16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

A NONLINEAR BLIND SOURCE SEPARATION SOLUTION FOR REMOVING THE SHOW-THROUGH EFFECT IN THE SCANNED DOCUMENTS

Farnood Merrikh-Bayat[1], Massoud Babaie-Zadeh[1], Christian Jutten[2]*

[1]Department of Electrical Engineering, Sharif University of Technology
Azadi Aven., Tehran, IRAN
phone: + 98-66165925, fax: +98-2166023261, emails: f_merrikhbayat@ee.sharif.edu, mbzadeh@yahoo.com
[2]GIPSA-LAB, Institut National Polytechnique de Grenoble (INPG), Grenoble., France
phone: + 33-476574351, fax: +33-476574790, email: Christian.Jutten@inpg.fr

ABSTRACT

Digital documents are usually degraded during the scanning process due to the printings existing on the backside of the scanning manuscript. This is often caused by the so called show-through effect, the image that interferes with the main picture due to the intrinsic opacity of the paper. This phenomenon is on type of degradation that one would like to remove.

In this paper, we propose a novel and general nonlinear model for show-through phenomenon. A nonlinear blind source separation (BSS) algorithm is used for this particular application in a new recursive and extendible structure for compensating show-through. Finally, we introduce a new structure for removing the show-through and the blurring effect which appears during the scanning process simultaneously.

1. INTRODUCTION

Libraries and archives usually need automatic methods for improving the readability of the ancient or printed documents without altering the original resources. This act is also recommended for the applications concerning with machine-readable version producing from ancient handwritings by using character recognition algorithms; these algorithms require a clean and apparent version of the original documents.

In this paper we consider one of the most common degradations, usually in ancient documents which are written or printed on both sides of the page, called print-through. Print-through is an undesired appearance of a printed image or text on the reverse side of the paper and can be divided into three additive components, each of them corresponding to a physical phenomenon [1]:

- The show-through component related to the paper's intrinsic opacity or low thicknesses;
- The pigment penetration component;
- The vehicle oil separation component, which is related to the loss of opacity due to the filling of pores with oil.

When the ink of the printer does not penetrate in paper considerably the role of pigment penetration and vehicle oil separation components are negligible and print-through can be approximated only by show-through. Such show-through can significantly impair the readability of the document and also cause visual fatigue for the reader. When the show-through degradation is significant (the darkness of the show-through is comparable to, or even greater than, that of some parts of the desired recto writing), then it is practically impossible to remove show-through by only using a simple thresholding operation.

Several approaches to show-through reduction have been investigated. Some authors used various features in document for distinguishing show-through from foreground image and presented show-through removal techniques involving one side of the document only [2], [3]. While these methods certainly perform better than simple thresholding, but there is no way to unambiguously differentiate foreground from show-through without comparing both sides of the document specially in the grayscale images. Some other works are done specially by using image processing algorithms on this specific problem. They processed both sides of the document simultaneously, so it would be possible to identify regions that are mainly show-though, and replace them by an estimate of the background [4], [5]. Most of these work deal with only texts or handwritings and the original images are distorted during the show-through removing procedure. Recent investigations are in process for applying Blind Source Separation (BSS) algorithms for solving this problem because the original sources and the combination style of the sources are unknown [6], [7]. They assume that the document is obtained by adding background, foreground and show-through approximately linearly and use registered version (that is approximate alignment of two images based on corresponding features in those images) of the recto (scanned image obtained from front side of the paper with show-through) and verso (scanned image obtained from back side of the paper with show-through) as an observed data and they try to make the

* This work has been partially supported by Iran NSF (INSF) under contract number 86/994, and also by center for International Research and Collaboration (ISMO) and French embassy in Tehran in the framework of a GundiShapour program..

