# From digitization to NLP: manuscript virtual restoration

Franca Debole, Muhammad Hanif, Emanuele Salerno, Pasquale Savino, Anna Tonazzini
Institute of Information Science and Technologies
Italian National Research Council, Pisa, Italy
Email: anna.tonazzini@isti.cnr.it

***Abstract*—Digitization of the documental heritage conserved in libraries and archives is a common practice, in order to ensure the preservation and fruition of this extended part of the human cultural and historical patrimony. For the most precious, fragile and difficult to read and decipher manuscripts, specialized though portable digitization equipment, such as high resolution multispectral/hyperspectral cameras, is nowadays available. Digitization made it possible the increasingly extensive use of digital image processing techniques, to perform a number of virtual restoration tasks, which constitute a first, often necessary step prior subsequent automatic analysis of the writing contents, with the ultimate goal to perform automatic transcription and/or natural language processing tasks. Here we report our experience in this field, referring, as a case study, to the problem of removing one of the most frequent and impairing degradation affecting many ancient manuscripts, i.e., the bleed-through distortion. In this case, virtual restoration gives also the immediate benefit to facilitate the work of philologists and paleographers interested in examining and transcribing the manuscript in a traditional way.**

***Index Terms*—Ancient manuscript restoration; recto-verso registration; bleed-through removal; blind source separation; sparse representation inpainting**

## I. Introduction

The virtual restoration of an ancient and degraded manuscript simulates, on the digital images, the process of physical and/or chemical restoration performed on the original, tangible manuscript. The aim is twofold: i) to remove or attenuate the interfering patterns caused by several and diverse degradations undergone by the manuscript during time, and due to bad storage environment, careless usage, and the natural ageing and deterioration of the ink and the support; ii) to enhance the interesting text against faded or blurred ink, fragmented characters, and incomplete words. The advantages of virtual restoration over physical restoration are apparent, since a number of different and reversible techniques can be attempted without harming the original manuscript. As an example, in the case of the very common bleed-through degradation, physical restoration is impossible, since the chemical substances needed to remove the seeped ink would also destroy the ink of the foreground text. Again, the reconstruction of fragmented characters and words can take advantage of similarity searching techniques using available dictionaries, both at the image and textual level.

Virtual restoration can be the ultimate goal of manuscript digital processing. Indeed, it can serve for providing the scholars with a help to a better and easier reading of the text, during its manual transcription and the study of the manuscript origin, history and contents. In this sense, while removing interferences and enhancing the writing, it must maintain as much as possible intact the original appearance of the manuscript itself. Thus, it is fundamental that virtual restoration preserves all marks and genuine features such as pencil annotations, stamps, paper watermarks and textures, miniatures, and so on.

On another hand, virtual restoration can be the first, intermediate step toward the automatic analysis of the writings, needed for the automatic or user-assisted textual transcription, and/or for natural language processing purposes. In this sense, it must facilitate subsequent tasks such as layout analysis, text binarization, word spotting, and OCR.

A comprehensive survey of the most advanced technologies and computer science tools applied to the study of manuscripts can be found in [1].

In this paper we consider the problem of bleed-through removal, which is one of the most urgent and challenging issues in the field of virtual restoration of ancient, degraded manuscripts. Indeed, as already highlighted, this kind of degradation is very frequent and difficult to treat. We will compare two different strategies, which are able to cope with this kind of degradation according to the above described two levels of usage of the enhanced manuscripts.

Although many works presented in the current literature address bleed-through removal for application to single-sided and grayscale digital manuscripts, we tackle the problem for recto-verso manuscripts acquired as RGB digital images. As a matter of fact, after the extensive and high quality digitization campaigns recently carried out or in course in the majority of libraries and archives, it is very likely that the digital version of a manuscript affected by ink seeping from the reverse side comes out in the recto-verso and at least RGB modalities. On one hand, diversity of acquisition increases information and, as such, should be exploited, when available. On the other hand, the color of a manuscript is an important cue, both for a more pleasant and authentic representation, and for the information that it can provide about the chemical nature of the substances (i.e. ink pigments), and the kind of degradation that the manuscript has undergone.

