Color Space Transformations for Analysis and Enhancement of Ancient Degraded Manuscripts[1]

A. Tonazzini

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

e-mail: anna.tonazzini@isti.cnr.it

Abstract—In this paper we focus on ancient manuscripts, acquired in the RGB modality, which are degraded by the presence of complex background textures that interfere with the text of interest. Removing these artifacts is not trivial, especially with ancient originals, where they are usually very strong. Rather than applying techniques to just cancel out the interferences, we adopt the point of view of separating, extracting and classifying the various patterns superimposed in the document. We show that representing RGB images in different color spaces can be effective for this goal. In fact, even if the RGB color representation is the most frequently used color space in image processing, it does not maximize the information contents of the image. Thus, in the literature, several color spaces have been developed for analysis tasks, such as object segmentation and edge detection. Some color spaces seem to be particularly suitable to the analysis of degraded documents, allowing for the enhancement of the contents, the improvement of the text readability, the extraction of partially hidden features, and a better performance of thresholding techniques for text binarization. We present and discuss several examples of the successful application of both fixed color spaces and self-adaptive color spaces, based on the decorrelation of the original RGB channels. We also show that even simpler arithmetic operations among the channels can be effective for removing bleed-through, refocusing and improving the contrast of the foreground text, and to recover the original RGB appearance of the enhanced document.

Key words: document Analysis and Enhancement, Color Space Transformations, Color Channel Decorrelation.

DOI: 10.1134/S105466181003017X

1. INTRODUCTION

Conservation, readability and interpretation of ancient manuscripts is often compromised by several and different damages that they have undertaken over time, and that continue to cause a progressive decay. In this way, we undergo the risk to lose much of our past memory during the next years. Furthermore, the fragility of rare or very important historical documents prevents their direct access by scholars and historians. Natural ageing, usage, poor storage conditions, humidity, molds, insect infestations and fires are the most diffuse degradation factors. In addition, the materials used in the original production of the documents, i.e. paper or parchment and inks, are usually highly variable in consistency and characteristics. All these factors concur to cause ink diffusion and fading, blurred or unfocused writings, seeping of ink from the reverse side (bleed-through effect), transparency from either the reverse side or from subsequent pages (show-through effect), spots, noise, low contrast of the characters with respect to the background, faint, fragmented or joined characters. Furthermore, these defects are usually varying across the document. These problems are common to the majority of the governmental, historical, ecclesiastic and commercial archives in Europe, so that seeking out for a remedy would have an enormous social and technological impact. Digital imaging can play a fundamental role in this respect. Indeed, it is an essential tool for generating digital archives in order to ensure the documents accessibility and conservation. Moreover, OCR processing for automatic transcription and indexing facilitates the access to the digital archives and the retrieval of information.

Often, these images are acquired only in the visible range of the spectrum, due to the larger diffusion of the dedicated acquisition equipments. However, owing to specific damages, some documents may be very difficult to read when acquired in the visible range. This particularly concerns documents produced during the XVI and XVII centuries, due to the corrosion, fading, seeping and diffusion of the ink used (iron-gall mostly), and those produced even more recently, due to the bad quality of the paper that started being used after the XIX century. Furthermore, interesting features are often barely detectable in the original color document, while revealing the whole contents is an important aid to scholars that are interested in dating or establishing the origin of the document itself, or

[1]The article is published in the original.

Received March 24, 2010

reading hidden text it may contain. Thus, additional information can sometimes be obtained from images taken at non-visible wavelengths, for instance in the near infrared and ultraviolet ranges. Alternatively, or in conjunction with multispectral/hyperspectral acquisitions, digital image processing techniques can be used for enhancing the readability of the document contents and seeking out new information.

In this paper we focus on ancient manuscripts, acquired in the RGB modality, which are degraded by the presence of complex background textures interfering with the foreground text. Often these artifacts are very strong, so that removing them is not simple. For instance, dealing with strong bleed-through degradation is practically impossible by any standard thresholding technique, since the intensities of the unwanted background can be very close to those of the main text. Thus, adaptive and/or structural approaches have to be adopted [11, 20]. Work done on the specific problem of bleed-through/show-through removal has mainly exploited information from the grayscale front and back pages, usually referred as recto and verso [1, 6, 18]. Besides requiring a preliminary registration of the two sides, these techniques are usually expensive, as they are based on steps of segmentation, to identify the bleed-through areas, followed by inpainting of estimated pure background areas [3]. More recently, variational approaches, based on nonlinear diffusion, have been proposed to model and then remove this kind of degradations [13]. In [15], only a color scan from a single side is required, but a thresholding technique based on multiresolution analysis and adaptive binarization must be employed. Other authors propose segmentation of the different color clusters in the image via adaptation of the k-means algorithm [4, 10]. In [26], for the grayscale scan of a single-sided document, a classification approach based on a double binary MRF, one for the recto pattern and another for the verso pattern, is proposed.

The main disadvantages of most of the techniques proposed so far can be summarized as follows: (i) usually, only grayscale documents can be treated; the availability of the scan of the verso side is often required; (iii) a preliminary registration of the two sides is necessary; this is presently a difficult task, since the common parts of the two images to be aligned are sparse and different in intensity; (iv) the application, in cascade, of several techniques is needed; (v) the computation times are usually high and depend on the document size; (vi) the correct tuning of several parameters is required; (vii) all the structured background is removed in order to obtain a clean foreground text; this causes the loss of the original appearance of an old manuscript or document, and of all those marks that often contain the imprints of its history and authenticity.

Rather than applying techniques to indiscriminately cancel out the interferences in the background, we adopt the point of view of separating, extracting and classifying the different patterns superimposed in the document. In this way, the artifacts are removed in a selective way, depending on their origin. Specific situations in which the interfering textures, or some of them, can be of interest per se, such as underwritings in palimpsests, stamps, or paper watermarks, can be addressed. These patterns, often representing the most significant information from a cultural and historical point of view, can be enhanced and recovered. We show that representing RGB images in different color spaces can be an efficient tool to reach this scope. Indeed, even if the RGB color representation is the most frequently used color space in image processing, it presents some limitations in terms of maximization of the image information contents. Hence, in the literature, many different color spaces have been developed for different image analysis tasks, such as object segmentation and edge detection [25]. We will show that some of them are particularly suitable for the analysis of degraded documents, often allowing for the enhancement of the document contents, the improvement of the text readability, and the extraction of partially hidden features. In addition, the methods we are going to propose are very fast, almost independent of the document size, and do not require the setting of parameters. They only require the RGB scan of a single side, so that registration is not necessary, and the color appearance of the original document can often be recovered for the restored document as well.

The paper is organized as follows. In Section 2, we will analyze the performance of fixed color spaces previously proposed in the literature, and of arithmetic operations between the channels. In Section 3, a technique for the blind, self-adaptive projection of a document image into document-dependent color spaces will be described. Section 4 will be devoted to the rendering of the original color of an enhanced document, and, finally, Section 5 will conclude the paper with comments and perspectives.

2. FIXED COLOR SPACES AND ARITHMETIC OPERATIONS BETWEEN CHANNELS

A reversible projection of an RGB image into a different color space amounts to the operation of multiplying the Red, Green and Blue components by an invertible matrix. Any invertible matrix defines a color space, where the new colors are weighted additive mixtures of the original RGB channels. It is intuitive that, since the RGB components contain the overlapped patterns in different percentage, simple difference operations between the colors, after suitable regulation of the levels, can cancel one pattern and enhance the other. In the following, we will review some of the most common color spaces proposed in the literature, with the aim of highlighting and explaining their ability to perform reduction of artifacts in ancient degraded manuscripts.