outputs as independent as possible by BSS techniques. Tonazzini *et al.* in [8] represent a method for removing show-through in colour images by using only one side of the paper; however, the method is applicable only for colour images. Although these methods give good results but the results are not perfect specially when the images of the front and back side of the paper have overlap with each others and the grayscale of the front side's image is near to black. In these areas it can be seen that in the final results the front image has become whiter in that areas compared to other sections which do not have overlap with the back side's image. This is because of the fact that this phenomenon is not linear as we will see in next section. Sharma in [9] considered the nonlinear model for this phenomenon and he tried to compensate this effect by using adaptive filters. However, using adaptive filters for compensating show-through has some disadvantages. First of all, unless the coefficients of the filter are not adapted, the outputs of the filter are not perfect and usable. On the other hand, choosing the size of the filter is important and it is done manually in [9]. Recently, a new line of research has emerged that focuses on nonlinear mixtures separation [10], [11], [12].

In this paper, at first, we will introduce a nonlinear model for show-through based on a simple experiment. The model that we acquired is general and can be extended easily for achieving required quality of the outputs. We modified the nonlinear BSS recursive structure presented by Hosseini and Deville [13] for this specific application and consequently we could compensate the degradation caused by show-through considerably as the results will demonstrate the effectiveness of our technique. Finally, by modifying the separating structure we propose a new structure which considerably enhances the quality of the removing show-through algorithm.

The paper is organized as follows. In Section 2, we introduce our nonlinear model. We devote Section 3 for demonstrating basic blind separation structure for dealing with this kind of nonlinear model. Modification on this structure for enhancing the show-through removing technique is explained in Section 4. Some experimental results with real printed or manuscript documents are presented in Section 5.

2. SHOW-THROUGH NONLINEAR MODELING

Show-through appears when a fraction of the verso is added to recto pixel by pixel in the scanning process. However as the results of the linear BSS show-through removal algorithms indicate, this fraction is proportional to the grayscale of the front image in that point i.e. as the front image become darker, the show-through will be lower. Therefore, it will be convenient to consider show-through effect as:

$$\begin{aligned} f_r^s(m,n) &= a_1 f_r^i(m,n) + b_1 f_v^i(m,n) \times g_1(f_r^i(m,n)) \\ f_v^s(m,n) &= a_2 f_v^i(m,n) + b_2 f_r^i(m,n) \times g_2(f_v^i(m,n)) \end{aligned} \quad (1)$$

where

- m and n : 2-D spatial coordinates on the paper being scanned;

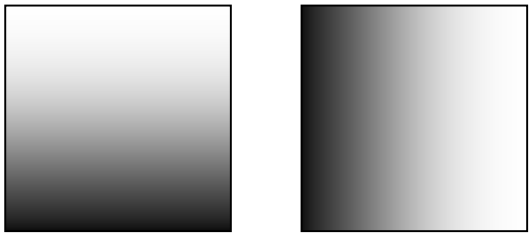


Figure 1 – Front and backside ideal images. Each one is printed on one side of the paper and the paper used for our experiment.

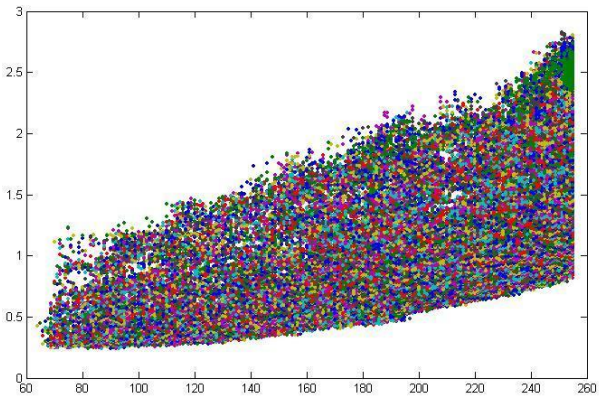


Figure 2 – The result of plotting left hand side of (2) versus $f_r^i(m,n)$ pixel by pixel for obtaining the shape of the nonlinear function involved in show-through phenomenon.

