Bleed-Through Removal from Degraded Documents Using a Color Decorrelation Method

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. A color decorrelation strategy to improve the human or automatic readability of degraded documents is presented. The particular degradation that is considered here is bleed-through, that is, a pattern that interferes with the text to be read due to seeping of ink from the reverse side of the document. A simplified linear model for this degradation is introduced to permit the application of very fast decorrelation techniques to the RGB components of the color data images, and to compare this strategy to the independent component analysis approach. Some examples from an extensive experimentation with real ancient documents are described, and the possibility to further improve the restoration performance by using hyperspectral/multispectral data is envisaged.

1 Introduction

Improving the readability of printed or manuscript documents is a common need in libraries and archives. The original documents should not be altered, but the availability of more readable digital versions is an important aid to the scholar. Furthermore, one of the main tasks in document analysis is to produce machine-readable versions of original texts. This task should be performed automatically, or, at least, minimizing human intervention. This is normally done by optical character recognition systems, whose performance, however, depends on the quality of the original documents. Ancient documents, in particular, are often affected by several types of degradations. Restoring the digital scans of the original documents is thus essential to improve human readability and to regain acceptable OCR performances.

The particular kind of degradation we are considering here is *bleed-through*, that is, presence of patterns interfering with the main text due to seeping of ink from the reverse page side. Removing the bleed-through pattern from a digital image of a document is not trivial, especially with ancient originals, where interferences of this kind are usually very strong. Indeed, dealing with strong bleed-through degradation is practically impossible by any simple thresholding technique. Some work done on this specific problem has exploited information

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

S. Marinai and A. Dengel (Eds.): DAS 2004, LNCS 3163, pp. 229–240, 2004.

from the front and back pages [11] [4] [2]. The drawbacks of this type of techniques are that the scans from both sides of the documents must be available, and a preliminary registration of the two sides is needed. In addition, they are usually expensive, as the processing may be quite complicated. In [9], a color scan from a single side is required, but a thresholding technique can only be used in the framework of multiresolution analysis and adaptive binarization.

Our approach to this problem is to model a document image as a linear combination of three independent patterns: the main foreground text, the bleed-through, and the background pattern, i.e. an image of the paper or any other support, which can contain various interfering features, such as stains, color inhomogeneities, textures, etc. Our scope is to obtain a clean main text, by reducing the interferences from bleed-through and background. This goal can be achieved by processing multiple "views" of the mixed object. When a color scan of the document is available, three different views can be obtained from the red, green, and blue image channels, but scans at nonvisible wavelengths can also be available. By processing the different color components, it is possible to extract the main text pattern, and, sometimes, even to achieve a complete separation of the overlapped patterns. Although our linear image model is known to be naïve [11], it has already proved to give interesting results for extracting the hidden texts from color images of palimpsests, assuming to evaluate by visual inspection the mixture coefficients [5]. Nevertheless, in general, the mixture coefficients are not known, and the separation problem becomes one of blind source separation (BSS). An effective solution to BSS can be found if the source patterns are mutually independent, by using separation techniques based on independent component analysis, or ICA [7]. We already proposed ICA for document processing [13], obtaining good results with real manuscripts.

Instead of enforcing independence, in this paper we only try to decorrelate the observed data. As is known, this requirement is weaker than independence, and can be satisfied by transforming the data in order to get zero cross-covariances. Conversely, independence implies that the cross-central-moments of all orders between each pair of estimated sources must be zero. In principle, no source separation can be obtained by only constraining second-order statistics, at least if no additional requirements are satisfied [3]. However, our present aim is not full separation but interference reduction, and this can often be achieved even by only constraining second-order statistics. Furthermore, the second-order approach is always less expensive than most ICA algorithms and, in many cases, it is even more effective for our purposes. Enforcing statistical uncorrelation is equivalent to orthogonalize the different data images. The result of orthogonalization is of course not unique. We experimentally tested the performances of different strategies, and compared the results to the ones obtained by the ICA approach.

The paper is organized as follows. In Section 2, we introduce our linear data model. In Section 3, we recall the properties of different orthogonalization matrices, and, in Section 4, we present some experimental results with real printed or manuscript documents. Some final remarks highlight the promises of applying this type of techniques to color document processing.