However, using both sides of the manuscript implies that an accurate matching between the information carried on by

corresponding pixels should be ensured. In other words, the two pixel values must be the spectral signatures in the two sides of the same geometrical point. This means that the two images must be aligned. To digitally acquire ancient manuscripts, usually professional cameras are used, either high resolution CCD cameras or multispectral cameras, mounted on special mechanical equipment that guarantee a stable setup. Despite that, misalignments between the images of the two sides are likely to occur, due to the human intervention needed for turning around and repositioning the leaf. Thus, registration algorithms must be used prior the restoration process. This issue will be discussed in the paper, as well.

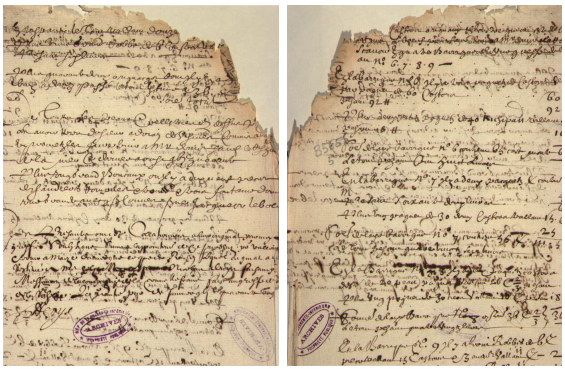
The paper is organized as follows. In Section II we present the registration algorithm that we designed for this specific application. Section III is devoted to the description and the analysis of the results of a first virtual restoration technique, based on linear models of patterns overlapping in the two views, and blind source separation algorithms. In Section IV we discuss the results of a second virtual restoration technique based on non-stationary patterns overlapping model, and image inpainting via sparse representation. We will also show that blind source separation, exploiting the spectral diversity of the restored images, can facilitate subsequent layout analysis. Section V concludes the paper.

## II. Recto-verso registration

Registration of recto-verso images is not an easy task, since the intensity of corresponding foreground and bleed-through areas are usually very different, bleed-through might only occur sparsely across the page, and the binding of the page in case of books may have different degrees of curvature in the two sides. The earliest recto-verso registration methods in the literature considered global affine transformations [2], [3], [4], [5], [6]. More recently, projective transformations have been proposed [7], as well as non-rigid registration methods [8], [9], [10], to cope with binding in the page.

In this paper, we consider flat manuscripts (e.g. letters), so that we assume a global rigid deformation of the horizontally reflected verso with respect to the recto. However, we extend the transformation to be projective, since, besides translations and rotation caused by turning around the leaf, accidental movements of the camera may also cause scale changes and projective deformations. We estimate the transformation parameters by least mean squares (LMSE), based on a suitable number of corresponding points in the two images. Rather than searching for matching points among the set of singular points, such as corners or crosses, they are automatically detected by looking at the maximum of the cross-correlation between small patches of same size and same location in the two sides. In other words, the center of each recto patch is assumed to correspond to that verso point having the coordinates of the peak of the cross-correlation function. Reasonably assuming a moderate misalignment at the local level enables performing this estimation by exploiting the shift property of the Fourier Transform, through the computation of the cross power spectrum of the two patches. Using FFT, this computation is very fast, allowing the quick detection of a very large number of relative translations and then corresponding points, which makes LMSE more robust [11].

When an RGB manuscript is acquired through a multispectral camera, it is likely that the three channels of each single side appear misaligned, as shown in Fig. 2 (a), where a detail of the manuscript recto in Fig. 1 (a) is shown. Consequently, the misalignment of the mirrored color planes of the verso is "inverted" (see Fig. 2 (b)). To correct the misalignment of the color planes and geometrically register the two sides, we start correcting the color channel misalignment of the RGB recto (Fig. 2 (c)). This is trivial if we assume the color channel misalignment only caused by global displacements between the three channels. Hence, choosing a reference channel, the relative displacements of the other two channels can be computed still via the shift property of the FT, by using a single pair of patches. Subsequently, we separately estimate and apply as described before the individual projective transformations from each channel of the corrected recto and the homologous channel of the acquired verso. The final result is shown in Fig. 2 (d). It is apparent that in the registered verso the correct RGB appearance has been restored as well.