- subscript r: front side (Recto);
- subscript v: back side (Verso);
- superscript i: ideal version of the image (without show-through);
- superscript s: the image acquired by scan (having show-through effect);
- function g: unknown nonlinear function representing the nonlinearity in the show-through effect.

Note that by eliminating the nonlinear function g, a well-known linear model will appear. By symmetry, we can let $a_1 = a_2$, $b_1 = b_2$ and $g_1 = g_2$ (this is like what done in [6]). However, in this paper we only consider the last condition which is $g_1 = g_2$ to preserve the generality of the model as much as possible. By rewriting (1) we obtain:

$$\begin{aligned} \frac{f_r^s(m,n) - a_1 f_r^i(m,n)}{b_1 f_v^i(m,n) + \varepsilon} &= g(f_r^i(m,n)) \\ \frac{f_v^s(m,n) - a_2 f_v^i(m,n)}{b_2 f_r^i(m,n) + \varepsilon} &= g(f_v^i(m,n)) \end{aligned} \quad (2)$$

where ε is a small number for avoiding division by zero. By plotting left hand side of (2) versus $f_r^i(m,n)$, it is possible to recognize the fundamental shape of the function g. As an experiment, we used images shown in Fig. 1 as f_r^i and f_v^i. These images are printed on both sides of a sheet of paper and then scanned. By choosing these images, we will be sure that in the final scanned pictures we will have all combinations of grayscales.

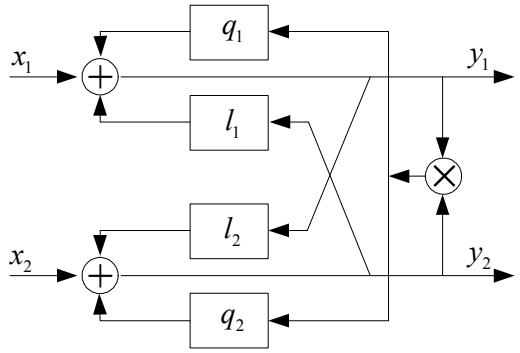


Figure 3 – The configuration proposed in [13] for separating linear-quadratic mixtures, parameters are updated in each iteration using maximum likelihood approach.

Figure 2 shows the plot of $g(f_r^i(m,n))$ versus $f_r^i(m,n)$. Inspiring from this figure, we model the function $g(\alpha)$ as $\gamma\exp(\beta\alpha)$. This means that as the front image becomes whiter, the backside image will be added to the front image with a higher amplitude (show-through can show itself better). Note that if show-through was a linear phenomenon then g was a constant (a horizontal line).

By substituting $\gamma\exp(\beta\alpha)$ instead of g in (1) we get (3):

$$f_r^s(m,n) = a_1 f_r^i(m,n) + b_1' f_v^i(m,n)\times\exp(c_1 f_r^i(m,n)) \quad (3)$$

Equation (3) can be simplified by replacing the exponential function by its first two elements of Taylor representation.

Taylor representation is accurate, as it can be seen at the end of the paper because the value of c_1 is small for most of the real scanned documents having show-through effect. By doing this simplification we obtain:

$$\begin{aligned} f_r^s(m,n) &\approx a_1 f_r^i(m,n) + b_1' f_v^i(m,n)\times\left[1 + c_1 f_r^i(m,n)\right] \\ &= a_1 f_r^i(m,n) + b_1' f_v^i(m,n) + d_1 f_v^i(m,n) f_r^i(m,n) \end{aligned} \quad (4)$$

Analogous to (4), the scanned image of the back side can be written as:

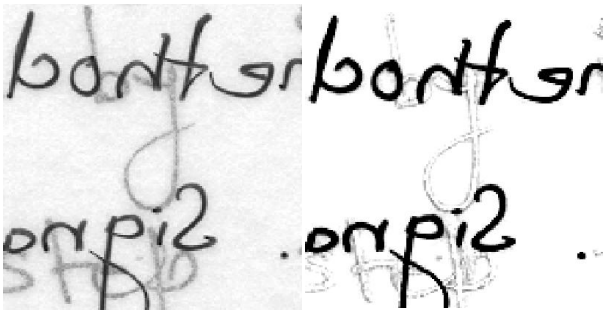
$$\begin{aligned} f_r^s(m,n) &= a_2 f_v^i(m,n) + b_2' f_r^i(m,n)\times\exp(c_2 f_v^i(m,n)) \\ &\approx a_2 f_v^i(m,n) + b_2' f_r^i(m,n)\times\left[1 + c_2 f_v^i(m,n)\right] \\ &= a_2 f_v^i(m,n) + b_2' f_r^i(m,n) + d_2 f_r^i(m,n) f_v^i(m,n) \end{aligned} \quad (5)$$

Note that our nonlinear model is in fact a generalization to the Sharma's model [9].

For separating the sources, first, we should have the mixing coefficients and then solve the nonlinear equations (4) and (5), which the first one is unknown and the second one is not applicable directly to more complicated nonlinear models. To deal with these difficulties, we use BSS techniques.

3. BASIC BLIND SEPARATING STRUCTURE

The nonlinear model (5) is a linear-quadratic mixing model, whose blind separation has already been addressed by Hosseini and Deville [13] based on a recurrent separating structure. It is well known that the independence hypothesis is not sufficient for separating general nonlinear mixtures because of the very large indeterminacies [14], [15]; however, these indeterminacies can be reduced by limiting the structure of the mixing model.



(a) (b)

Figure 4 – show-through removing using the structure of Fig. 3. (a) Recto of the real scanned document showing show-through (from http://www.site.uottawa.ca/~edubois/documents) (b) Output of the structure shown in Fig. 3.

Hosseini and Deville proposed the separating structure of Fig. 3, inspired from the early work of Hérault and Jutten [16]. Note that by setting $q_1 = 0$ and $q_2 = 0$ this structure is reduced to the basic network proposed by Hérault and Jutten.

In this model, parameters are estimated by maximum likelihood approach (or they can be estimated by minimizing the mutual information of the outputs). This configuration can be generalized to arbitrary polynomial models [13].

The computation of the structure's outputs requires the realization of the following recurrent iterative expression:

$$\begin{aligned} y_1(n+1) &= x_1 + l_1 y_2(n) + q_1 y_1(n) y_2(n) \\ y_2(n+1) &= x_2 + l_2 y_1(n) + q_2 y_2(n) y_1(n) \end{aligned} \quad (6)$$

It has been proven in [13] that this model is locally stable at the separating point $(y_1, y_2) = (f_r^i, f_v^i)$ provided that the parameters vector $\boldsymbol{p} = [l_1, l_2, q_1, q_2]$ is chosen properly.

Consequently, the process of show-through correction based on the model of Fig. 3 is described as follows. Firstly, approximate alignment of the front and the back side scans is determined by identifying corresponding image features in the front and the back side scan or by minimizing the square error between these two images. Front and back side images are then applied to the inputs of the structure of Fig. 3, which is initialized by $(y_1(n), y_2(n))|_{n=0} = (\mathbf{0},\mathbf{0})$ and $\boldsymbol{p} = [0, 0, 0, 0]$. At each iteration, the outputs of the structure are considered as the independent original sources and the parameters vector $\boldsymbol{p}$ is updated through the maximum likelihood optimization algorithm presented in [13]. The iteration is continued until the convergence is achieved.

4. MODIFIED BLIND SEPARATION STRUCTURE

Structure of Fig. 3 has an ability to effectively remove the show-through. As an experiment, we applied the algorithm described in the previous section to the scanned image of Fig. 4-(a) and the result of the algorithm is shown in Fig. 4-(b) (only one of the input and output images are shown). Fig. 4-(b) shows that the show-through has been removed perfectly only in that parts which two writings have overlap.

