A DUAL ESTIMATION APPROACH FOR REMOVING THE SHOW-THROUGH EFFECT IN THE SCANNED DOCUMENTS

Sabita Langkam and Alok Kanti Deb

Dept. of Electrical Engineering, IIT Kharagpur, West Bengal, INDIA

ABSTRACT

The digital scans of double sided documents suffer from distortions because the contents on the back side of the document often shows up on the front side in the scans and vice-versa either due to transparency of the paper or due to ink-bleeding. This is show-through effect. In this paper a state-space based approach is proposed for removing this commonly found contamination in the scans of duplex printed documents. Separate state-space representation for signals and parameters are defined and a dual state-parameter estimation approach is employed to alleviate the degradation in the scans. The proposed framework functions with two Kalman filters running simultaneously, Kalman state filter to estimate states and Kalman parameter filter for estimating parameters. The simulation results show the effectiveness of the algorithm in removing the show-through.

***Index Terms*—**Dual estimation; Kalman filter; Show-through; Show-through

1. INTRODUCTION

The degradation due to show-through is a problem commonly encountered when digital scans of duplex printed documents are taken. The contents on the back side (verso) of the document often shows up on the front side (recto). This contamination is often attributed to either paper transparency or ink-bleeding. For low quality images, show-through may not be a problem but for quality images it becomes essential to remove the degradation. Several algorithms, from simple thresholding to linear-quadratic nonlinear modelling have been proposed so far of which nonlinear mixture models [1] - [8] have gained popularity. In most of the cases, the scanning process is considered a complex nonlinear process and linear-quadratic modelling of the phenomenon caught attention of many working towards show-through removal. In [6] - [7] least squares adaptive filtering methods are proposed. The image on one side is cleaned considering the image on the other side as reference noise. The show-through cancellation in [2] is framed as a blind source separation (BSS) problem. In [3] a nonlinear nonnegative matrix factorization is defined incorporating both recto and verso giving rise to a nonlinear model. Fornés et al. [4] proposed a multiresolution contrast decomposition based method that estimates the contrast of features at different spatial scales and by thresholding the low contrast components removes the show-through. The model described in [5] employs a two-layer bidirectional neural networks. The stochastic behavior of the networks simulate the linear-quadratic mixing. A nonlinear BSS algorithm is designed in [9] to obtain clean images. Tonazzini et al. [10] assumed mutual independence between the clean recto and verso and applied independent component algorithm to the show-through distorted scans. Merikh-Bayat et al. [11] introduced gradient based Non-negative Matrix Factorization (NMF) algorithm for removing show-through.

In this paper, a state-space approach is proposed for removing the degradation of the scan due to show-through. A dual estimation method based on Kalman filters learns the show-through process and extracts clean images from the degraded ones. There are no pre- or post-processing steps in the proposed method. The simulation with real images show that the proposed simple linear modelling is effective in encountering show-through. The remainder of the paper is organized as follows. Section 2 discusses the state-space description of the problem. Section 3 describes the dual estimation that learns both the unknown signals and parameters of a dynamical system simultaneously. Section 4 explains in brief the tuning and initialization of Kalman filter. Section 5 gives the simulation results to validate the proposed approach and Section 6 summarizes the conclusion.

2. PROBLEM DESCRIPTION

The problem of show-through creeps in when a portion of the verso interferes with the recto during scanning and a pixel-by-pixel added image appears on scans of both sides. To achieve the objective of removing show-through in the scans of double sided documents, the following linear mixing model is considered.

$$\mathbf{y}_k = \mathbf{A}\mathbf{x}_k \tag{1a}$$

$$\begin{bmatrix} y_k^1 \\ y_k^2 \end{bmatrix} = \mathbf{A}\begin{bmatrix} x_k^1 \\ x_k^2 \end{bmatrix} \tag{1b}$$

$$\begin{bmatrix} y_k^1 \\ y_k^2 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_k^1 \\ x_k^2 \end{bmatrix} \quad (1c)$$

