of noisy mixtures of degraded images. Assuming statistical independence of the sources, we considered MRF image models to describe local spatial autocorrelation. In particular, we proposed convex edge-preserving models, and an alternating maximization scheme for the joint MAP estimation of the mixing and the sources, which employs gradient ascent algorithms. We presented preliminary results of the application of this technique to the separation of overlapped texts in documents showing bleed-through or show-through. Our method involves a linear data model, where each color channel of the input image is a mixture of all the patterns to be extracted, but we introduced a blur model to account for the typical degradation of the bleed-through pattern, and for the optical blur and noise due to the scanning process or to smearing and diffusion of the ink. The possibility of separating and restoring interfering patterns in documents is crucial for improving readability, both by a human reader and by an OCR system. Another possible application is the enhancement of traces of erased texts, as in ancient palimpsests. Current research is towards automatic selection of the regularization parameters, treatment of non-stationary noise, typical of real ancient documents, and joint estimation of the blur operators.

6. REFERENCES

[1] S. Amari and A. Cichocki, "Adaptive blind signal processing - neural network approaches", Proc. IEEE, vol. 86, pp. 2026-2048, 1998.

[2] L. Bedini, I. Gerace, E. Salerno and A. Tonazzini, "Models and algorithms for edge-preserving image reconstruction", Adv. Imaging and Electron Physics, vol. 97, pp. 86-189, 1996.

[3] L. Bedini and A. Tonazzini, "Fast fully data driven image restoration by means of edge-preserving regularization", Real-Time Imaging, vol.7, pp. 3-19, 2001.

[4] A. Hyvärinen, "Fast and Robust Fixed-Point Algorithms for Independent Component Analysis", IEEE Trans. NN, vol.10, pp. 626-634, 1999.

[5] T. Lee, M. Lewicki and T. Sejnowski, "Independent component analysis using an extended infomax algorithm for mixed sub-gaussian and super-gaussian sources", Neural Computation, vol. 11, pp. 409-433, 1999.

[6] S.Z. Li, "On discontinuity-adaptive smoothness priors in vision", IEEE Trans. PAMI, vol. 17, pp. 576-586, 1995.

[7] G. Sharma, "Show-through cancellation in duplex printed documents", IEEE Trans. IP, vol. 10, pp. 736–754, 2001.

[8] D. Shulman and J.Y. Herve, "Regularization of discontinuous flow fields", in Proc. Work. Visual Motion, pp. 81-86, 1989.

[9] C.L. Tan, R. Cao, and P. Shen, "Restoration of archival documents using a wavelet technique", IEEE Trans. PAMI, vol. 24(10), pp. 1399–1404, 2002.

[10] A. Tonazzini, L. Bedini, E. Kuruoglu, and E. Salerno, "Blind separation of auto-correlated images from noisy mixtures using MRF models", in Proc. ICA2003, pp. 675–680, 2003.

[11] A. Tonazzini, L. Bedini, and E. Salerno, "Independent Component Analysis for document restoration", Int. J. on Document Analysis and Recognition, in press, 2004.