Using Non-Negative Matrix Factorization for Removing Show-Through

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

[1] Department of Electrical Engineering, Sharif University of Technology, Azadi Avenue, Tehran, Iran
[2] GIPSA-lab, Depart. of Images and Signal, Grenoble, France and Institut Universitaire de France
f_merrikhbayat@ee.sharif.edu, mbzadeh@sharif.edu, christian.jutten@gipsa-lab.grenoble-inp.fr

Abstract. Scanning process usually degrades digital documents due to the contents of the backside of the scanned manuscript. This is often because of the show-through effect, *i.e.* the backside image that interferes with the main front side picture mainly due to the intrinsic transparency of the paper used for printing or writing.

In this paper, we first use one of Non-negative Matrix Factorization (NMF) methods for canceling show-through phenomenon. Then, non-linearity of show-through effect is included by changing the cost function used in this method. Simulation results show that this proposed algorithm can remove show-through effectively.

1 Introduction

In this paper, we are going to consider one of the most common degradations, called show-through usually appearing in ancient documents which are written or printed on both sides of the page. Show-through is an undesired appearance of a printed image or text of the reverse side of the paper which can significantly degrade the readability of the document.

Several approaches for show-through reduction have been already investigated. Some authors have used only one side of the document and they have tried to distinguish show-through from foreground image by using various features in document and presented show-through removal techniques [1]. Although these methods certainly perform better than simple thresholding, there is no way to unambiguously differentiate foreground from show-through without comparing both sides of the document. In [2], authors process both sides of the document simultaneously, in order to identify regions that are mainly show-though, and replace them by an estimate of the background. Most of these works have the drawback that they can only be used with texts or handwritings. Interests

⋆ This work has be partially supported by center for International Research and Collaboration (ISMO) and French embassy in Tehran in the framework of a GundiShapour collaboration program.

V. Vigneron et al. (Eds.): LVA/ICA 2010, LNCS 6365, pp. 482–489, 2010.

in applying Blind Source Separation (BSS) algorithms for solving this problem have been increased noticeably nowadays. In [3,4], authors modeled the show-through effect as a superimposition of back and front sides printed images. Then, by assuming that the scanned front side image (corrupted by backside) and the scanned back side image (corrupted by front side) are linear mixtures of the *independent* front and back sides images, they used BSS techniques for estimating the pure sources. Tonazzini *et al.* in [5] represent an effective method for removing show-through in color images by using only one side of the paper. One of the most disadvantages of these methods in spite of their good results is that the results are not perfect especially in regions where the images of the front and back sides of the paper have overlaps and the front side's image is near to black. In such regions, the recovered front image is whiter than other sections where there is no overlap. The main reason of having poor results in these areas is that show-through is in fact a nonlinear phenomenon as illustrated in [6,7]. Sharma in [8] considered a nonlinear model for this phenomenon and proposed to compensate this effect by using adaptive filters. Almeida suggested in [9] MISEP method based on Multi-Layer Perceptron networks for separating real-word nonlinear image mixtures. The main drawback of using MISEP or other universal nonlinear networks for BSS is the separability issue: the ICA does not lead necessarily to BSS using such networks. In addition to BSS techniques, wavelet-based method has also been recently proposed for separation of nonlinear show-through and bleed-through image mixtures [10].

In this paper, we first consider linear mixing model for show-through and use one of non-negative matrix factorization methods for separating the pure sources. Then, by performing simple simulations, we show that using linear model for show-through is not correct. Therefore, non-linearity of show-through is included in our method by changing the cost function used in selected NMF algorithm.

The paper is organized as follows. In Section 2, we assume linear model for show-through and use NMF for its removal. Section 3 describes a procedure of modifying the selected NMF algorithm for taking into account non-linearity of show-through. Finally, a few experimental results with real printed or manuscript documents are presented in Section 4, before conclusions and perspectives in Section 5.

2 Modeling Show-through as a Linear Phenomenon and Using NMF for Separating the Sources

In this section, assuming that show-through is a linear phenomenon, we will try to apply NMF to this particular application as another method for separating the sources and decreasing the degradation caused by show-through.

Let vectors $\mathbf{s}_1$ and $\mathbf{s}_2$ be the front side and backside pure images written or printed on sides of the paper respectively, and vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ be the scanned images of sides of the paper. Note that these vectors are obtained by

concatenating rows of the corresponding image matrixes. Then, assuming the show-through is linear, we have

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \times \begin{bmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \end{bmatrix} = \mathbf{AS}. \tag{1}$$

If the size of images obtained through the scanning process is $m \times n$, the size of vectors $\mathbf{s}_1$, $\mathbf{s}_2$, $\mathbf{x}_1$ and $\mathbf{x}_2$ will be $1 \times mn$. Here, the objective is to find the original sources, *i.e.* $\mathbf{s}_1$ and $\mathbf{s}_2$, and it should be noted that in (1), both of the mixing matrix, $\mathbf{A}$, and pure sources, $\mathbf{S}$, are unknown.