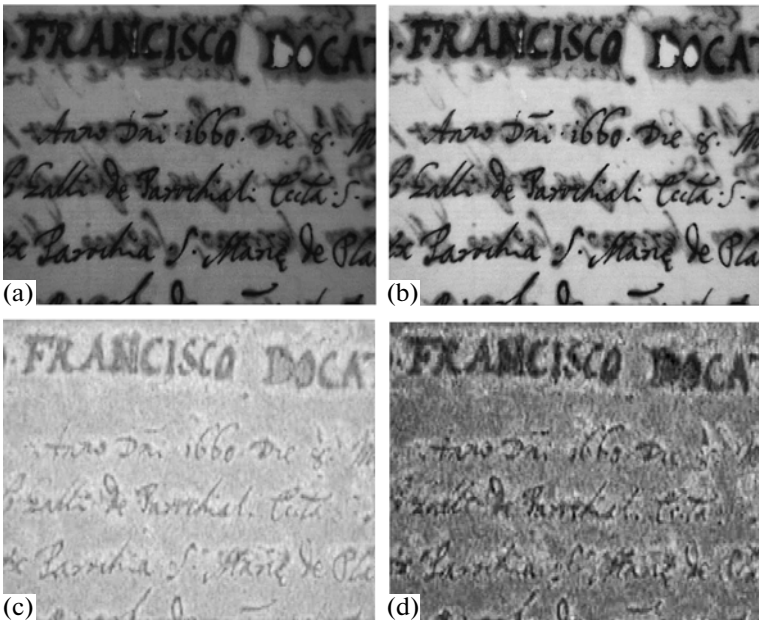

Fig. 1. YES color space applied to a manuscript (iron-gall ink, XV century): (a) RGB image; (b) *Y* channel; (c) *E* channel; (d) *S* channel.

2.1. The YES Color Space

In 1989, Xerox Corporation proposed a color encoding standard called YES [27]. The YES color space is a linear transformation of the RGB vector that matches the physiology of the human visual system. YES provides a separation of the color and intensity information. The three coordinates are an achromatic luminance channel, that is a weighted sum of the RGB values, called *Y*, and two opponent-color chromatic coordinates given by spectral differences: the *E* channel is proportional to Red minus Green, while the *S* channel is proportional to Yellow minus Blue. The RGB-to-YES transformation is:

$$\begin{bmatrix} Y \\ E \\ S \end{bmatrix} = \begin{bmatrix} 0.253 & 0.684 & 0.065 \\ 0.5 & -0.5 & 0 \\ 0.25 & 0.25 & -0.5 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}. \quad (1)$$

YES has been specifically employed for the enhancement of degraded ancient manuscripts. In particular, when imaging the Dead Sea Scrolls, Keith Knox et al. (1997) [7] found that in the *E* map the contrast was significantly augmented, and hidden characters were revealed. Other authors claim that subtracting the Green from the Red is able to reveal hidden characters in charred documents, with performance similar to that obtained through the acquisition in the near infrared band [19].

We experimentally found that the YES color space, and in particular the Red minus Green operation, is useful also for removing bleed-through in reddish documents. Figure 1 shows the YES components of a manuscript affected by a strong bleed-through. The reddish appearance of this document is due to oxidation of the iron-gall ink used. We can appreciate the extraction of a clean foreground text pattern in the *E* image and in the *S* image, though this latter is worse. The *Y* image, instead, is similar to the gray level version of the original color image, as expected. We tested the same techniques on other reddish images from our data set, and found similar performances. The rationale for this will be given later on in this section. However, we expect (and indeed verified) a dependence on the document color.

2.2. The OHTA Color Space

The OHTA color space has been derived by Ohta et al. [16] in 1980 as a way to approximate the Principal Component Analysis (PCA) of the RGB components of a color image. The fixed coefficients of the OHTA matrix were experimentally found by a statistical study of the uncorrelated color components on a large population of images of typical real-world scenes. The three coordinates are an achromatic luminance channel, that is a homogeneous weighted sum of the RGB values, called *O*, and two chromatic coordinates given by spectral differences: the *H* channel is proportional to Red minus Blue, while the *T* channel is proportional to Green minus Magenta. The transformation is given by the following equation:

$$\begin{bmatrix} O \\ H \\ T \end{bmatrix} = \begin{bmatrix} 0.33 & 0.33 & 0.33 \\ 0.5 & 0 & -0.5 \\ -0.25 & 0.5 & -0.25 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}. \quad (2)$$

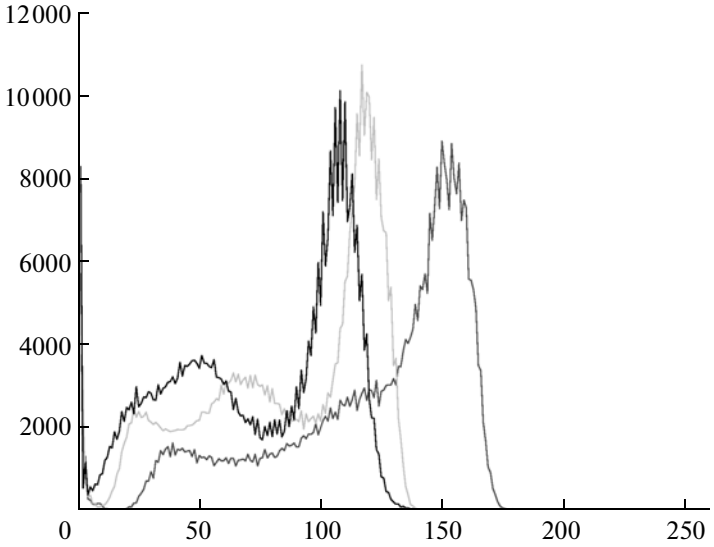

Fig. 2. Histogram of the color manuscript in Fig. 1.

The results obtained with the OHTA color space transformation are quite similar to those that we obtained with the YES representation. For instance, when applied on the manuscript of Figure 1(a), OHTA produces a H image that represents a good enhancement of the foreground text. This time, the performing operation is the subtraction between Red and Blue, which, again, works well on most reddish documents.

To explain why Red-Green and Red-Blue work on this kind of images, let us consider the histogram of the original color image in Fig. 1 (see Fig. 2). From this histogram, we can observe that in the background/bleed-through areas Red and Green (or Red and Blue) are well separated, i.e. their difference is high. Thus, Red-Green/Red-Blue returns almost equal, high values for both the background and the bleed-through pixels, so that these merge; conversely, the same operation gives much lower values for the text, that results enhanced.

2.3. The CMYK Color Space

The CMYK color space (Cyan, Magenta, Yellow, and BlacK) [5, 17, 24] is a variation on the CMY model and is commonly used in color printers, due to the subtractive properties of inks. Cyan, Magenta, and Yellow are the complements of Red, Green, and Blue, respectively, and are subtractive primaries because their effect is to subtract some color from white light.