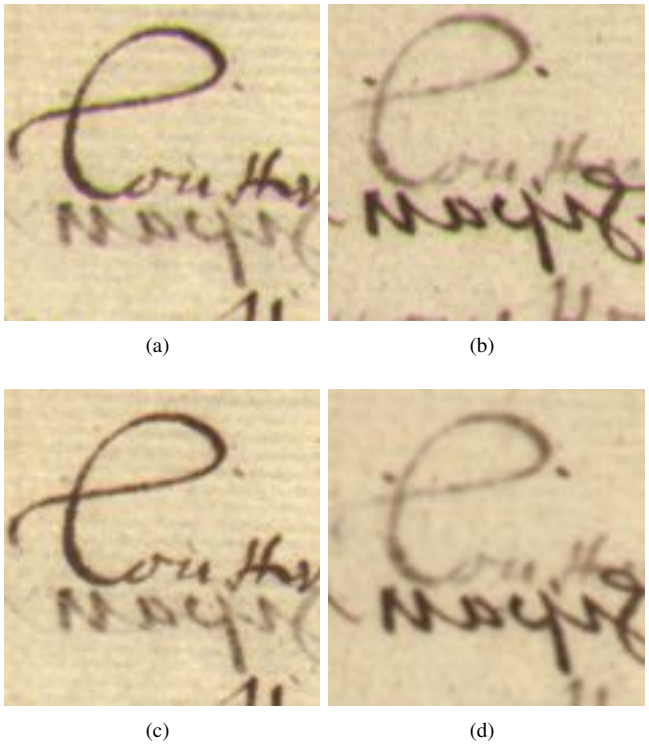
(a) (b)

Fig. 1. The manuscript used for our experiments: (a) recto; (b) verso.

## III. Blind Source Separation

In the literature, bleed-through removal is mainly addressed as a classification problem, where the document image is subdivided into three components: background (the paper support), foreground (the main text), and bleed-through. The existing methods can be divided into two main categories: blind approaches ( [12], [14], [13], where the image of a single side is used, and the far more adopted non-blind approaches ( [15], [16], [17], [18], [19], [20], [21], [22], [23]), where the information of both the recto and verso sides of the document is required.

By adopting a point of view different from classification, any degraded manuscript, including those affected by bleed-through) appears as a mixture of layers of different texts and possibly of other diverse information (paper texture and watermarking, stains, stamps, pencil annotations, etc.). Some of them are interferences (e.g. bleed-through and stains) and



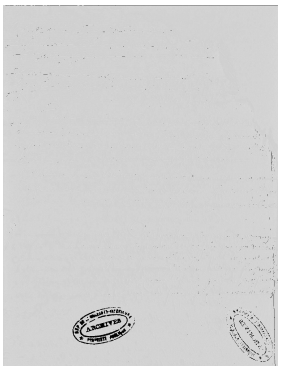
(a) (b) (c) (d)

Fig. 2. Illustration of the whole registration process on a small portion of the recto-verso pair in Fig. 1 : (a) original recto; (b) original verso after horizontal flipping; (c) recto after color plane alignment; (d) verso after geometrical registration on the recto.

should be removed. Others, such as pencil annotations and stamps, may provide useful information to the scholars, and then should be preserved. In any case, the individual channels of the RGB acquisition, or possible multispectral acquisition, still show the various layers overlapped. This suggests that, at each pixel, such channels can be modeled as linear mixtures of the individual layers, weighted by coefficients that are related to the spectral responses of the corresponding object. If the different layers exhibit different spectral responses, that is, they are of different colors, the mixing matrix will be non-singular, and then invertible, if known. On the other hand, when viewed as spatial signals, while the channels are clearly highly correlated, the layers are likely to be uncorrelated.