Instead of somewhat complex BSS techniques, here we will show how NMF can be used for separating the sources in linear mixtures [11].

The main idea behind NMF methods is to factorize a non-negative matrix $\mathbf{X}$ as a product of two other non-negative matrices, $\mathbf{A}$ and $\mathbf{S}$ [11]. Sometimes, these non-negativity constraints are only placed on matrices $\mathbf{S}$ and $\mathbf{X}$ as we will do in this paper, too. Note that these non-negativity constraints are essential in our case because we are trying to extract the original sources which are non-negative pure images. A conventional approach to find $\mathbf{A}$ and $\mathbf{S}$ is by minimizing the difference between $\mathbf{X}$ and $\mathbf{AS}$ [12]

$$\min_{\mathbf{S},\mathbf{A}} J_1(\mathbf{S},\mathbf{A}) = \|\mathbf{X} - \mathbf{AS}\|_F^2 \qquad \text{subject to:} \quad s_{ij} \text{ and } a_{kl} > 0,\ \forall i,j,k,l, \tag{2}$$

where in this cost function $\|\cdot\|_F$ denotes the Frobenius norm and s_{ij} and a_{ij} are respectively the element of matrixes $\mathbf{S}$ and $\mathbf{A}$ in the ith row and the jth column. The gradient descent techniques has been previously proposed for minimization of this cost function [12] and we restate it here for convenience as follows

$$\mathbf{A}^{k+1} = \mathbf{A}^k - \mu \nabla_{\mathbf{A}} J_1(\mathbf{S}^k, \mathbf{A}^k) \quad \text{and} \quad \mathbf{S}^{k+1} = \mathbf{S}^k - \mu \nabla_{\mathbf{S}} J_1(\mathbf{S}^k, \mathbf{A}^k), \tag{3}$$

where k and μ are the index of iteration and fixed step size, respectively, and gradients of J_1 with respect to $\mathbf{S}$ and $\mathbf{A}$ are

$$\nabla_{\mathbf{S}} J_1(\mathbf{S},\mathbf{A}) = \mathbf{A}^T(\mathbf{AS} - \mathbf{X}) \quad \text{and} \quad \nabla_{\mathbf{A}} J_1(\mathbf{S},\mathbf{A}) = (\mathbf{AS} - \mathbf{X})\mathbf{S}^T, \tag{4}$$

where T stands for transposition. Due to the subtraction operation used in each iteration, the non-negative property of $\mathbf{S}$ and $\mathbf{A}$ cannot be guaranteed, and a projection has to be introduced to project any negative elements back to non-negative regions. This has be done in each iteration in this method by

$$\mathbf{A}^{k+1} = \max(0, \mathbf{A}^{k+1}) \quad \text{and} \quad \mathbf{S}^{k+1} = \max(0, \mathbf{S}^{k+1}), \tag{5}$$

Therefore, this algorithm is essentially a simple Projected Gradient (PG) method [12]. Hence, we can summarize the application of the projected gradient based NMF algorithm of [12] for removing show-through as:

- Initialize matrices $\mathbf{A}^0$ and $\mathbf{S}^0$ by positive random values;
- Consider a suitable value for step size, *i.e.* μ;
- For $k = 0, 1, \ldots$ until the convergence of the algorithm:

- Calculate the gradients of cost function J_1 presented in (4);
- Update matrices $\mathbf{A}$ and $\mathbf{S}$ using equation (3);
- Replace the negative elements of $\mathbf{A}$ and $\mathbf{S}$ by zero.

Figures 1(c) and 1(d) show the result of applying this NMF method to Figs. 1(a) and 1(b) which belong to a manuscript degraded by show-through effect. The results illustrate the capability of NMF algorithms in removing show-through and improving the readability of the document. However, a new degradation is still visible in the results of this simulation. This degradation is associated with linearly modeling of show-through and for better visibility of its effect on the outputs, an interesting region of one of the output images is magnified which is shown in Fig. 1(e). Here, it can be seen that in areas where two images printed on sides of the paper have overlap with each other (shown by arrows), recovered pixels are whiter than what they should be. This is because of the fact that show-through is a nonlinear phenomenon as we have shown in our previous paper experimentally either [6]. Using linear model for show-through means that during the scanning process, all of the pixels of the backside image will be added with their corresponding pixels in the front side image by the same factor. Whereas in reality, this factor is different for each pixel and depends on the grayscale of the front side pixels; as the front side image becomes darker, backside image is added to the front image with lower factor.