2 Data Model

Let us assume to have a collection of T samples from a random N-vector $\mathbf{x}$, which is generated by linearly mixing the components of a random M-vector $\mathbf{s}$ through an $N \times M$ mixing matrix A:

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

Our present aim is to describe a color document image by means of a model of the type (1). Let us assume that each pixel (of index t) of a color scan of a document has a vector value $\mathbf{x}(t)$. Normally, the color vector has dimension 3 (it is composed, for example, by the red, green, and blue channels), thus we have $N = 3$. In our case, let us also assume that the image can be modeled as the superposition of three different sources, or classes, that we will call "background", "main text" and "bleed-through". Thus, we also have $M = 3$. In general, by using 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 $M = N = 3$, although in principle there is no difference with the general case. Since we consider images of documents containing text, we can also reasonably assume that the color of each source 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 main text and (r_3, g_3, b_3) for the bleed-through. In this application, we assume that the three sources mix linearly at each pixel, and that noise and blur can be neglected. By this approximated model, the reflectance indices $(x_r(t), x_g(t), x_b(t))$ of a generic point t of the document are given by:

$$\begin{bmatrix} x_r(t) \\ x_g(t) \\ x_b(t) \end{bmatrix} = \begin{bmatrix} r_1 & r_2 & r_3 \\ g_1 & g_2 & g_3 \\ b_1 & b_2 & b_3 \end{bmatrix} \begin{bmatrix} s_1(t) \\ s_2(t) \\ s_3(t) \end{bmatrix} \tag{2}$$

where functions $s_i(t)$, $i = 1, 2, 3$ denote the "quantity" of background, main text and bleed-through, respectively, that concur to form the color at point t. For instance, a pure background point t will be represented as $\mathbf{s}(t) = (1, 0, 0)$. Eq. (2) has the same form of 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.

However, this model does not perfectly account for the phenomenon of bleed-through. Just to mention one aspect, in the pixels where the bleed-through is superimposed to the main text, 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 [11], a nonlinear model is also derived for the phenomenon of show-through (interfering pattern from the reverse side due to transparency of the paper). Although the linear model is only a rough approximation, it has already been useful in several applications [5] [13]. As already mentioned, if both $\mathbf{s}$ and A in (1) are unknown, estimating them from $\mathbf{x}(t)$ alone is called a problem of blind source separation. For it to be uniquely solvable, some additional information is

needed. For example, if the components of vector $\mathbf{s}$ are statistically independent from each other, it can be solved by the ICA techniques [7]. Some good result can also be obtained even if the independence assumption is not completely verified. We already addressed this issue in [13]. In our work, we have found that, in many cases where ICA was not able to separate the individual sources, in some of our outputs the interfering patterns were greatly reduced. The failure of ICA can be ascribed to both the approximated model and the significant correlation between different sources. We found that, to reduce the presence of the interfering patterns, it suffices to decorrelate the signals, without requiring their mutual independence. Decorrelation is faster than ICA, but it is not unique, and it is interesting to assess experimentally the performances of different decorrelation strategies in terms of interference reduction.

3 Processing Strategy

Since we have no physical grounds to justify the ICA assumptions, the ICA output processes will not be guaranteed to be replicas of the original sources. However, one can try to maximize the information content in each component of the data vector by decorrelating the observed image channels. This amounts to force the cross-covariances of the data to zero. To avoid cumbersome notation, and without loss of generality, let us assume to have zero-mean data vectors. To obtain zero cross-covariances, we seek for a linear transformation

$$\mathbf{y}(t) = W\mathbf{x}(t) \tag{3}$$

such that:

$$< y_i y_j >= 0, \;\; \forall i, j = 1, \ldots, M, \;\; i \neq j \tag{4}$$

where the notation $< \cdot >$ means expectation, and W is generally an $M \times N$ matrix. 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 any orthonormal basis, any rigid rotation yields another orthonormal basis spanning the same subspace.