All the above observations lead to infer that the layers could somehow be separated from each other. Since the mixing matrix is not known, a layer uncorrelation constraint can be exploited to this purpose through statistical Blind Source Separation (BSS) techniques [24], [25]. Indeed, BSS techniques, such as principal component analysis (PCA) and independent component analysis (ICA), linearly combine the highly correlated spectral images to produce a different set of images that are uncorrelated and with decreasing variance, i.e. carrying on decreasing amounts of information. Furthermore, the output channels of ICA are statistically independent. Thus, they can actually produce images each showing one single layer separated from the others [26], [27]. Being the requirement of statistical independence of ICA a stronger condition than the assumption of uncorrelation of PCA, it may also happen that signals that are not well segmented by PCA may be separable by ICA, or by ICA applied on a set of principal components of highest eigenvalues, as done in [28].

In case of perfect separation of the layers, each layer would dominate the grayscale range in the related output channel, while pixels belonging to the other layers would exhibit the same gray value and thus merge with the background in that channel. More realistically, PCA or ICA may not succeed in separating the layers if their statistics are not truly orthogonal. For example, when applying ICA to the image of Figure 1(a), since foreground and bleed-through exhibit the same spectral behaviour, they cannot be fully separated, whereas the stamps can be extracted as isolated patterns (Figure 3). However, the front stamp and the verso, seeping stamp remain mixed, because they have the same color.



(a)

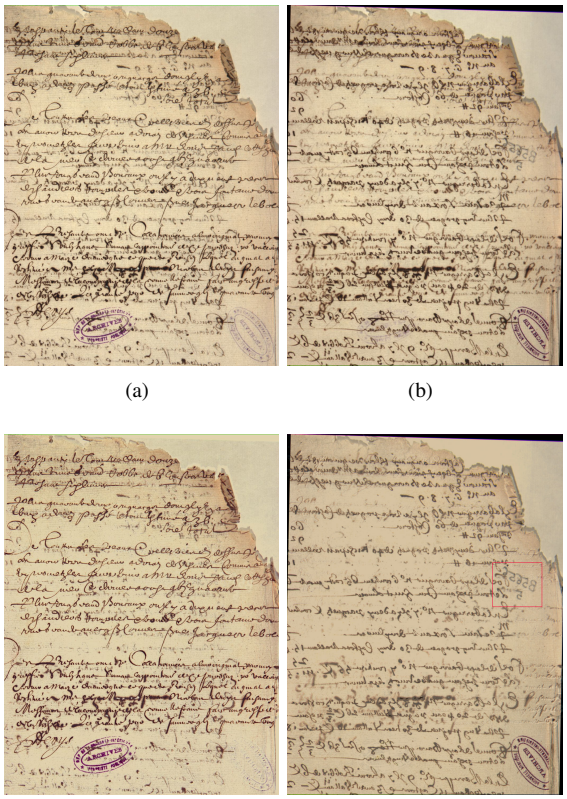
Fig. 3. The unique layer that ICA can isolate from the image in Figure 1(a).

Some other form of diversity of information must thus be sought, through the use of other acquisition modalities. In case of recto-verso acquisitions, the modality considered in this paper, diversity of information is constituted by the different intensities of a same text pattern in the two observations. A linear, instantaneous overlapping model, similar to that devised for a single-side RGB acquisition, can be adopted also in this case. This is a $2 \times 2$ model, whose mixing matrix is related to the percentage of ink seeping from a side to the other. Since it is reasonable to assume that the ink penetrates the paper in the same way from recto to verso and from verso to recto, the mixing matrix can be assumed to be symmetric, so that separation can be equivalently achieved by ICA or the faster symmetric whitening [29].

The superposition model holds for each pair of recto-verso homologous channels, i.e. $(R_r, R_v)$, $(G_r, G_v)$ and $(B_r, B_v)$, where the subscripts $r$ and $v$ stands for recto and verso, respectively, so that, once the three $2 \times 2$ systems have been inverted, the restored channels can be recomposed to furnish the restored RGB sides [5].