The subscript k denotes pixel location, the vector y_k^1 is obtained from the scanned image of one side of the double sided document, y_k^2 is obtained from the scanned image of other side of the same document, matrix $\mathbf{A} \in \Re^{2\times 2}$ denotes the mixing matrix, and x_k^1 and x_k^2 are recto and verso of the document respectively. The scanned images are considered linear combinations of the recto and verso. The objective would be to estimate $\mathbf{x}_k$ when only $\mathbf{y}_k$ is known. The parameters (elements of matrix $\mathbf{A}$) causing show-through would be unknown. The vectors in (1) are obtained by concatenating the rows of the matrices obtained by reading the images. The reformulation of (1) in a state-space formulation can be written as

$$\mathbf{x}_k = \mathbf{F}\mathbf{x}_{k-1} + \mathbf{p}_{k-1} \quad (2a)$$

$$\mathbf{y}_k = \mathbf{H}\mathbf{x}_k + \mathbf{m}_k \quad (2b)$$

assuming that the image pixels have temporal correlations which can be modelled as an AR(1) process. $\mathbf{x}_k$, the original source vector in (1), is the state vector in (2) and the mixing matrix $\mathbf{A}$ in (1) is the observation matrix $\mathbf{H}$ in (2). The matrix $\mathbf{F} = diag\left(f_{11}, f_{22}\right)$ is 2×2 diagonal matrix. where $\mathbf{p}_{k-1} \sim N\left(\mathbf{0}, \mathbf{Q}\right)$ and $\mathbf{m}_k \sim N\left(\mathbf{0}, \mathbf{R}\right)$. The elements of $\mathbf{F}$ and $\mathbf{H}$ are unknown. The basic idea is to employ a dual estimation approach to obtain the clean scanned images $\hat{x}_k^1$ and $\hat{x}_k^2$. For known matrices $\mathbf{F}$ and $\mathbf{H}$ the Kalman filter can be applied to estimate $\mathbf{x}_k$. Table 1 shows the algorithm for standard Kalman Filter (KF).

Table 1 Kalman Filter

Initialization	$\hat{\mathbf{x}}_0 = E\left[\mathbf{x}_0\right],$ $\mathbf{P}_{\mathbf{x}_0} = E\left[\left(\mathbf{x}_0 - \hat{\mathbf{x}}_0\right)\left(\mathbf{x}_0 - \hat{\mathbf{x}}_0\right)'\right]$
State Prediction	$\hat{\mathbf{x}}_k^- = \mathbf{F}\hat{\mathbf{x}}_{k-1}$ $\mathbf{P}_{\mathbf{x}_k}^- = \mathbf{F}_{k-1}\mathbf{P}_{\mathbf{z}_{k-1}}\mathbf{F}_{k-1}' + \mathbf{Q}$
Measurement Prediction	$\hat{\mathbf{y}}_k^- = \mathbf{y}_k - \mathbf{H}\hat{\mathbf{x}}_k^-$ $\mathbf{M}_k^- = \mathbf{H}_k\mathbf{P}_{\mathbf{x}_k}^-\mathbf{H}_k' + \mathbf{R}$
Gain	$\mathbf{K}_k = \mathbf{P}_{\mathbf{x}_k}^-\mathbf{H}'\left(\mathbf{M}_k^-\right)^{-1}$
State Update	$\hat{\mathbf{x}}_k = \hat{\mathbf{x}}_k^- + \mathbf{K}_k\left(\mathbf{y}_k - \hat{\mathbf{y}}_k^-\right)$ $\mathbf{P}_{\mathbf{x}_k} = \mathbf{P}_{\mathbf{x}_k}^- - \mathbf{K}_k\mathbf{M}_k^-\mathbf{K}_k'$

A recursive algorithm alternating between two steps viz. *predict* and *update* forms the working of a Kalman filter. The first step predicts the state $\hat{\mathbf{x}}_k^-$ and the state error covariance matrix $\mathbf{P}_{\mathbf{x}_k}^-$ when the available information is in the form of $\mathbf{x}_{k-1}$ and $\mathbf{Q}$. The second step updates the state estimate obtained in the *predict* step using the information contained in the new measurement $\mathbf{y}_k$ to get $\hat{\mathbf{x}}_k$. The Kalman filter algorithm requires that the process and measurement noise matrices $\mathbf{Q}$ and $\mathbf{R}$ are known. The state and error covariance matrix must be initialized, $\hat{\mathbf{x}}_0$ and $\mathbf{P}_{\mathbf{x}_0}$.