What we want to stress here is that linear data processing can help to restore color text images, although the linear model is not fully justified. This strategy resembles the standard color space transformations in image processing, which are able to enhance specific perceptual information in color images [10] [14] [1] [8]. In [10], the authors compare many fixed color transformations on the basis of their effects on the performance of a particular recursive segmentation algorithm. They argue that, among linear transformations, the ones that obtain maximum-variance components are the most effective. They thus derive a criterion for the evaluation of fixed transformations in comparison with the Karhunen-Loeve transformation, which is known to give orthogonal output vectors. This approach is also called principal component analysis (PCA), and one of its purposes is to find the most useful (the ones that have dominant variances) among a number of variables [3]. Our data covariance matrix is the $N \times N$ matrix:

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

where superscript T means transposition. Since we do not have the probability density function of vector $\mathbf{x}$, we are not able to compute the expectations needed. An approximated estimate of the covariance matrix can be drawn from the sample at our disposal, that is, from the RGB components of our image:

$$R_{\mathbf{xx}} \approx \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}(t)\mathbf{x}^T(t) \tag{6}$$

Since the data are normally correlated, matrix $R_{\mathbf{xx}}$ will be nondiagonal. The covariance matrix of vector $\mathbf{y}$ defined in (3) is

$$R_{\mathbf{yy}} =< W\mathbf{x}\mathbf{x}^T W^T >= WR_{\mathbf{xx}}W^T \tag{7}$$

To obtain property (4) for $\mathbf{y}$, we should require that matrix $R_{\mathbf{yy}}$ is diagonal.

$$R_{\mathbf{yy}} = WR_{\mathbf{xx}}W^T = D_M \tag{8}$$

where D_M is any diagonal matrix of order M. Let us now perform the eigenvalue decomposition of matrix $R_{\mathbf{xx}}$

$$R_{\mathbf{xx}} = V_{\mathbf{x}}\Lambda_{\mathbf{x}}V_{\mathbf{x}}^T \tag{9}$$

where $V_{\mathbf{x}}$ is the matrix of the eigenvectors of $R_{\mathbf{xx}}$, and $\Lambda_{\mathbf{x}}$ is 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 $R_{\mathbf{yy}}$:

$$W_o = V_{\mathbf{x}}^T \tag{10}$$

$$W_w = \Lambda_{\mathbf{x}}^{-\frac{1}{2}}V_{\mathbf{x}}^T \tag{11}$$

$$W_s = V_{\mathbf{x}}\Lambda_{\mathbf{x}}^{-\frac{1}{2}}V_{\mathbf{x}}^T \tag{12}$$

Matrix $\Lambda_{\mathbf{x}}^{-\frac{1}{2}}$ is a diagonal matrix whose elements are the reciprocals of the square roots of the elements of $\Lambda_{\mathbf{x}}$. Matrix W_o produces a set of vectors $y_i(t)$ that are orthogonal to each other; indeed, from (7), (9) and (10):

$$W_oR_{\mathbf{xx}}W_o^T = V_{\mathbf{x}}^TV_{\mathbf{x}}\Lambda_{\mathbf{x}}V_{\mathbf{x}}^TV_{\mathbf{x}} = \Lambda_{\mathbf{x}} \tag{13}$$

Vectors y_i are thus mutually orthogonal, and their 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. From (7), (9) and (11), we have:

$$W_wR_{\mathbf{xx}}W_w^T = \Lambda_{\mathbf{x}}^{-\frac{1}{2}}V_{\mathbf{x}}^TV_{\mathbf{x}}\Lambda_{\mathbf{x}}V_{\mathbf{x}}^TV_{\mathbf{x}}\Lambda_{\mathbf{x}}^{-\frac{1}{2}} = I_N \tag{14}$$

the orthogonal vectors $y_i(t)$ are thus on a spherical surface (*whitening*, or *Mahalanobis transform*). Note that any whitening matrix can be multiplied from the left by an orthogonal matrix, and relation (14) still holds true. In particular, if

we use matrix W_s defined in (12), 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. Our experimental work has consisted in applying the above matrices to typical images of ancient documents, assuming a linear model of type (2) to be valid, with the aim at emphasizing the main text in the whitened vectors and reducing the influence of background and bleed-through.

For each test image, the results are of course different for different whitening matrices. However, it is interesting to note that, for bleed-through reduction, the symmetric whitening matrix often performs better than ICA, which applies a further rotation to the output vectors, based on higher-order statistics. On the other hand, in some cases, whitening can also achieve a separation of the three different components, which is the final aim of ICA processing.