However, experience with various types of inks and papers has shown that when equal high values of Cyan, Magenta, and Yellow inks are mixed, the result is usually a dark brown, but not a true black. Adding some black to the mixture solves this problem. The specific black map to be used for this purpose is called BlacK (K) and is computed as the minimum among Cyan, Magenta and Yellow. The basic transformation RGB-to-CMYK is given by the following equations:

$$\begin{aligned} C &= 1 - R \\ M &= 1 - G \\ Y &= 1 - B \\ K &= \min(C, M, Y), \end{aligned} \tag{3}$$

where the RGB data are gamma-corrected prior to converting them to the CMY color space. However, more accurate transformations can be used to account for the dependency of the inks and paper used. Slight adjustments, such as mixing the CMY data, are usually necessary, according to:

$$\begin{bmatrix} C \\ M \\ Y \end{bmatrix} = \begin{bmatrix} m_1 & m_2 & m_3 \\ m_4 & m_5 & m_6 \\ m_7 & m_8 & m_9 \end{bmatrix} \begin{bmatrix} 1-R \\ 1-G \\ 1-B \end{bmatrix}, \tag{4}$$

where coefficients m_1, m_5, and m_9 are values near unity and the other coefficients are small in comparison.

We have found that subtracting the K channel from a suitably chosen grayscale component (either one of the RGB maps, or a map from another representation) of a document containing a dark text is often effective to increase the contrast. Indeed, when the text appears as black or dark, and the background lighter, the K channel is light in the foreground text pixels and dark in the background pixels. Thus if we subtract pixel by pixel the K channel from the original image, and clip to zero all the values resulting negative, we obtain an image where the text is darker. Conversely, in the light background this operation subtracts low values from high values, so that the new background remains almost the same, i.e., light. In this way, the text is

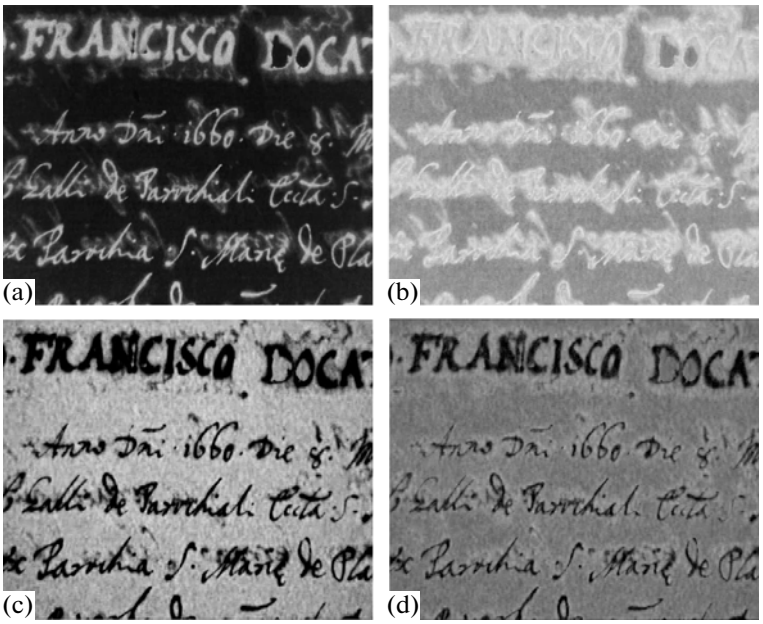

Fig. 3. Subtraction of the *K* channel of the CMYK color space: (a) *K* channel of the RGB image in Fig. 1a; (b) Magenta channel of the same image; (c) *E* channel of the YES representation of the same image after subtraction of the *K* channel; (d) Magenta channel after subtraction of the *K* channel.

enhanced and better contrasted with respect to the background.

Besides its use for contrast improvement and text refocusing in low quality document images, when no artifacts are present in the background, this simple enhancement technique can also be used for the further improvement of document already restored through some technique of artifacts reduction. This latter application is illustrated in Fig. 3c, where the *K* subtraction is applied to the text pattern extracted from the original RGB document of Fig. 1a, by means of the YES or OHTA transformations. The contrast improvement is apparent. The same technique has provided some good result also when applied in presence of bleed-through. For example, Fig. 3d shows the effect of subtracting the *K* channel from the Magenta channel. Here, the performing cause can be easily found looking at the intensity values of pixels belonging to text, bleed-through and background areas, in the Magenta and *K* maps, respectively. In particular, the equal intensity of the main text pixels in both the Magenta and *K* map returns almost zero, i.e. dark, from their difference.

2.4. The YCbCr Color Space

Y'CbCr or YCbCr is a color space that provides a luminance channel, the *Y'* channel (luma), and two chrominance channels, *Cb* and *Cr*, respectively. Luma represents the brightness in an image, i.e. the achromatic image without any color, while the chroma components represent the color information. Converting RGB sources into luma and chroma allows for chroma subsampling. Indeed, since the human vision system is more sensitive to luminance detail than color detail, video systems can optimize bandwidth for luminance over color. The transformation is given by the following equations:

$$\begin{aligned} Y' &= K_r * R' + (1 - K_r - K_b) * G' + K_b * B', \\ C_b &= \frac{0.5}{1 - K_b}(B' - Y'), \\ C_r &= \frac{0.5}{1 - K_r}(R' - Y'), \end{aligned} \tag{5}$$

where R', G' and B' are the gamma-compressed color components, and K_r and K_b are coefficients in [0, 1].

Suppose that the interferences to be removed have a color that is different from that of the main text, usually dark, but the same color, possibly more intense, of the background. In other words, the interference pixels will have luminance values that are different from those of the main text and the background, and chrominance values that are different from those of the main text but similar to those of the background. In these conditions, the *Cb* and *Cr* maps could represent themselves an enhanced version of the document. Figure 4 shows the original RGB and the negative of the *Cb* map of a manuscript. As it can be appreciated, the *Cb* map is a quite clean image of the main text alone, although the contrast is very poor.

In Section 4, we will show that the chrominance channels of the YCbCr representation can be used also to recover the original color appearance of a restored, grayscale document.

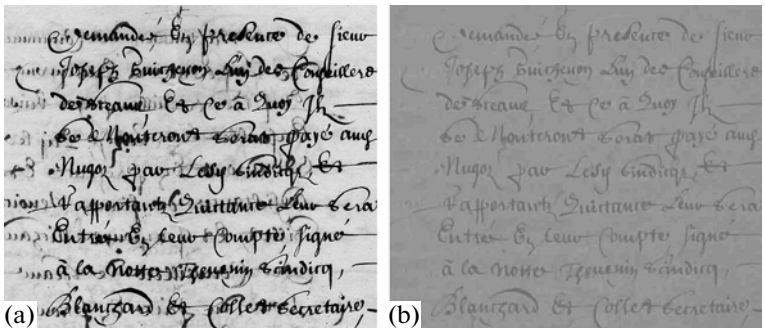

Fig. 4. YCbCr representation: (a) original RGB manuscript; (b) the Cb channel (negative).

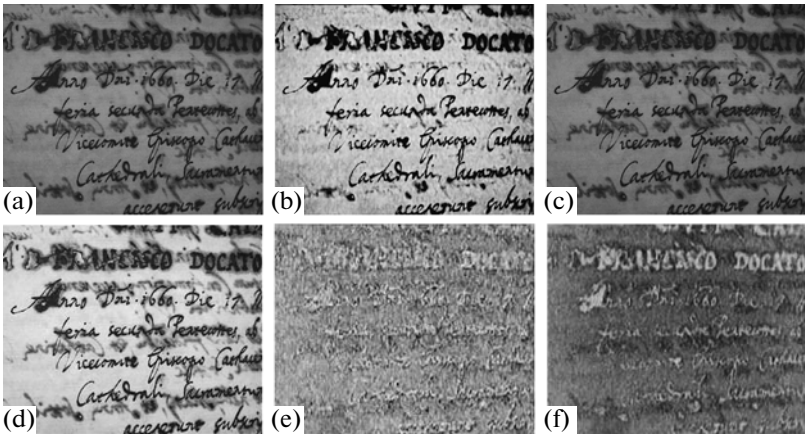

Fig. 5. YES applied to a bluish manuscript: (a) the original reddish manuscript; (b) the E channel of the reddish manuscript; (c) the bluish manuscript; (d) the Y channel of the bluish manuscript; (e) the E channel of the bluish manuscript; (f) the S channel of the bluish manuscript.