For known $\mathbf{x}_k$ and $\mathbf{y}_k$ a Kalman weight filter can estimate either $\mathbf{F}$ or $\mathbf{H}$. The KF for estimating the parameters of dynamical system in Eq. (2) uses the following state-space representation

$$\begin{aligned} \mathbf{w}_{k+1} &= \mathbf{w}_k + \mathbf{r}_k \\ \mathbf{d}_k &= \mathbf{G}\left(\mathbf{x}_k, \mathbf{w}_k\right) + \mathbf{e}_k \end{aligned} \quad (3)$$

The output $\mathbf{d}_k$ represents nonlinear observation on $\mathbf{w}_k$. The vector $\mathbf{w}_k$ contains either the elements of matrix $\mathbf{F}$ or $\mathbf{H}$. The substitutions $\mathbf{G} \rightarrow \mathbf{F}$ and $\mathbf{d}_k \rightarrow \mathbf{x}_k$ are made for learning the state dynamics and the substitutions $\mathbf{G} \rightarrow \mathbf{H}$ and $\mathbf{d}_k \rightarrow \mathbf{y}_k$ are used for learning the measurement function. KF weight estimation assumes that the clean state $\mathbf{x}_k$ is available. In the KF weight filter algorithm, the matrix $\mathbf{C}_k^w$ is

$$\mathbf{C}_k^w \sim \left.\frac{\partial \mathbf{g}\left(\mathbf{x}_{k-1}, \mathbf{w}\right)^T}{\partial \mathbf{w}}\right|_{\mathbf{w}=\hat{\mathbf{w}}_k^-}$$

Table 2 shows the algorithm for Kalman weight filter. For unknown $\mathbf{F}$ or $\mathbf{H}$ and known $\mathbf{y}_k$, the estimation of $\mathbf{x}_k$ would call for a dual estimation approach.

Table 2 The KF weight filter

Initialization	$\hat{\mathbf{w}}_0 = E\left[\mathbf{w}\right], \mathbf{P}_{\mathbf{w}_0} = E\left[\left(\mathbf{w} - \hat{\mathbf{w}}_0\right)\left(\mathbf{w} - \hat{\mathbf{w}}_0\right)^T\right]$
Weight Prediction	$\hat{\mathbf{w}}_k^- = \hat{\mathbf{w}}_{k-1}$ $\mathbf{P}_{w_k}^- = \mathbf{P}_{w_{k-1}} + \mathbf{R}_{k-1}^r$
Weight Update	$\mathbf{K}_k^w = \mathbf{P}_{w_k}^-\left(\mathbf{C}_k^w\right)^T\left(\mathbf{C}_k^{w-}\mathbf{P}_{w_k}^-\left(\mathbf{C}_k^w\right)^T + \mathbf{R}^e\right)^{-1}$ $\hat{\mathbf{w}}_k = \hat{\mathbf{w}}_k^- + \mathbf{K}_k^w\left(\mathbf{d}_k - \mathbf{g}\left(\hat{\mathbf{w}}_k^-, \mathbf{x}_{k-1}\right)\right)$ $\mathbf{P}_{w_k} = \left(\mathbf{I} - \mathbf{K}_k^w\mathbf{C}_k^w\right)\mathbf{P}_{w_k}^-$

3. DUAL ESTIMATION

The reformulation of BSS in a state-space framework gives unknown states and unknown process model and measurement model parameters. This resembles the problem of dual estimation where both the dynamical system and its parameters are unknown and are estimated simultaneously using only the available observations. The dual estimation algorithm learns both the hidden signals $\mathbf{x}_k$ and parameters $\mathbf{w}$ of the discrete-time dynamical system

$$\mathbf{x}_k = \mathbf{f}\left(\mathbf{x}_{k-1}, \mathbf{w}\right) + \mathbf{p}_k \tag{4a}$$

$$\mathbf{y}_k = \mathbf{h}\left(\mathbf{x}_k, \mathbf{w}\right) + \mathbf{m}_k \tag{4b}$$

The vector $\mathbf{w}$ contains the elements of either $\mathbf{F}$ or $\mathbf{H}$. The dual estimation method alternates between estimating the signal using the model and estimating the model using the signal. Two filters viz. state filter and weight filter would run simultaneously. In this work state KF is concatenated with weight KF to estimate both the signal and the model from the available observations. The dual KF equations for the dynamical system in Eq. (4) are shown in Table 3.