An example of the results of this procedure applied to the registered versions of the pair in Figure 1 is shown in Figure 4. Note how the bleed-through is reduced, especially in the verso side, and the color and other manuscript features are preserved

as well. In particular, note how each side has maintained its own stamp only, and the pencil annotation in the verso side, highlighted by the red box, has not been removed. Given the extremely high computational efficiency of such a procedure, this could be used as a routine in libraries and archives.



(a) (b)

(c) (d)

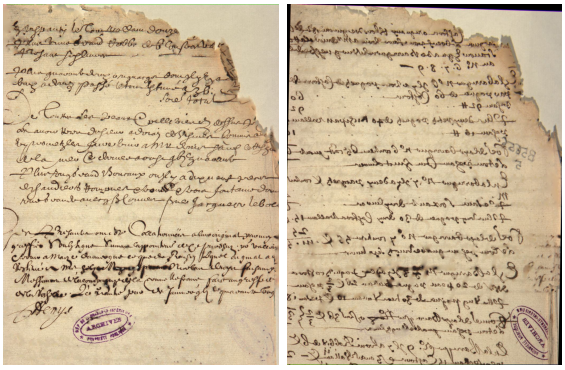
Fig. 4. Application of symmetric whitening to the pair in Figure 1 after their registration: (a) recto; (b) verso; (c) restored recto; (d) restored verso.

## IV. Pixel-wise bleed-through identification and image inpainting

In [30], we proposed another bleed-through removal algorithm based on a data model where the observed optical density of each side is given by the linear combination of the density of the undegraded side and an ink-smeared version of the ideal density of the opposite side. This linear combination is weighted by a pixel-dependent positive parameter, representing the degree of attenuation of the text that shows through, thus making the model non-stationary. Based on simple considerations, the attenuation levels are estimated from the data, and the data model is then inverted in a single step, making the algorithm very fast. In case of RGB images, this model, as the previous, instantaneous one, holds independently for each pair of the three channels, and, as before, the algorithm can be separately applied to each pair of recto-verso color channels to obtain the restored RGB sides. Also this algorithm is able to remove only the unwanted interferences, while preserving other patterns, such as stamps or pencil annotations, that are peculiar of each side, as well as the original colors of foreground and background.

Nevertheless, as the optical density is defined with respect to a unique, average value of the background, in the practice the algorithm substitutes the identified bleed-through pixels with values around the used average background value. In cases where the background is textured and non-uniform, this makes some unpleasant imprints of the bleed-through pattern to be still visible. Thus, on the light of the principle to obtain a virtual restored image which is faithful to the original one as much as possible, we recently proposed an improvement to the original algorithm. In this improved version, the identified bleed-through pattern is inpainted in continuity with its surrounding, by using techniques of sparse representation based image inpainting [31], [32].

Figure 5 shows the results of this second algorithm, to be compared with those of the BSS based algorithm in Figure 4. The much better removal of bleed-through with this algorithm can be clearly appreciated.



(a) (b)

Fig. 5. Application of pixel-wise bleed-through identification and sparse image inpainting: (a) restored recto; (b) restored verso.

Observe that the restored image of Figure 5(a) contains now only two layers of information, the foreground text and the stamp, which exhibit different spectral behaviours. To this image we can then apply BSS tecniques, e.g. ICA, to attempt to separate the two layers. The excellent result is shown in Figure 6. This experiment shows how preliminary virtual restoration of a degraded manuscript can facilitate the task of document layout analysis.

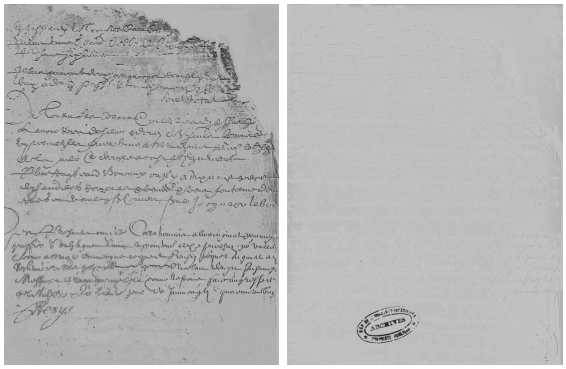
Figure 7(b) shows instead the result of the binarization of the restored recto with the Sauvola algorithm [33], to be compared with Figure 7(a), showing the binarization with the same algorithm of the original degraded recto.