2.5. Fixed Color Spaces Versus the Document Color

We inferred that the success of the YES and OHTA color transformations, at least for removing front-to-back interferences, depends on the color of the document. This was confirmed by a simple experiment. Figure 5 shows another reddish manuscript, and its bluish version obtained by suitably rearranging the RGB components of the original image. While YES applied to the reddish document provides the E map shown in Fig. 5b, the YES components of the bluish document are clearly useless. In particular, observe that no bleed-through removal is achieved in this case.

Another example of failure of the YES or OHTA representation due to the unsuitability of the color of the manuscript at hand is shown in Fig. 6. The main problem with this manuscript was still bleed-through.

In Section 3 we will explain our idea for attempting separation of superimposed patterns in documents of any color, provided that some hypotheses are verified. The approach is based on the principle of decorrelating the color components, as for the OHTA color space. However, in our case, the method is fully blind and adaptive, in the sense that the matrix to be used for the transformation is jointly estimated with the patterns, and hence self-adaptive to the document image at hand. Furthermore, the method can be applied to any number of color components, for example it can exploit possible available views in non-visible bands, and, at least in principle, can be used to separate more than three overlapped patterns.

3. ADAPTIVE COLOR SPACES

The main idea behind the enhancement technique we are going to explain for RGB images, or its generalization to a lower/higher dimension, is that while the color components of an image are usually spatially correlated, the individual patterns (or classes, or sources) superposed into the image are usually much less correlated. Hence, decorrelating the color components gives a different representation where the now orthogonal components of the image could coincide with the single classes. It is worth to say that our problem can be seen as a particular instance of a more general class of problems in signal and image processing, namely, Blind Source Separation (BSS) [2].

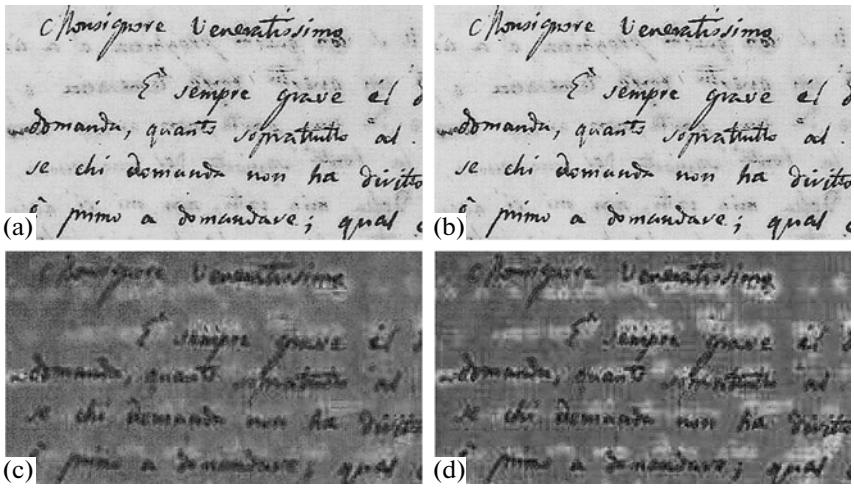

Fig. 6. Example of failure of the YES representation: (a) RGB image; (b) the Y channel; (c) the E channel; (d) the S channel.

3.1. The Data Model

We assume that the multispectral scan of a document has a vector value $\mathbf{x}(i)$ of N components, where t is the pixel index, and the total number of pixels in the images is T. Similarly, we 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, in its pristine state, i.e. undegraded, is almost uniform. We call A_{ij} the mean reflectance index for the j-th source at the i-th observation channel. Thus, 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. In formulas, it is:

$$\mathbf{x}(t) = A\mathbf{s}(t) \quad t = 1, 2, \ldots, T, \tag{6}$$

where the $N \times M$ matrix A has the meaning of a mixing matrix. A special instance of the model arises when considering the document as constituted of three only sources and three observations. In most ancient documents, indeed, we have a main text superimposed with textures coming from the support and from just another underlying pattern, due to bleed-through/show-through, or watermarks, or old erased texts, or whatever. Similarly, in the most frequent situations, we have the RGB view of the document, which can be always split in the Red, Green and Blue components. The reflectance indices $(x_r(t), x_g(t), x_b(t))$ of a generic pixel t of the document can be seen as given by the following equation:

$$\begin{bmatrix} x_r(t) \\ x_g(t) \\ x_b(t) \end{bmatrix} = \begin{bmatrix} r_1 & r_2 & r_3 \\ g_1 & g_2 & g_3 \\ b_1 & b_2 & b_3 \end{bmatrix} \begin{bmatrix} s_1(t) \\ s_2(t) \\ s_3(t) \end{bmatrix} \quad t = 1, 2, \ldots, T, \tag{7}$$

where, in this notation, $s_i(t)$, $i = 1, 2, 3$, denote respectively the "quantity" of background, bleed-through and foreground that concur to form the color at pixel t. The mean reflectance indices (r_i, g_i, b_i), $i = 1, 2, 3$, are the entries of the mixing matrix A. Since, in principle, the reflectance indices are unknowns, they need to be estimated jointly with the ideal sources from the data alone.

It is easy to see that this model does not perfectly account for the phenomenon of interfering texts in documents, related to complicated chemical processes of ink diffusion and paper absorption. For instance, it does not account for color saturation in correspondence of occlusions. This effect could be better described by a nonlinear model, as done in [18] for the show-through. In addition, ink diffusion or light spreading through the support can produce a sensible blur effect, so that the true source patterns superimposed in the document should be modelled as blurred version of the ideal ones. However, we can reasonably assume that the linear approximation holds when the patterns do not occlude each other, and that the occlusion areas are a few. Similarly, it is reasonable to assume that each source pattern is affected by its own blur operator that is approximately the same for a given source in the different observations. At this stage, we then assume that blur can be disregarded.

3.2. Solution Through Decorrelation of the Color Channels

Under the fundamental hypothesis of a non-singular mixing matrix, which means linear independence of the views and, in our case, different colors for the different patterns, it has been proved that blind separation can only be achieved by assuming the mutual statistical independence of the sources themselves. Independent Component Analysis (ICA) techniques have been developed for finding a linear transformation W that, when applied to the observations, produces a set $\mathbf{y}(t) = W\mathbf{x}(t)$ of independent sources [9]. If the mixed

patterns are really independent, the estimated sources $\mathbf{y}$ are replicas of the true patterns $\mathbf{s}$, and W is the inverse of the mixing matrix A. To force statistical independence, the cross-central moments of all orders between each pair of estimated sources must be constrained to zero. Although based on an oversimplified model, we satisfactorily applied ICA algorithms for document analysis [21].