Table 3 Dual KF Algorithm

Initialization	$\hat{\mathbf{x}}_0 = E[\mathbf{x}_0],$ $\mathbf{P}_{\mathbf{x}_0} = E\left[\left(\mathbf{x}_0 - \hat{\mathbf{x}}_0\right)\left(\mathbf{x}_0 - \hat{\mathbf{x}}_0\right)'\right]$ $\hat{\mathbf{w}}_0 = E[\mathbf{w}],$ $\mathbf{P}_{\mathbf{w}_0} = E\left[\left(\mathbf{w} - \hat{\mathbf{w}}_0\right)\left(\mathbf{w} - \hat{\mathbf{w}}_0\right)^T\right]$
Weight Prediction	$\hat{\mathbf{w}}_k^- = \hat{\mathbf{w}}_{k-1}$ $\mathbf{P}_{w_k}^- = \mathbf{P}_{\mathbf{w}_{k-1}} + \mathbf{R}_{k-1}^{\mathbf{r}} = \lambda^{-1}\mathbf{P}_{\mathbf{w}_{k-1}}$
State Prediction	$\hat{\mathbf{x}}_k^- = \mathbf{F}\hat{\mathbf{x}}_{k-1}$ $\mathbf{P}_{\mathbf{x}_k}^- = \mathbf{F}_{k-1}\mathbf{P}_{\mathbf{z}_{k-1}}\mathbf{F}_{k-1}' + \mathbf{Q}$
State Update	$\hat{\mathbf{y}}_k^- = \mathbf{y}_k - \mathbf{H}\hat{\mathbf{x}}_k^-$ $\mathbf{M}_k^- = \mathbf{H}_k\mathbf{P}_{\mathbf{x}_k}^-\mathbf{H}_k' + \mathbf{R}$ $\mathbf{K}_k = \mathbf{P}_{\mathbf{x}_k}^-\mathbf{H}'\left(\mathbf{M}_k^-\right)^{-1}$ $\hat{\mathbf{x}}_k = \hat{\mathbf{x}}_k^- + \mathbf{K}_k\left(\mathbf{y}_k - \hat{\mathbf{y}}_k^-\right)$ $\mathbf{P}_{\mathbf{x}_k} = \mathbf{P}_{\mathbf{x}_k}^- - \mathbf{K}_k\mathbf{M}_k^-\mathbf{K}_k'$
Weight Update	$\mathbf{K}_k^{\mathbf{w}} = \mathbf{P}_{\mathbf{w}_k}^-\left(\mathbf{C}_k^{\mathbf{w}}\right)^T\left[\mathbf{C}_k^{\mathbf{w}}\mathbf{P}_{\mathbf{w}_k}^-\left(\mathbf{C}_k^{\mathbf{w}}\right)^T + \mathbf{R}^{\mathbf{e}}\right]^{-1}$ $\hat{\mathbf{w}}_k = \hat{\mathbf{w}}_k^- + \mathbf{K}_k^{\mathbf{w}}\mathbf{e}$

4. KALMAN FILTER TUNING AND INITIALIZATION

The original formulation of Kalman filter assumes that the mean value and the variance of the initial state is known. When this information is missing the initial phase of the filtering would show transients in the value of the estimates. The unconditional expectation and variance-covariance matrix can be calculated in the initialization phase.

The Kalman filter assumes that the filter noise statistics are known and performs only state estimation. However, these statistics are seldom known. Hence a proper filter tuning in terms of process and measurement noise covariances becomes important. The knowledge of how filter behaves for different values of $\mathbf{Q}$ and $\mathbf{R}$ helps in achieving proper filter tuning. If $\mathbf{P}_0 = 0$ and $\mathbf{Q} = 0$, the filter will ignore the measured data and will not evolve. The state error covariance matrix $\mathbf{P}_0$ is generally considered to be a diagonal matrix whose off-diagonal elements are zeros and diagonal elements are some large values. A large value of $\mathbf{P}_0$ will make the filter believe measured values more and ignore state model while a small value of $\mathbf{P}_0$ would go the reverse way. This error matrix controls the initial transients in the filter behavior. The noise matrix $\mathbf{Q}$ takes care of the unmodelled errors in the state equation. The measurement noise covariance matrix is usually kept close to zero. A careful study of the Kalman filter equations reveal that the effect of $\mathbf{Q}$ is opposite to that of $\mathbf{R}$. Hence a careful combination needs to be chosen.