4 Experimental Results

In this section, we show some examples from our extensive experimentation on real degraded documents. From this set of experiments, we can already draw some general considerations. A quantitative assessment of the technique should rely on "ground truth" data, which of course are not available when treating real images. To assess the technique against synthetic images, these should be generated from a realistic mixture model, since evaluating the results obtained from a model that perfectly fits the data would be almost meaningless. Figure 1 shows, in grayscale, an ancient manuscript affected by strong bleed-through, one output from symmetric whitening, and one output of the FastICA algorithm [6] [13], both obtained from the RGB components of the original color image. The PCA result is not shown, since its best output is similar to a graylevel version of the original; no text enhancement has been achieved. Conversely, the result from symmetric whitening shows an effective bleed-through reduction. Instead, in this case, FastICA failed to achieve a complete class separation, and was not able to produce one output with a clean main text pattern. As mentioned above, if both the data model is accurate and the image classes are mutually independent, FastICA processing should effectively separate the classes from the mixed RGB data. In this case, none of the two hypotheses is verified, and symmetric orthogonalization apparently performs better than ICA. In Figure 2, we report an example where the three methods perform similarly, though the result of symmetric orthogonalization is slightly better than those of FastICA and PCA. However, FastICA produced one output with a clean main text pattern, even if it still failed to achieve a complete class separation. In Figure 3, the grayscale image of another ancient manuscript with strong bleed-through and one of the RGB orthogonalization outputs are shown. The apparent bleed-through suppression clearly improves the human legibility of the document. To show also the advantages of orthogonalization versus an improvement of the performance of OCR systems, in Figure 4 we report the results of applying the QIR optimal thresholding technique [12] to the grayscale document scan and to the orthogo-