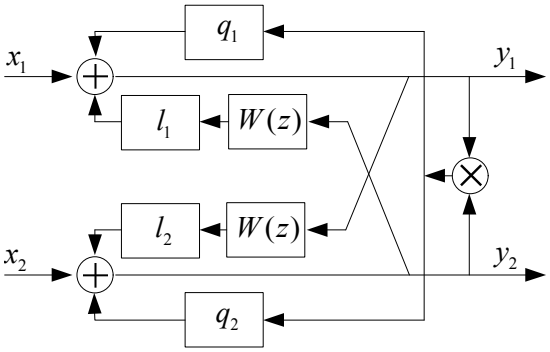


Figure 5 – New Configuration for removing show-through and blurring effect simultaneously.

In scanning process, if the page is not completely opaque and the scanner uses a white backing behind the page, the sensor receives some light that is transmitted through the paper, reflected from the backing and transmitted back through the paper. In this procedure, the light may scatter to different direction because of several reasons such as unsmooth surface of the paper. This scattering phenomenon acts as a low-pass filter and degrades the show-through by eliminating the high frequency parts of it and expanding the boundaries of the show-through image. This effect is known as blurring effect as it has been mentioned and compensated in [7].

The structure of the previous section (Fig. 3) is trying to compensate show-through in one of the scanned image by using the other one and since the blurring effect in one image does not have any correspondence in other one, the configuration of Fig. 3 is unable to remove this effect as can be seen in Fig. 4-(b) as well.

For considering this effect in our model without altering the recovered signal we modify the structure of Fig. 3 and propose a new structure which is shown in Fig. 5.

In this configuration the structure's outputs can be computed as follow:

$$\begin{aligned} y_1(n+1) &= x_1 + l_1 w(n) * y_2(n) + q_1 y_1(n) y_2(n) \\ y_2(n+1) &= x_2 + l_2 w(n) * y_1(n) + q_2 y_2(n) y_1(n) \end{aligned} \quad (7)$$

where $*$ is the convolution operator and $w(n)$ is $N_1 \times N_2$ 2-D low-pass filter such as a median or an averaging filter.

Note that (N_1 , N_2) is proportional to the severity of the blurring effect and depends on the physical structure of the paper such as its sicknesses. For example, as the blurring effect becomes worth, dimensions of the filter should be larger.

In structure of Fig. 5 two low-pass filters are inserted to simulate the blurring effect. Actually, in this configuration, a show-through effect in one of the inputs is removed by the blurred version of the other one and vice versa. By this way, we will be sure that other parts of the images excluding show-through will not encounter blurring or any other degradations.