In principle, no source separation can be obtained by only constraining second-order statistics, i.e. by enforcing uncorrelation alone. However, decorrelation has several advantages over ICA in our application: (i) it is always less computationally expensive than most ICA algorithms, and no parameters need to be set; (ii) in some cases, it is able to separate as well [2], and we experimentally verified that, for our document images, often it outperforms ICA; (iii) our main aim is not full separation but interference reduction, and this can sometimes be achieved even if only one of the overlapped patterns possesses the desired properties. We then propose herein to perform separation based on second-order statistics.

To enforce statistical uncorrelation on the basis of the available color channels, we must estimate the sample covariance matrices and diagonalize them. This is equivalent to orthogonalize the different color channels. The result of orthogonalization is of course not unique. We experimentally tested the performances of two different strategies, namely Principal Component Analysis (PCA) and symmetric whitening or symmetric orthogonalization [23]. More formally, our data covariance matrix is the 3×3 matrix:

$$R_{\mathbf{xx}} = \langle \mathbf{x}\mathbf{x}^T \rangle, \tag{8}$$

where the superscript T means transposition. Since we do not have the probability density function of vector $\mathbf{x}$, we approximate an estimate of the covariance matrix from the RGB components of our image:

$$R_{\mathbf{xx}} \approx \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}(t)\mathbf{x}^T(t). \tag{9}$$

Since the data are normally correlated, matrix $R_{\mathbf{xx}}$ will be nondiagonal. Let us now perform the eigenvalue decomposition of matrix $R_{\mathbf{xx}}$

$$R_{\mathbf{xx}} = V_{\mathbf{x}}\Lambda_{\mathbf{x}}V_{\mathbf{x}}^T, \tag{10}$$

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. It is easy to verify that the two following choices for W yield a diagonal $\mathbf{R}_{\mathbf{yy}}$:

$$W_o = V_{\mathbf{x}}^T, \tag{11}$$

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

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, and whose Euclidean norms are equal to the eigenvalues of the data covariance matrix. Indeed, it is:

$$R_{\mathbf{yy}} = W_o R_{\mathbf{xx}} W_o^T = V_{\mathbf{x}}^T V_{\mathbf{x}}\Lambda_{\mathbf{x}}V_{\mathbf{x}}^T V_{\mathbf{x}} = \Lambda_{\mathbf{x}}. \tag{13}$$

This is what PCA does [2]. By using matrix W_s, we have:

$$\begin{aligned} R_{\mathbf{yy}} &= W_s R_{\mathbf{xx}} W_s^T \\ &= V_{\mathbf{x}}\Lambda_{\mathbf{x}}^{-\frac{1}{2}}V_{\mathbf{x}}^T V_{\mathbf{x}}\Lambda_{\mathbf{x}}V_{\mathbf{x}}^T V_{\mathbf{x}}\Lambda_{\mathbf{x}}^{-\frac{1}{2}}V_{\mathbf{x}}^T = I_N. \end{aligned} \tag{14}$$

Vectors $y_i(t)$ are thus orthonormal. Note that matrix W_s has the further property of being symmetric, and then its application is equivalent to ICA when the mixing matrix A (see Eq. (6)) is symmetric as well [2].

We observed that PCA never performs full separation on our document images, and rarely produces at least one output where the bleed-through interference is reduced. Conversely, it is interesting to note that symmetric orthogonalization almost always performs bleed-through reduction, and often achieves a full separation of the superimposed patterns.

These techniques, in the practice, amount to different, adaptive color representations of the document, where the new colors are mutually spatially uncorrelated. As an observation, the original RGB representation of the individual recovered sources can also be restored, at least in principle, by exploiting the estimated mixing coefficients. Finally, these techniques can be extended to more than three channels and three patterns, including non visible bands.

3.3. Applications of color Decorrelation to Document Analysis

The self-adaptive color transformation described above can be successfully applied to several instances of document analysis and enhancement, which will be illustrated in the following, through some experiments and examples.

The first two experiments were conceived to show how decorrelation, in a fully blind way, includes and often outperforms fixed color transformations, like YES and OHTA. Let us consider the reddish document where the YES and OHTA transformations produced a good bleed-through removal. When applying the orthogonalization process to this image (Fig. 1a), we obtain the results shown in Fig. 7. It is doubtful in this case whether the process has achieved full separation of the three classes or not, since the second output, that could correspond to the bleed-through pattern, presents strong residuals from the foreground text. However, note how the first output is very similar

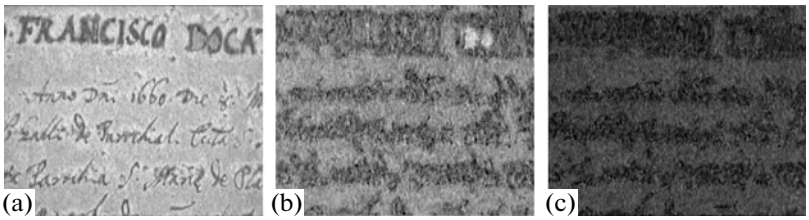

Fig. 7. Symmetric orthogonalization of the RGB components of the image shown in Fig. 1a: (a) first output; (b) second output; (c) third output.

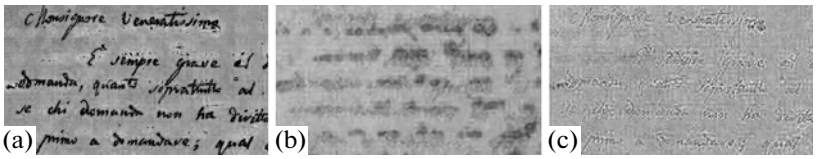

Fig. 8. Symmetric orthogonalization of the RGB image in Fig. 6: (a) first output: (b) second output; (c) third output.

to the E and H images of the YES and OHTA transformations of the same image (see Fig. 1c). Furthermore, it is interesting, in this case, to look at the demixing matrix W estimated by the process:

$$W = \begin{bmatrix} 0.078 & -0.062 & -0.009 \\ -0.062 & 0.163 & -0.079 \\ -0.009 & -0.079 & 0.109 \end{bmatrix}.$$

The first row of this matrix, which gives the first output, clearly performs a Red-Green operation. This is exactly what the E image of the YES space corresponds to. Thus, on this image, we have obtained the same result of the YES space, in terms of bleed-through reduction, in a fully blind way.

At this point, it is immediate to appreciate that, by applying color decorrelation to the bluish manuscript of Fig. 5c, we will obtain the same result of Fig. 5b, i.e. the E map (or, equivalently, the orthogonalized map) of the corresponding reddish manuscript. By looking at the demixing matrix W, we observed that, as expected, the performing operation was not Red minus Green, in this case. Hence, color decorrelation can produce bleed-through reduction in documents of any color, provided that the bleed-through pattern exhibits a spectral diversity with respect to the main text.

The example in Fig. 8 shows the result of the application of symmetric orthogonalization to the RGB image of Fig. 6, for which the YES color space was proven useless.

As it can be clearly appreciated, the decorrelation process produced a quite satisfactory separation of the three overlapped patterns, although the bleed-through is unreadable due to the strong blur, and the background pattern presents some residuals from the other two classes.

From the experiments above, and the many other we executed on a large data set, we can thus conclude that, in general, self-adaptive color decorrelation outperforms fixed color transformations for document enhancement and bleed-through reduction, and, what is more, it can also reach full separation of the overlapped patterns.