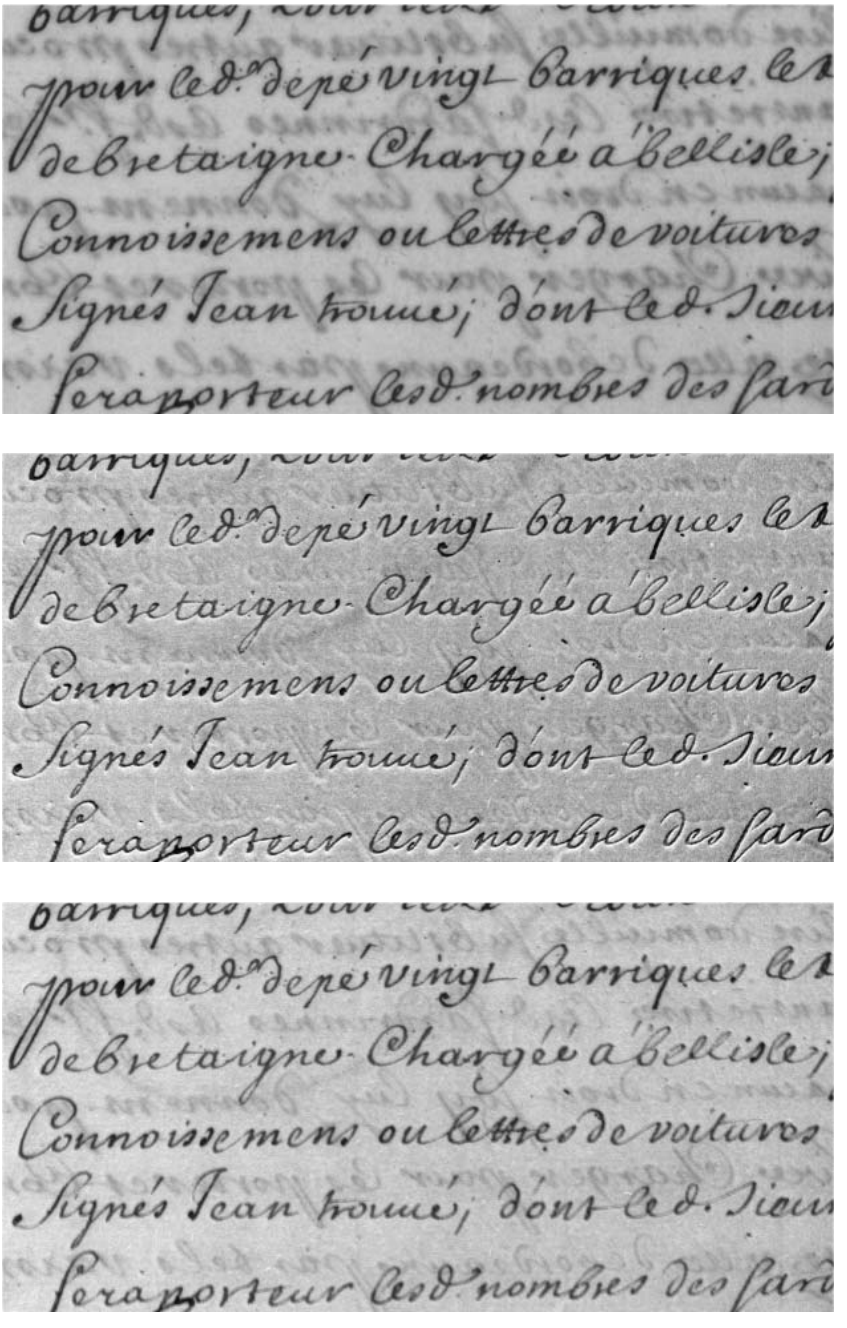

Fig. 1. From top to bottom: Scan of an ancient manuscript with bleed-through interference; Selected symmetric orthogonalization output from the RGB components of the color image; Selected FastICA output from the same data set.

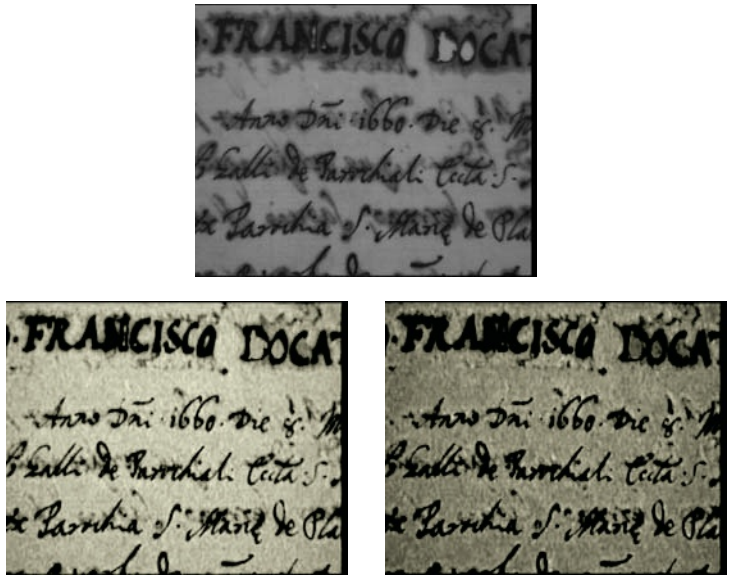

Fig. 2. Left: Scan of a manuscript with bleed-through interference. Middle: Selected symmetric orthogonalization output from the RGB components. Right: Selected FastICA output from the same data set.