## Acknowledgment

This work has been partially supported by the European Research Consortium for Informatics and Mathematics (ERCIM), within the Alain Bensoussan Fellowship Programme.

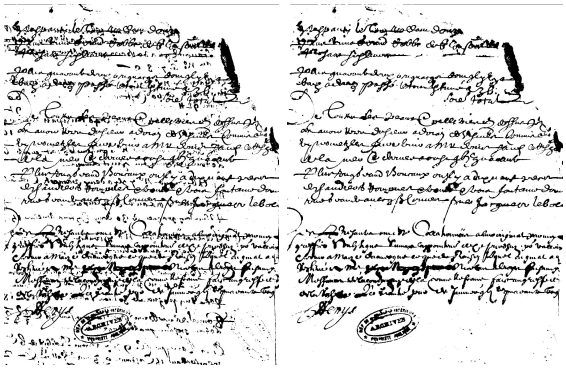
## References

[1] C. Brockmann, M. Friedrich, O. Hahn, B. Neumann, and I. Rabin, Eds., *Natural Sciences and Technology in Manuscript Studies*, ser. Manuscript Cultures. Hamburg: University of Hamburg, 2014, vol. 7.



(a) (b)

Fig. 6. Application of ICA to the restored recto of Fig. 5(a) : (a) extracted foreground text; (b) extracted stamp.



(a) (b)

Fig. 7. Application of the Sauvola binarization before and after virtual restoration: (a) binarized original recto; (b) binarization of the recto restored with the second proposed method.


[2] E. Dubois and A. Pathak, "Reduction of bleed-through in scanned manuscript documents," in *Proc. IS&T Image Processing, Image Quality, Image Capture Systems Conference*, 2001, pp. 177–180.

[3] Q. Wang and C. L. Tan, "Matching of double-sided document images to remove interference," in *Proc. IEEE CVPR 2001*, 2001, p. 1084.

[4] J. Wang, M. S. Brown, and C. L. Tan, "Accurate alignment of double-sided manuscripts for bleed-through removal," in *Proc. 8-th IAPR Workshop on Document Analysis Systems*, 2008, pp. 69–75.

[5] A. Tonazzini, G. Bianco, and E. Salerno, "Registration and enhancement of double-sided degraded manuscripts acquired in multispectral modality," in *Proc. 10th International Conference on Document Analysis and Recognition ICDAR 2009*, 2009, pp. 546 – 550.

[6] V. Rabeux, N. Journet, and J. P. Domenger, "Document recto-verso registration using a dynamic time warping algorithm," in *Proc. Int. Conf. on Document Analysis and Recognition (ICDAR)*, 2011, pp. 1230–1234.

[7] B. Li, W. Wang, and H. Ye, "Multi-sensor image registration based on algebraic projective invariants," *Optics express*, vol. 21, pp. 9824–9838, 2013.

[8] A. Myronenko and S. Xubo, "Intensity-based image registration by minimizing residual complexity," *IEEE Transactions on Medical Imaging*, vol. 29, p. 18821891, 2010.

[9] J. Wang and C. L. Tan, "Non-rigid registration and restoration of double-sided historical manuscripts," in *Proc. Int. Conf. on Document Analysis and Recognition (ICDAR)*, 2011, p. 13741378.

[10] R. Rowley-Brooke, F. Piti, and A. Kokaram, "Nonrigid recto-verso registration using page outline structure and content preserving warps," in *Proc. 2nd International Workshop on Historical Document Imaging and Processing, HIP 2013*, 2013, p. 813.

[11] P. Savino and A. Tonazzini, "Digital restoration of ancient color manuscripts from geometrically misaligned recto-verso pairs," *Journal of Cultural Heritage*, vol. 19, pp. 511–521, 2016.

[12] D. Fadoua, F. L. Bourgeois, and H. Emptoz, "Restoring ink bleed-through degraded document images using a recursive unsupervised classification technique," *Document Analysis Systems VII, Lecture Notes in Computer Science, vol. 3872. Springer*, pp. 27–38, 2006.