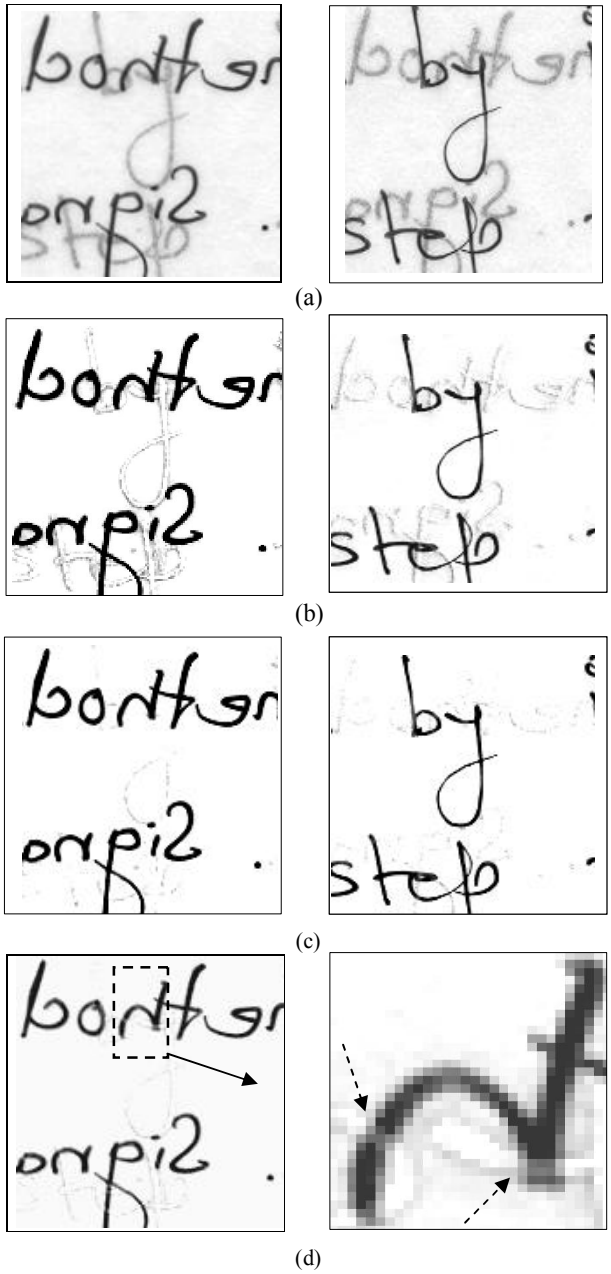


Figure 6 – (a) Two registered image distorted by show-through (b) Obtained results using the structure shown in Fig. 3 (c) Obtained results using the structure shown in Fig. 5 (d) Obtained Result using linear BSS structure and averaging filter.

5. EXPERIMENTAL RESULTS

We applied the algorithm of Section 4 to several real and synthetic images. Images are first registered by minimizing the square error between them and then applied to the structure of Fig. 5.

A typical experiment among the ones we carried out is shown in Fig. 6, in which we used the images of [6]. In this experiment the parameters vector $\boldsymbol{p}$ and $(y_1(n), y_2(n))|_{n=0}$ are initialized as $\boldsymbol{p} = [0, 0, 0, 0]$ and $(\mathbf{0},\mathbf{0})$ respectively and

an averaging filter is used for $w(n)$. We ran the algorithm in MATLAB using a computer characterized by 1.7 GHz CPU and 256 MB ram. The operating system of the computer was Windows XP.

Figure 6-(b) and Fig. 6-(c) show the results acquired by using the structure of Fig. 3 and Fig. 5 respectively. In this experiment, the structure converged after 46 iterations which took about 31 minutes (the most time consuming part is the calculation of the score functions of the outputs which is required in the algorithm of [13] to estimate the gradient of their maximum likelihood cost function). For this experiment, the parameters vector was equal to [-0.4823 -0.4285 0.0424 0.0284] after 46 iterations which confirms the approximate symmetry property mentioned in Section 2. Smallness of the last two parameters confirms that the approximation we used for exponential function in Section 2 was accurate.

Figure 6-(d) shows one of the outputs of the structure of Fig. 5 when it is configured as a linear BSS ($q_1 = 0$ and $q_2 = 0$). An interesting part of the figure is magnified for better understanding. Arrows emphasize the whitening effect produced by using linear models discussed in the introduction. This figure confirms that in the structure of Fig. 5, the nonlinearity of the show-through effect is compensated by the nonlinear model itself and not by the averaging filter.

Although the computational complexity of the algorithm is high, the results are very satisfactory compared to those obtained from linear BSS or adaptive filter techniques [6], [9] and still is simpler than [7].

We tested several kinds of low-pass filters and we recognized that until the show-through is not severe, the dimension and the shape of the low-pass filter are not so much important.