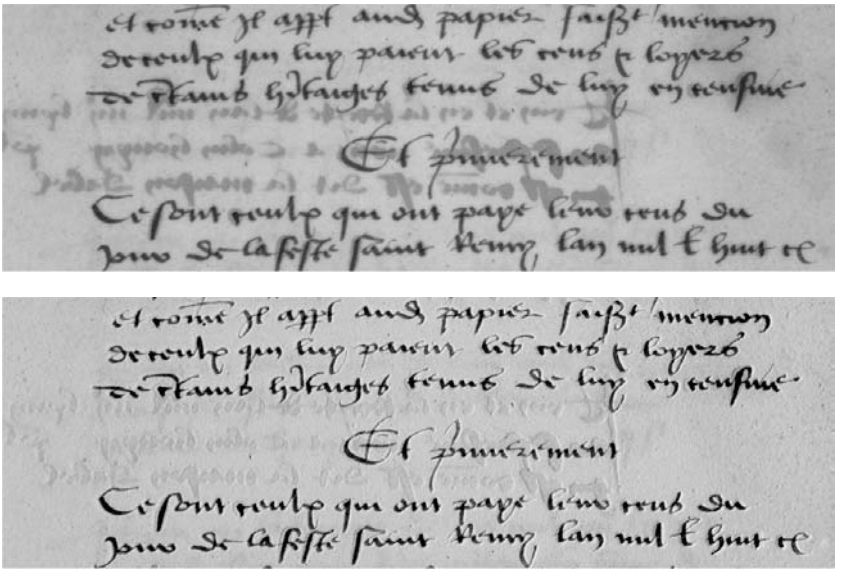

Fig. 3. Top: Scan of an ancient manuscript with bleed-through. Bottom: One of the symmetric orthogonalization outputs from the RGB components.

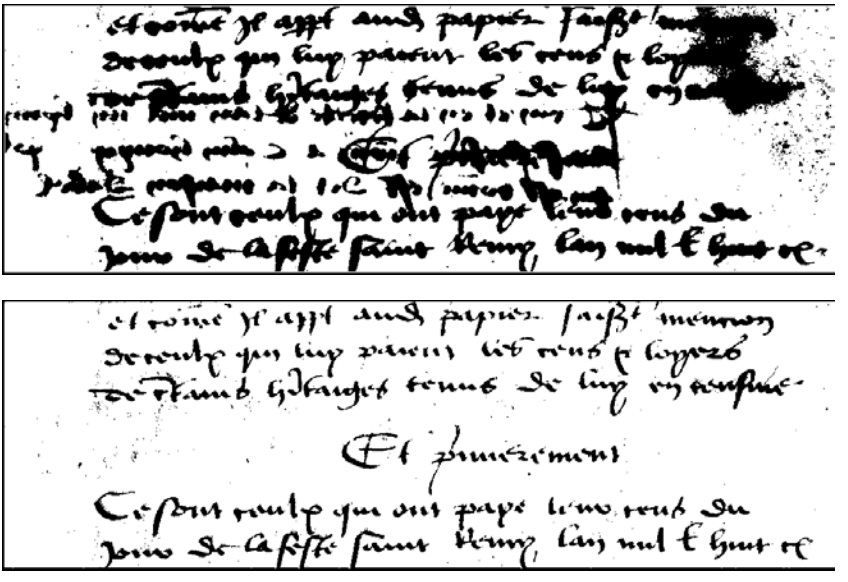

Fig. 4. Top: QIR thresholding on the grayscale original in Fig. 3. Bottom: QIR thresholding on the orthogonalization output in Fig. 3.

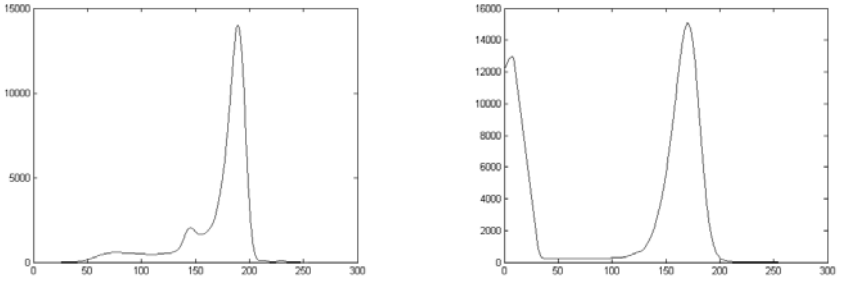

Fig. 5. Left: Histogram of the grayscale version of the original in Fig. 3. Right: Histogram of the orthogonalization output in Fig. 3.

nalization output. As is seen, optimal thresholding has been much more effective on the orthogonalized image. A reason for this can be found by observing the histograms in Figure 5. The sensible peak of the histogram of the graylevel original (on the left) at about level 150 is due to bleed-through. This peak does not appear in the histogram of the processed image, shown in the right-hand panel of Figure 5. The optimal threshold established by the QIR procedure is thus able to avoid artifacts due to bleed-through in the binarized image. Figure 6 shows an example where symmetric orthogonalization achieved full separation of the two overlapped texts, the main foreground text and the bleed-through pattern, which