In Figs. 9 and 10 we show another couple of examples where the orthogonalization process achieves a complete separation of the overlapped patterns. In the first example of Fig. 9, the three overlapped patterns are background, foreground text, and a stamp in place of the bleed-through pattern. This is barely detectable on the center-right of the RGB image. Note how, in the second output of the orthogonalization process, the stamp is now well visible and cleansed from the other patterns. On the other hand, the foreground text class is refocused and the contrast enhanced. Thus, we can hypothesize that color decorrelation can also be useful for detecting hidden texts/textures, and for augmenting the contrast [22].

In the second example of Fig. 10, the RGB document contained four different patterns. With three views available, we can only hope to extract three individual patterns, and, indeed, Fig. 10b shows one of the three orthogonal components were two patterns remain overlapped (the main handwritten text and the paper watermark). However, it is interesting to note that both the paper watermark and the typewritten text were not visible in the original RGB image. If another acquisition channel was available, e.g. a non-visible one, perhaps it could even be possible to separate handwriting and watermark, provided that these pos-

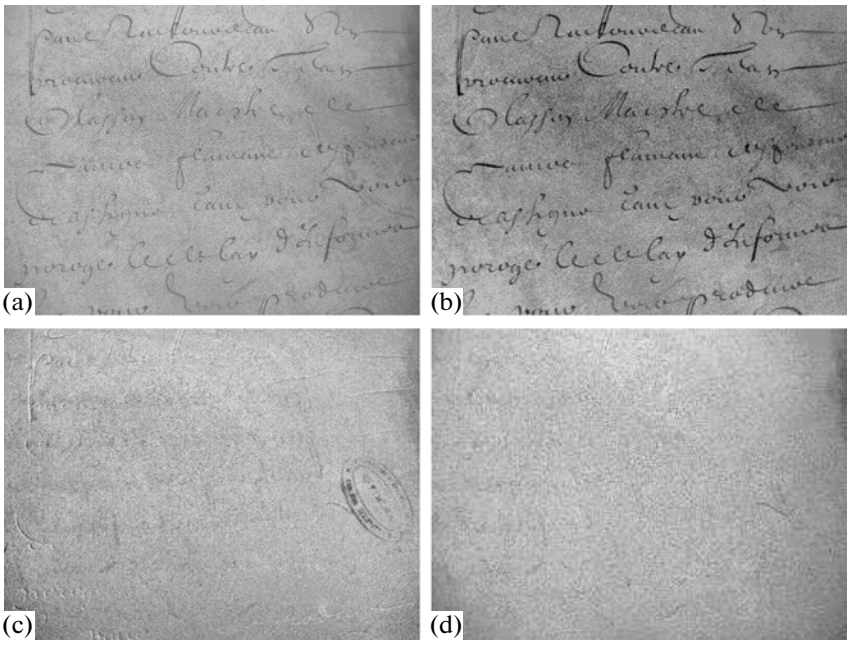

Fig. 9. Symmetric orthogonalization: (a) RGB image; (b) the written text; (c) the stamp; (d) the background.

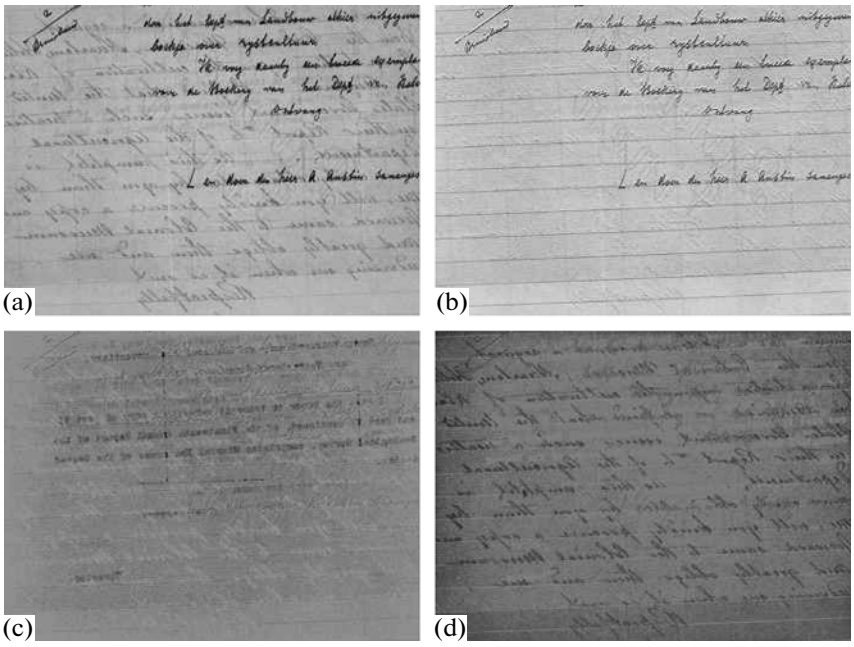

Fig. 10. Symmetric orthogonalization: (a) RGB image; (b) the main handwritten text plus an enhanced paper watermark; (c) an enhanced typewritten text; (d) the show-through text.

sess a genuine spectral diversity, and augmenting to four dimensions the model in Eq. (7).

Despite the two examples above, it is clear that color decorrelation, as well as fixed color space transformations, do not always produce a good separation of all the classes. This may be due to the actual correlation of the classes themselves, the lack of spectral diversity of the inks, and certainly the rough linear approximation for the model adopted. However, very often each orthogonalized image shows one of the classes predominant over the others. In this way we can use some techniques, such as optimal thresholding or further arithmetic operations between the orthogonal-

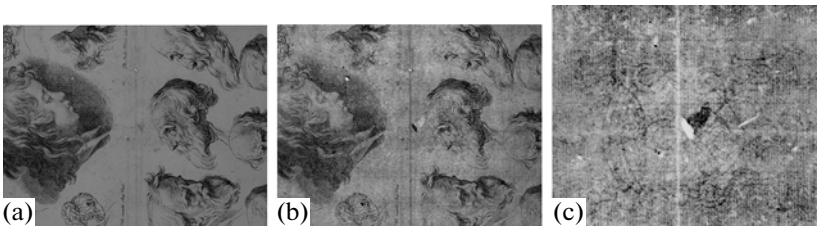

Fig. 11. Extraction of a watermark from an IR pair: (a) front and back illumination; (b) back illumination; (c) a detail of one output of symmetric orthogonalization.

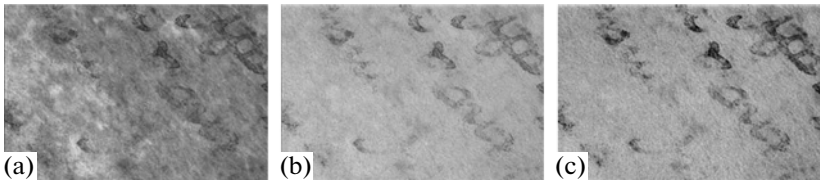

Fig. 12. Comparison between IR imaging and orthogonalization of the RGB channels: (a) RGB image; (b) IR image; (c) one output of symmetric orthogonalization of the RGB components of the image.