In the next section, we will change the cost function used in the NMF method to compensate the degradation caused by linearly modeling of show-through.

3 Compensating the Degradation Caused by Using Linear Model for Show-through

Figure 1(e) demonstrated the problem of using linear model for show-through. We propose here that this degradation can be compensated by adding some regularizing terms to the cost function of (2). These added terms should have some properties such as:

- Where there is no overlap between the printed images on sides of the paper, the results obtained through the minimization of non-regularized cost function, J_1, are fine. So minimization of these added terms should have no effect on output results;
- Where there is overlap between the printed images on sides of the paper, the results obtained through the minimization of non-regularized cost function, J_1, are whiter than what they should be. Therefore, minimization of these added terms should cause the recovered images to become darker in these areas (see such areas that are shown by arrows in Fig. 1(e)).

Here, we will show that the following suggested regularized cost function has the above mentioned properties (its reason will be explained after the calculation of the gradients of this new cost fuction)

$$J_2(\mathbf{S},\mathbf{A}) = \|\mathbf{X}-\mathbf{AS}\|_F^2 + w_1\left(\mathbf{s}_1\mathbf{v}^T\right) + w_2\left(\mathbf{s}_2\mathbf{v}^T\right) \qquad \text{subject to:} \quad s_{ij} > 0,\ \forall i,j, \tag{6}$$

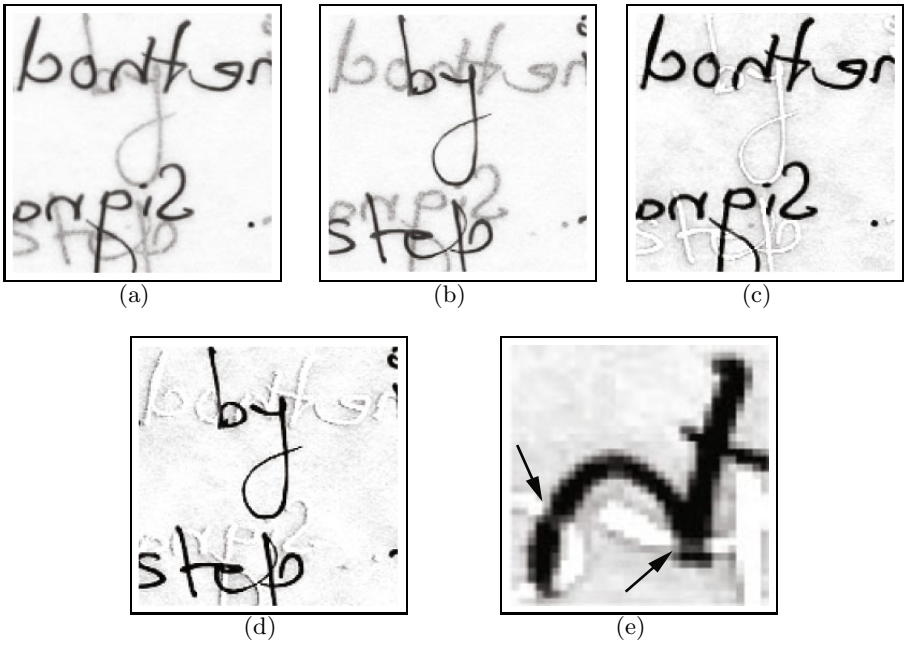


Fig. 1. Results of NMF method while assuming linear model for show-through. (a) and (b) are recto and verso images having show-through (obtained from *http://www.site.uottawa.ca/∼edubois/documents*). (c) and (d) are recovered recto and verso images. Figure (e) shows that using linear model for show-through has some problems.