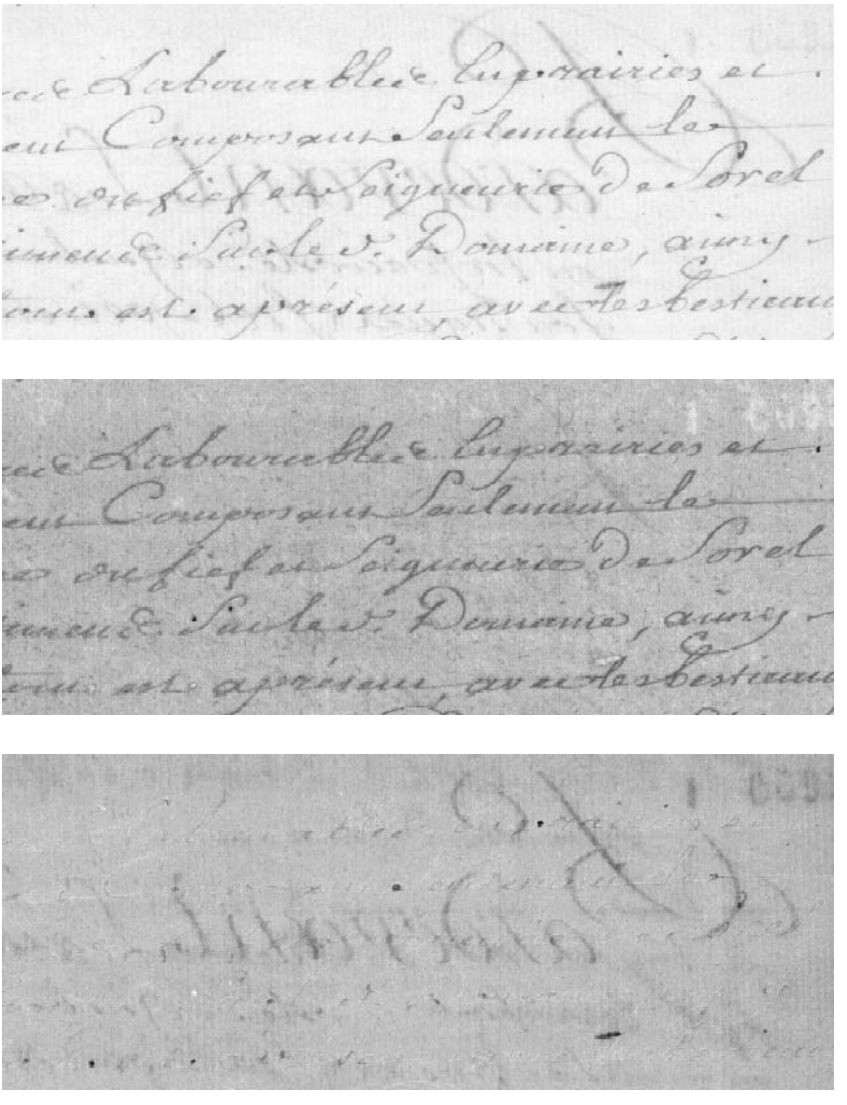

Fig. 6. Top: Scan of an ancient manuscript (from: http://www.site.uottawa.ca/~edubois/documents). Middle and Bottom: Main text and bleed-through patterns extracted through symmetric orthogonalization of the RGB channels.

was not possible via optimal thresholding. As a last example, we show a case where an effective bleed-through cancellation has been added with a focusing of the main text pattern. In Figure 7, the grayscale original and the processed image of an ancient printed page are shown. In this case, bleed-through is weaker than in the cases presented above. It can be seen, however, that the main text appears more focused in the processed output than in the original.