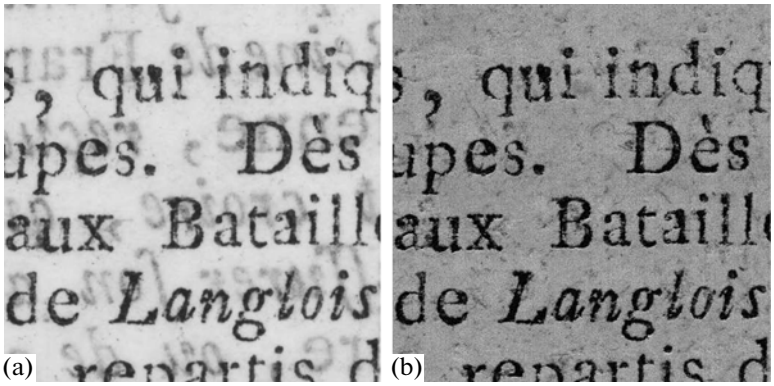

Fig. 13. Subtraction of the *K* channel of the CMYK color space from a decorrelated image: (a) original RGB image; (b) enhanced image.

ized channels, to obtain the extraction of the class that we want to enhance.

The already mentioned ability of color decorrelation in detecting hidden texts or textures, as well as its applicability to multisensor/multiview images, is illustrated in the example of Fig. 11, where a pair of infrared (IR) images of a drawing were acquired, under different illumination conditions. The aim was to highlight the paper watermark. This is barely visible in the back illumination modality, but is very faint and still overlapped to the drawings on the paper. A clean map of the watermark alone can be obtained by applying symmetric orthogonalization to the infrared pair.

We showed above that the subtraction between Red and Green, weighted by suitable coefficients, is very useful for removing complex backgrounds from reddish documents. We also mentioned that this operation has been previously used to enhance historical manuscripts. In particular, Knox et al. [7], by highly stretching the contrast of the difference between the Red and Green separations of the color image, were able to reveal characters never seen before in the Dead Sea Scroll. The mechanism behind this method is still not clear to Knox and co-authors. The best explanation they could determine is that the Red separation may have recorded some infrared information, which would make characters more visible in a damaged region of the parchment [8]. Since color decorrelation also performs weighted arithmetic operations between the color channels, along the hypothesis above, in Fig. 12 the IR image of a document with very faint characters is compared with the result of applying symmetric orthogonalization to the RGB components of the visible image.

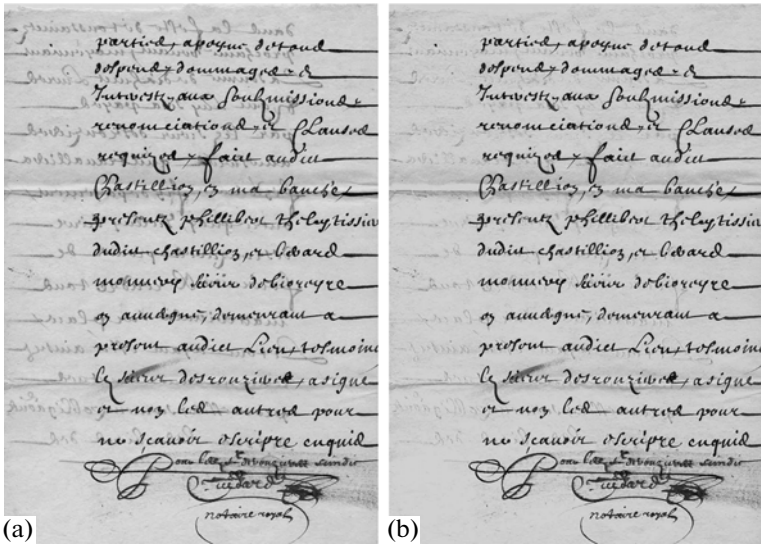

Fig. 14. Reduction of bleed-through preserving the original document appearance: (a) original image; (b) enhanced image.

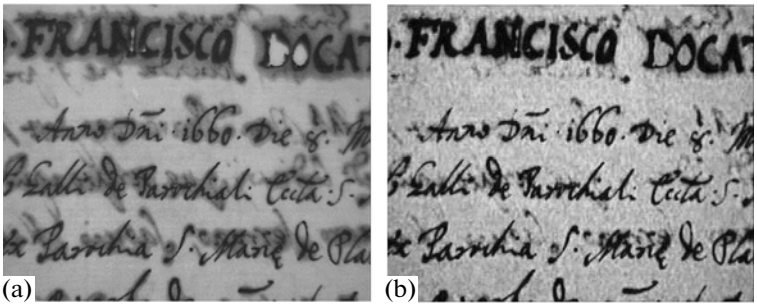

Fig. 15. Reduction of bleed-through preserving the original document appearance: (a) original image; (b) enhanced image.

Finally, in the example of Fig. 13, color decorrelation followed by *K* subtraction has showed to perform very well for the removal of a sensible show-through even if the spectral diversity among the overlapped text is minimal.

4. RECOVERING THE ORIGINAL COLOR OF THE ENHANCED DOCUMENTS

As already mentioned, BSS techniques would enable, at least in principle, the recovery of the original color of the extracted patterns, since their mean RGB reflectance indices are jointly estimated. However, to be effective, this approach requires the achievement of a full separation, or, at least, a perfect extraction of the pattern of interest, which is not always the case, as seen in the many examples above. In the practice, some residuals from the other patterns are still present in the enhanced document, which means that the estimated reflectance indices are affected by the reflectance indices of these residuals. We have found that an even simpler technique can be exploited to recover the global RGB appearance of an enhanced document, when the color of the removed interference is not too much different from that of the background. This technique is based on the YCbCr decomposition of the original, degraded RGB document.

Let us suppose that a grayscale restored image of the main foreground text has been obtained by using one of the techniques described in the previous sections. In this image, the main text, usually dark, will be well contrasted and embedded in a now almost homogeneous background. In addition, let us suppose that in the original RGB document the removed interferences had the same color of the background. Our idea has then been to substitute the *Y* channel of the YCbCr decomposition of the degraded document with the grayscale map of the restored document, leaving unchanged the chrominance components. Going back to the RGB representation, we observe that very often the original color appearance of the enhanced document is recovered well, as it can be appreciated in the examples shown in Figs. 14 and 15.

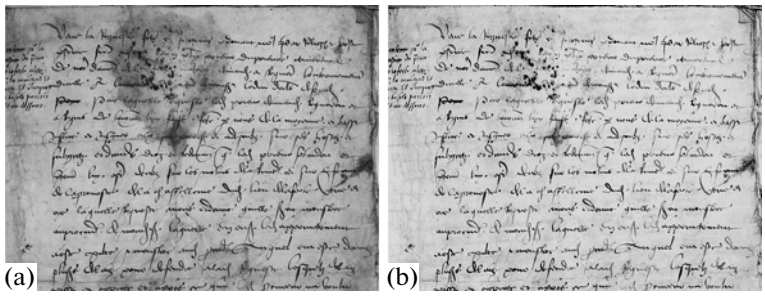

Fig. 16. Reduction of a spot preserving the original document appearance: (a) original image; (b) enhanced image.

In Fig. 16, the performance of RGB decorrelation with color recovery is shown with application to the reduction of spots due for instance to humidity. Note that the sensible attenuation of the large and dark spot allows for an easier readability of the text, while the original color is satisfactorily recovered.

5. CONCLUSIONS

Simple and fast procedures, based on the projection of the RGB image of a document into alternative color spaces, have been reviewed and proposed.

These color spaces, both fixed and self-adaptive, in many cases allow for the enhancement of the main text and the extraction of various features of the document. In particular, the proposed color decorrelation method has been shown to be able to let emerge all the document contents, such as hidden texts and textures. We believe that this achievement could be of great importance for scholars and historians.