[13] B. Sun, S. Li, X. P. Zhang, and J. Sun, "Blind bleed-through removal for scanned historical document image with conditional random fields," *IEEE Trans. Image Process.*, pp. 5702–5712, 2016.

[14] C. Wolf, "Document ink bleed-through removal with two hidden markov random fields and a single observation field," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 431–447, 2010.

[15] B. Ophir and D. Malah, "Show-through cancellation in scanned images using blind source separation techniques," in *Proc. Int. Conf. on Image Processing ICIP*, vol. III, 2007, pp. 233–236.

[16] G. A. Hanasusanto, Z. Wu, and M. S. Brown, "Ink-bleed reduction using functional minimization," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2010, pp. 825–832.

[17] Y. Huang, M. S. Brown, and D. Xu, "User assisted ink-bleed reduction," *IEEE Transactions on Image Processing*, vol. 19, no. 10, pp. 2646–2658, 2010.

[18] R. F. Moghaddam and M. Cheriet, "A variational approach to degraded document enhancement," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, no. 8, pp. 1347–1361, 2010.

[19] R. Rowley-Brooke and A. Kokaram, "Bleed-through removal in degraded documents," *Proc. SPIE 8297 Document Recognition and Retrieval XIX, 82970T-10*, 2012.

[20] F. Merrikh-Bayat, M. Babaie-Zadeh, and C. Jutten, "Using non-negative matrix factorization for removing show-through," in *Proc. LVA/ICA*, 2010, pp. 482–489.

[21] F. Martinelli, E. Salerno, I. Gerace, and A. Tonazzini, "Non-linear model and constrained ml for removing back-to-front interferences from recto-verso documents," *Pattern Recognition*, vol. 45, pp. 596–605, 2012.

[22] E. Salerno, F. Martinelli, and A. Tonazzini, "Nonlinear model identification and seethrough cancellation from recto-verso data," *Int. J. on Document Analysis and Recognition*, vol. 16, pp. 177–187, 2013.

[23] R. Rowley-Brooke, F. Piti, and A. Kokaram, "A non-parametric framework for document bleed-through removal," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2013, pp. 2954–2960.

[24] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing*. New York: Wiley, 2002.

[25] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: Wiley, 2001.

[26] A. Tonazzini, E. Salerno, M. Mochi, and L. Bedini, "Blind source separation techniques for detecting hidden texts and textures in document images," in *Proc. International Conference on Image Analysis and Recognition ICIAR 2004*, 2004, pp. 241–248.

[27] A. Tonazzini, L. Bedini, and E. Salerno, "Independent component analysis for document restoration," *Int. Journal on Document Analysis and Recognition*, vol. 7, pp. 17–27, 2004.

[28] E. Salerno, A. Tonazzini, and L. Bedini, "Digital image analysis to enhance underwritten text in the archimedes palimpsest," *Int. Journal on Document Analysis and Recognition*, vol. 9, pp. 79–87, April 2007.

[29] A. Tonazzini, E. Salerno, and L. Bedini, "Fast correction of bleed-through distortion in grayscale documents by a blind source separation technique," *Int. Journal on Document Analysis and Recognition*, vol. 10, pp. 17–25, June 2007.

[30] A. Tonazzini, P. Savino, and E. Salerno, "A non-stationary density model to separate overlapped texts in degraded documents," *Signal, Image and Video Processing*, vol. 9, pp. 155–164, 2015.

[31] T. Ogawa and M. Haseyama, "Image inpainting based on sparse representations with a perceptual metric," *EURASIP J. Adv. Signal Process.*, vol. 179, pp. 1200–1212, 2013.

[32] M. Hanif, A. Tonazzini, P. Savino, and E. Salerno, "Sparse representation based inpainting for the restoration of document images affected by bleed-through," *Proceedings MDPI*, vol. 2, p. 93, 2018.

[33] J. Sauvola and M. Pietikäinen, "Adaptive document image binarization," *Pattern Recognition*, vol. 33, p. 225236, 2000.