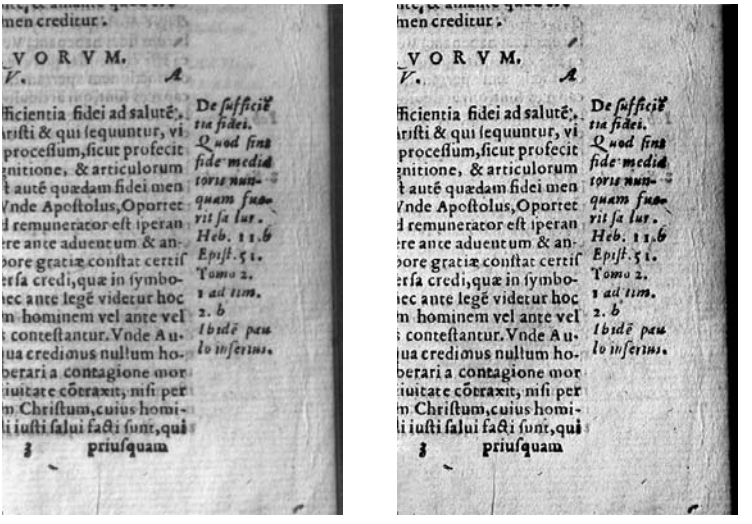

Fig. 7. Left: Scan of an ancient printed page with bleed-through. Right: One of its symmetric orthogonalization outputs.

5 Conclusions

We demonstrated a strategy that has proved to be effective in bleed-through cancellation from color or hyperspectral scans of degraded documents. The advantage of this technique over many other strategies is that it is quite simple and fast, and it does not require reverse side scans or registration prior to cancellation. Our approach lacks a true theoretical justification, however, an extensive experimentation has shown that one of the output channels from symmetric orthogonalization is always a more or less "clean" image of the main text pattern in the original document. Moreover, the application of other standard enhancement techniques is very effective if performed after orthogonalization. More general and quantitative results could be obtained from an experimentation with ground truth data available, that is, with synthetic images. However, a necessary step towards this goal is the development of an accurate numerical model for the bleed-through interference.

Another necessary step towards a complete evaluation of this restoration strategy is to assess the potentiality of using more channels than the common RGB components from data collected by a color camera. In particular, it would be interesting to know which is the optimum number of data images. A viable approach to this problem could be the analysis of the spectra of the data covariance matrices [3]. However, we do not expect that the number of significant data images is independent of the type of document, and we have already evidence towards this conclusion. A complete assessment will be made as soon as a richer database of hyperspectral images will be available to us.

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 (Hengelo, The Netherlands), Art Conservation (Vlaardingen, The Netherlands), Transmedia (Swansea, UK), Atelier Quillet (Loix, France), Acciss Bretagne (Plouzane, France), ENST (Brest, France), CNR-ISTI (Pisa, Italy), CNR-IPCF (Pisa, Italy).

References

1. van Assche, S., Denecker, K. N., Philips, W. R., Lemahieu, I., Proc. SPIE **3653** (1998) 1376–1383.
2. Chew, L.T., Cao, R., Peiyi, S.: IEEE Trans. Pattern Analysis and Machine Intelligence **24** (2002) 1399–1404.
3. Cichocki, A., Amari, S.-I.: Adaptive Blind Signal and Image Processing. (2002) Wiley, New York.
4. Dubois, E., Pathak, A.: Proc. IS&T Image Processing, Image Quality, Image Capture Systems Conf. (Montreal, Canada, 2001) 177–180.
5. Easton, R.L.: Simulating Digital Image Processing used for the Archimedes Palimpsest (2001) http://www.cis.rit.edu/people/faculty/easton/k-12/index.htm
6. Hyvärinen, A., Oja, E.: Neural Networks **13** (2000) 411–430.
7. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. (2001) John Wiley, New York.
8. Liang, Y., Simoncelli, E. P., Lei Z.: Computer Vision and Pattern Recognition (CVPR'00 - Hilton Head, South Carolina, 2000) 1606.
9. Nishida, H., Suzuki, T.: Proc. 16th Conf. Pattern Recognition (Quebec City, Canada, 2002).
10. Ohta, Y., Kanade, T., Sakai, T.: Computer Graphics, Vision, and Image Processing, **13** (1980) 222–241.
11. Sharma, G.: IEEE Trans. Image Processing **10** (2001) 736–754.
12. Solihin, Y., Leedham, C. G.: IEEE Trans. PAMI **21** (1999) 761–768.
13. Tonazzini, A., Bedini, L., Salerno, E.: Int. J. Document Analysis and Recognition (2004) in press.
14. Vertan, C., Boujemaa, N.: Proc. Int. Conf. on Pattern Recognition ICPR'00 (Barcelona, Spain, 2000) 3584–3587.