Image enhancement methods based on alternative color space representation are very simple, fast and user friendly. In particular, they do not require any intervention from the user side, in that no parameters need to be set. This could be an interesting feature, especially for a routinely application on wide scale, on the large amount of documents stored in the Libraries and Archives of the various States. This also enables their use by the side of users that, in general, are not expert in the field of digital image processing.

A point to be highlighted regards the generality of the approach. In other words, the method is independent of the characteristics of the documents at hand. It can be applied to documents of any type, regardless of their color, text font (e.g. manuscripts or printed texts, language, century, etc.), support (parchment or paper) and ink used, resolution of the acquisition (i.e. character dimension). Generality is an important merit, but it can also be a limitation, in that we cannot expect a single algorithm to perform in the same way (i.e. satisfactorily) on every documents. However, obtaining the best result for a given document would necessarily require the development of specific methods and algorithms, possibly more complex, both in terms of execution times and user intervention. In addition, they could be applied, at best, to a restricted class of documents all having same characteristics.

ACKNOWLEDGMENTS

This work has been supported by the Calabria Region PIA 2008 project no. 1220000119 AMMIRA (Multispectral acquisition, enhancement, indexing and retrieval of artworks)—http://www.aminira.eu.

REFERENCES

1. L. T. Chew, R. Cao, and S. Peiyi, "Restoration of Archival Documents Using a Wavelet Technique," IEEE Trans. Pattern Analysis Machine Intelligence **24**, 1399–1404 (2002).
2. A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing* (Wiley, New York, 2002).
3. P. Dano, "Joint Restoration and Compression of Document Images with Bleed-through Distortion," Master Thesis (Ottawa-Carleton Institute for Electrical and Computer Engineering, School of Information Technology and Engineering, University of Ottawa, 2003).
4. F. Drida, F. LeBourgeois, and H. Emptoz, "Restoring Ink bleed-Through Degraded Document Images Using a Recursive Unsupervised Classification Technique," in *Proc. 7th Workshop on Document Analysts Systems, 2006*, pp. 38–49.
5. R. L. De Queiroz, "On Independent Color Space Transformations for the Compression of CMYK Images," IEEE Trans. Image Proc. **8**, 1446–1451 (1999).
6. E. Dubois and A. Pathak, "Reduction of Bleed-Through in Scanned Manuscript Documents," in *Proc. IS&T Image Processing, Image Quality, Image Capture Systems Conf., Montreal, Canada, 2001*, pp. 177–180.
7. K. Knox, R. Johnston, and R. L. Easton, Jr., "Imaging the Dead Sea Scroll," Opt. Photonics News **8**, 30–34 (1997).
8. K. T. Knox, R. L. Easton, Jr., and W. Christens-Barry, "Image restoration of damaged or Erased Manuscripts," *EUSIPCO*, 2008.
9. A. Hyvärinen and E. Oja, "Independent Component Analysis: Algorithms and Applications," Neural Networks **13**, 411–430 (2000).
10. Y. Leydier, F. LeBourgeois, and H. Emptoz, "Serialized Unsupervised Classifier for Adaptive Color Image Seg-

mentation: Application to Digitized Ancient Manuscripts," in *Proc. Int. Conf. on Pattern Recognition, 2004*, pp. 494–497.

11. G. Leedham, S. Varma, A. Patankar, and V. Govindaraju, "Separating Text and Background in Degraded Document Images—a Comparison of Global Thresholding Techniques for Multi-Stage Thresholding," in *Proc. 8th Int. Workshop on Frontiers in Handwriting Recognition, Niagara on the Lake, Canada, 2002*, pp. 244–24.
12. R. D. Lins, and J. M. Monte da Silva, "Quantitative Method for Assessing Algorithms to Remove Back-to-Front Interference in Documents," in *Proc. ACM Symposium on Applied computing, 2007.*
13. R. F. Moghaddam and M. Cheriet, "Low Quality Document Image Modeling and Enhancement," Int. J. Document Analysis Recognition **11** (4), 183–201 (2009).
14. H. Nishida and T. Suzuki, "Correction Show-Through Effects in Document Images by Multiscale Analysi," in *Proc. 16th Conf. on Pattern Recognition, Quebec City, Canada, 2002*, Vol. 3.
15. H. Nishida and T. Suzuki, "A Multiscale Approach to Restoring Scanned Color Document Images with Show-Through Effects," in *Proc. 7th Int. Conf. on Document Analysis and Recognition, 2003.*
16. Y. Ohta, T. Kanade, and T. Sakai, "Color Information for Region Segmentation Computer," Comput. Vision, Graphics Image Processing **13**, 222–241 (1980).
17. G. Sharma, "Digital Color Imaging," IEEE Trans. Image Processing **6**, 901–932 (1997).
18. G. Sharma, "Show-Through Cancellation in Scans of Duplex Printed Documents," IEEE Trans. Image Processing **10** (5), 736–754 (2001).
19. R. Swift, *Analysis of the Spectra of Degraded Documents* (Center for Imaging Science, Rochester Inst. Tech., 2001), http://www.cis.rit.edu/research/thesis/bs/2001/swift/thesis.html.
20. C. L. Tan, R. Cao, and P. Shen, "Restoration of Archival Documents Using a Wavelet Technique," IEEE Trans. Pattern Analysis Machine Intelligence **24** (10), 1399–1404 (2002).
21. A. Tonazzini, L. Bedini, and E. Salerno, "Independent Component Analysis for Document Restoration," Int. J. Document Analysis Recognition **7**, 17–27 (2004).
22. A. Tonazzini, L. Bedini, and E. Salerno, "Image Analysis on the Archimedes Palimpsest," Ercim News, No. 58, pp. 53–54, 2004.
23. A. Tonazzini, E. Salerno, M. Mochi, and L. Bedini, Bleed-Through Removal from Degraded Documents Using a Color Decorrelation Method (Lecture Notes in Computer Science), in *Proc. 6th Int. Workshop on Document Analysis Systems*, Ed. by S. Marinai and A. Dengel (Florence, 2004), Vol. 3163, pp. 229–240.
24. M. Tkalcic and J. Tosic, "Color Spaces-Perceptual, Historical and Application Background," Thesis, Faculty of Electrical Engineering, University of Ljubljana, 2000.
25. C. Vertan and N. Boujemaa, "Color Texture Classification by Normalized Color Space Representation," in *Proc. Int. Conf. on Pattern Recognition, Barcelona, Spain, 2000*, pp. 3584–3587.
26. C. Wolf, *Document Ink Bleed-Through Removal with Two Hidden Markov Random Fields and a Single Observation Field* (Laboratoire d'Informatique en Images et Systémes d'Information, INSA de Lyon, France, 2006), Tech. Rep. RR-LIRIS-2006-019, Nov. 2006.
27. *Color Encoding Standard* (Xerox Corporation, Xerox Systems Institute, 475 Oakmead Parkway, Sunnyvale, Calif., Mar, 1989).

Anna Tonazzini received the degree in Mathematics (cum laude) from the University of Pisa, Italy. She is a researcher at the Signals and Images Laboratory of the Istituto di Scienza e Tecnologie dell'Informazione, Italian National Research Council (CNR) in Pisa. She was involved in a number of CNR special programs and European research projects, for basic and applied research on image processing and computer vision, and is co-author of over 80 scientific papers. Her present interests include applied inverse problems, multichannel image restoration and reconstruction, document analysis and recognition, blind source separation, computational biology.