5. SIMULATION RESULTS

In this section the proposed dual estimation method is applied on real images, snapshots taken from [2] and [3], and mixtures of MATLAB images (color). For all the cases discussed the following were the parameters assumed:

$$\mathbf{Q} = 10^6\mathbf{I}_2, \mathbf{P}_{\mathbf{x}_0} = 10^{-3}\mathbf{I}_2, \mathbf{R} = 3\mathbf{I}_2, \mathbf{R}^{\mathbf{e}} = \mathbf{R} + \mathbf{Q}, \mathbf{P}_{\mathbf{w}_0} = 0.01$$

Case 1. *Real images*

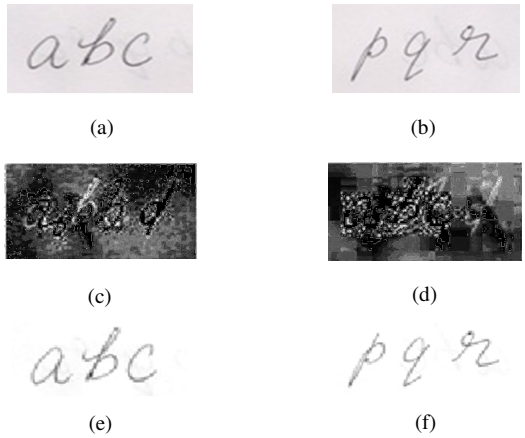
The experiment was performed on scans obtained from HP Scanjet 200. The performance was compared with FastICA performed on the same set of images. The size of the input images is 54 × 100 pixels. The results in Fig. 1 show that the show-through has been removed effectively by the proposed algorithm. FastICA fails miserably. One of the most probable reason may be that in the basic FastICA algorithm, the objective is to separate independent sources from their mixtures. The sources may not always be independent.

Case 2. *Snapshots from [2]*

The method in [2] is a BSS technique. The authors introduced a mean square error cost function defining a nonlinear mixing model and a regularization term based on total variation. The size of the input images is 128 × 137 pixels. Figure 2 shows that the proposed dual estimation technique removes show-through and the performance is better than that in [2].

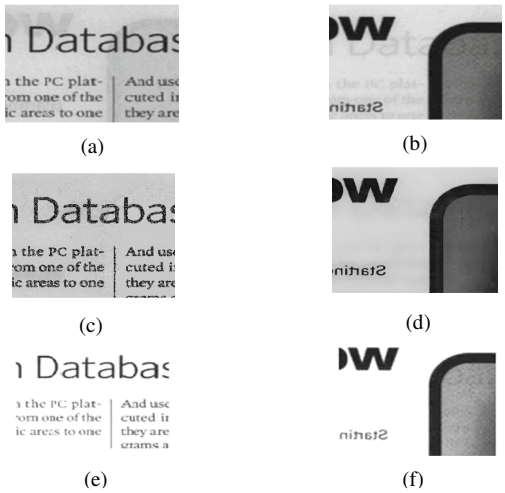
Case 3. *Snapshots from [3]*

Authors in [3] developed a nonlinear NMF algorithm with adaptive smoothing. The simulation results in Fig. 3 confirm better performance achieved by the proposed dual estimation technique.



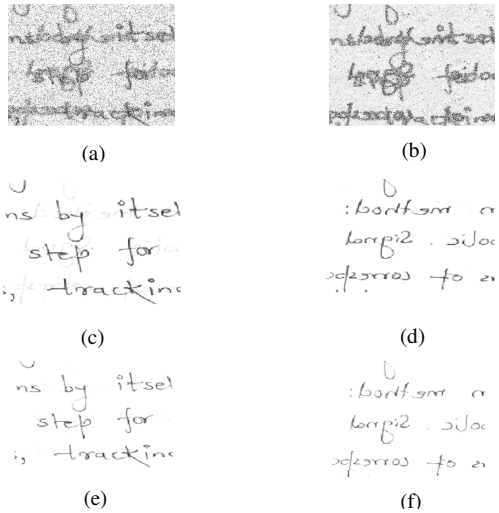
(a) (b) (c) (d) (e) (f)

Fig. 1. (a) and (b) are scanned images. (c) and (d) are outputs of FastICA. (e) and (f) are outputs of proposed algorithm.



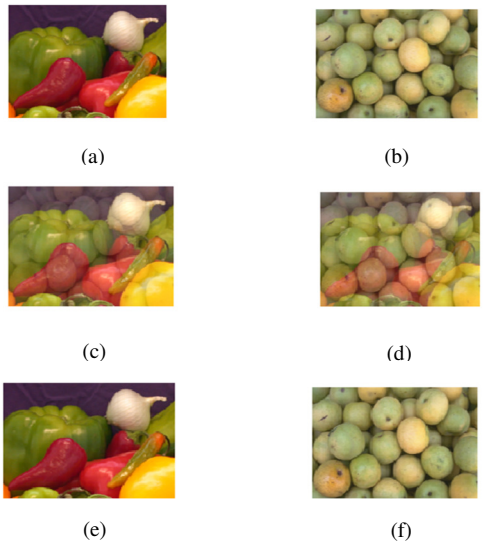
(a) (b) (c) (d) (e) (f)

Fig. 2. (a) and (b) are scanned images. (c) and (d) are outputs from BSS developed in [2]. (e) and (f) are outputs of proposed algorithm.