6. CONCLUSION

In this paper we modelled show-through as a nonlinear phenomenon based on an experiment. Then, the model is simplified without loosing the generality of the model. For this specific model, the nonlinear separating structure proposed in [13] based on BSS is used for recovering the original sources. In the next step, we compensate the blurring effect made by scanning process due to the intrinsic opacity of the paper and show-through simultaneously by introducing new nonlinear blind source separating structure. Finally, we justified the effectiveness of our method with several experiments.

The results were very satisfactory with comparison to other show-through removal methods but the computational complexity of the method is rather high and the dimension of the filter should be selected manually. We are now working on adaptive estimation of the filter coefficients, which eventually enhance the quality of separation.

REFERENCES

[1] Larsson, L.O. and Trollsås, P.O., "Print-Through as an Ink/Paper Interaction Effect in Newsprint", The Fundamental Properties of Paper Related to its Uses, Trans. Of the Cambridge Symposium 1,2:600-612, F. Bolam edition, London, 1976.

[2] H. Nishida and T. Suzuki, "A Multi-Scale Approach to Restoring Scanned Colour Documents with Show-Through Effects", in *Proc. Seventh International Conference on Document Analysis and Recognition*, vol. 1, pp. 584–588 (2003).

[3] Q. Wang, T. Xia, L. Li, and C.L. Tan, "Document Image Enhancement Using Directional Wavelet", in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, vol. 2, pp. II–534–II– 539 (2003).

[4] K. Knox, "Show-Through Correction for Two-Sided Documents." United States Patent 5,832,137, Nov. 1998.

[5] Q. Wang and C. L. Tan, "Matching of double-sided document images to remove interference", *IEEE CVPR2001*, Dec 2001.

[6] A. Tonazzini, E. Salerno, and L. Bedini, "Fast correction of bleed-through distortion in grayscale documents by a blind source separation technique", *Int. Journal of Document Analysis, IJDAR(10)*, No. 1, pp. 17-25, June 2007.

[7] B. Ophir and D. Malah, "Show-through cancellation in scanned images using blind source separation techniques", *IEEE Int. Conf. on Image Processing*, vol. 3, pp. 233-236, 2007.

[8] A. Tonazzini, E. Salerno, M. Mochi, and L. Bedini, "Bleed-Through Removal from Degraded Documents Using a Color Decorrelation Method", *Proc. Document Analysis Systems* VI: 6th International Workshop, Springer-Verlag GmbH, vol. 3163 of Lecture Notes in Computer Science, pp. 229–240 (2004).

[9] G. Sharma, "Cancellation of Show-Through in Duplex Scanning," in *Proc. IEEE Int. Conf. Image Processing*, vol. 2, pp. 609–612, Sept. 2000.

[10] C. Jutten, B. Babaie-Zadeh, and S. Hosseini, "Three easy ways for separating nonlinear mixtures?", *Signal Processing*, 84 (2), pp. 217-229, February 2004.

[11] S. Hosseini and C. Jutten, "On the separability of nonlinear mixtures of temporally correlated sources", *IEEE Signal Processing Letters*, Vol. 10, no. 2, pp. 43-46, February 2003.

[12] L. Almeida, "Linear and nonlinear ICA based on mutual information", in Proc. IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (AS-SPCC), Lake Louise, Canada, October 2000, pp. 117-122.

[13] S. Hosseini and Y. Deville, "Blind maximum likelihood separation of a linear-quadratic mixture", in *Proc. 5th Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA'04)*, pp. 694-701, Granada, Spain, September 2004.

[14] A. Taleb and C. Jutten, "Source separation in post-nonlinear mixtures", *IEEE Trans. On Signal Processing,*47(10), pp.2807-2820, 1999.

[15] M. Babaie-Zadeh, "On blind source separation in convolutive and nonlinear mixtures", Ph.D. thesis, INP Grenoble, 2002.

[16] C. Jutten and J. Hérault, "Blind separation of sources", part I: An adaptive algorithm based on neuromimetic architecture, *Signal Processing*, 24:1-10, 1991.