(a) (b) (c) (d) (e) (f)

Fig. 3. (a) and (b) are scanned images. (c) and (d) are outputs of Nonlinear NMF with adaptive smoothing developed in [3]. (e) and (f) are outputs of proposed algorithm.



(a) (b) (c) (d) (e) (f)

Fig. 4. (a) and (b) are MATLAB images. (c) and (d) are mixture images. (e) and (f) are outputs of proposed algorithm.

Case 4. *MATLAB Images (Color)*
MATLAB's pre-packaged demo images (color) *onion.png* and *pears.png* were mixed and the proposed algorithm was applied to obtain the original images. The simulation results in Fig. 4 confirm better performance achieved by the proposed dual estimation technique.

The performance comparison in terms of mutual information is shown in Table 4. The values do not change significantly but it does indicate towards achieving visually improved images.

Table 4 Similarity measure: Mutual Information

	Case 1	*Case 2*	*Case 3*
a, b	0.4096	1.4649	1.3557
c, d	0.3406	0.2889	0.2340
e, f	**0.2968**	**0.2151**	**0.0615**

6. CONCLUSION

The paper presents state-space approach for obtaining clean images out of contaminated scans of double sided documents. It has been showed that show-through can be effectively represented as a linear model and a dual estimation method with two Kalman filters can be applied to extract the clear images thus removing show-through. Simulations show the effectiveness of the proposed method and suggests potential in the dual estimation approach for removing show-through. An intelligent choice of noise parameters and proper initialization would be necessary for efficient separation of sources.

REFERENCES

[1] G. Sharma, "Cancellation of show-through in duplex scanning," *IEEE Int. Conf. on Image Processing*, vol. 2, 2000.

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

[3] Qingju Liu and Wenwu Wang, "Show-through removal for scanned images using non-linear NMF with adaptive smoothing," *IEEE China Summit & Int. Conf. on Signal and Information Processing*, 2013.

[4] Alicia Fornés, Xavier Otazu, and Josep Lladós, "Show-through cancellation and image enhancement by multiresolution contrast processing," *IEEE Int. Conf. on Document Analysis and Recognition*, pp. 200-204, 2013.

[5] Mikio Oda, and Hiromi Miyajima, "Show-through cancellation in scanned documents using two-layer bidirectional neural networks," *Joint IEEE Int. Conf. on Soft Computing and Intelligent Systems and Int. Symp. on Advanced Intelligent Systems (ISIS)*, pp. 71-74, 2014.

[6] G. Sharma, "Show-through cancellation in scans of duplex printed documents," *IEEE Trans. on Image Processing*, vol. 10, no. 5, pp. 736-754, May 2001.

[7] Boaz Ophir and David Malah, "Improved cross-talk cancellation in scanned images by adaptive decorrelation," *IEEE Convention of Electrical and Electronics Engineers*, pp. 388-391, 2004.

[8] Farnood Merrikh-Bayat, Massoud Babaie-Zadeh, and Christian Jutten, "Linear-quadratic blind source separating structure for removing show-through in scanned documents," *Int. Journal on Document Analysis and Recognition (IJDAR)*, vol. 14, no. 4, pp. 319-333, 2011.

[9] Merrikh-Bayat, Farnood, Massoud Babaie-Zadeh, and Christian Jutten, "A nonlinear blind source separation solution for removing the show-through effect in the scanned documents," *IEEE European Signal Processing Conference*, pp. 1-5, 2008.

[10] Tonazzini, Anna, Emanuele Salerno, and Luigi Bedini, "Fast correction of bleed-through distortion in grayscale documents by a blind source separation technique," *International Journal of Document Analysis and Recognition (IJDAR)*, vol. 10, no. 1, pp. 17-25, 2007.

[11] Merrkh-Bayat, Farnood, Massoud Babaie-Zadeh, and Christian Jutten, "Using non-negative matrix factorization for removing show-through," *International Conference on Latent Variable Analysis and Signal Separation*, Springer Berlin Heidelberg, pp. 482-489, 2010.