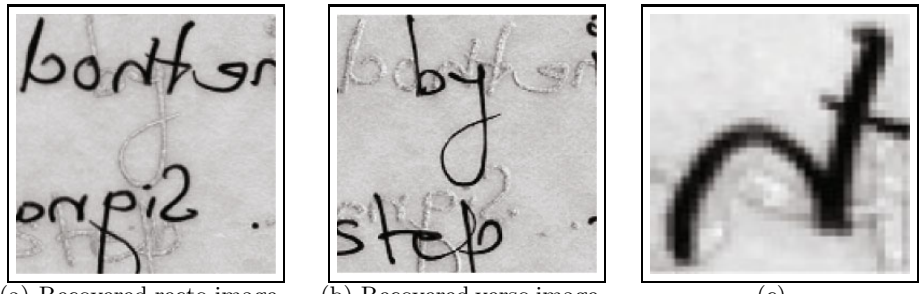
where $w_1 > 0$ and $w_2 > 0$ are regularization weight parameters and $\mathbf{v}$ is a spatially varying weight vector whose ith element is defined as

$$v(i) = exp(-\frac{\left(x_1^2(i) + x_2^2(i)\right)}{2\sigma^2}) \qquad \text{for} \qquad i = 1, 2, \ldots, nm, \tag{7}$$

where σ is a constant that determines the variance of the above 2-dimensional Gaussian function and $x_1(i)$ and $x_2(i)$ are the ith element of vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ respectively. It is clear that $v(i)$ will have high value only if both of $x_1(i)$ and $x_2(i)$ have small values.

Again, $J_2(\mathbf{S}, \mathbf{A})$ can be minimized by using a PG approach. Gradient of the above cost function, *i.e.* $J_2(\mathbf{S}, \mathbf{A})$, with respect to $\mathbf{A}$ is the same as (4). Gradient of this cost function with respect to $\mathbf{S}$ is

$$\nabla_{\mathbf{S}} J_2(\mathbf{S}, \mathbf{A}) = \mathbf{A}^T(\mathbf{AS} - \mathbf{X}) + \begin{bmatrix} w_1\mathbf{v} \\ w_2\mathbf{v} \end{bmatrix}. \tag{8}$$



(a) Recovered recto image. (b) Recovered verso image. (c)

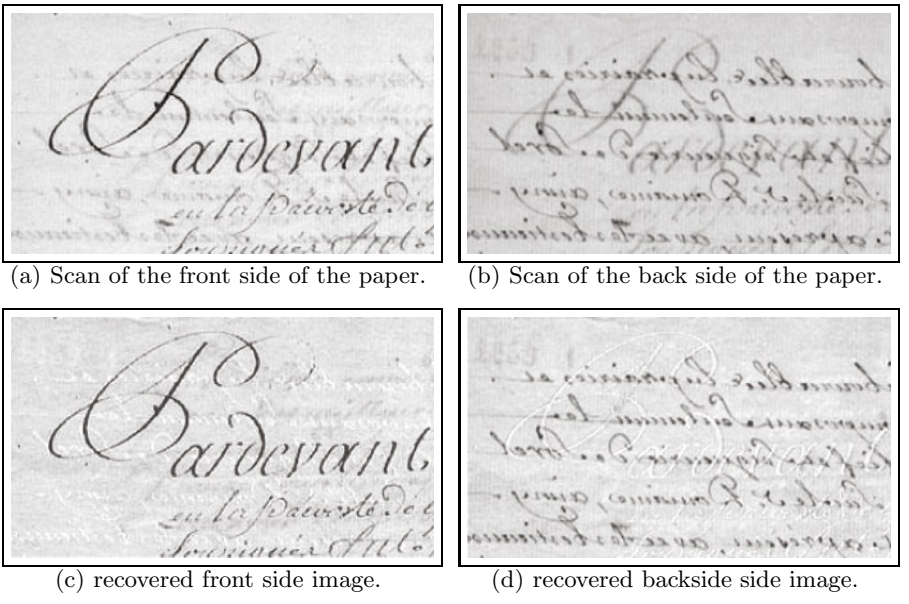
Fig. 2. Results obtained by using NMF method while assuming non-linear model for show-through. Figure (c) confirms that using non-linear mixing model for show-through has improved the output results.

Therefore, updating matrices $\mathbf{A}$ and $\mathbf{S}$ in each iteration before projection can be written as

$$\begin{aligned} \mathbf{A}^{k+1} &= \mathbf{A}^k - \mu \nabla_{\mathbf{A}} J_2(\mathbf{S}^k, \mathbf{A}^k) \\ \mathbf{S}^{k+1} = \mathbf{S}^k - \mu \nabla_{\mathbf{S}} J_2(\mathbf{S}^k, \mathbf{A}^k) &= \mathbf{S}^k - \mu \left[(\mathbf{A}^k)^T \left(\mathbf{A}^k \mathbf{S}^k - \mathbf{X} \right) \right] - \mu \begin{bmatrix} w_1 \mathbf{v} \\ w_2 \mathbf{v} \end{bmatrix}. \end{aligned} \tag{9}$$

The reason that using cost function of (6) instead of (2) will compensate the degradations cause by linearly modeling show-through becomes clear from this equation and the definition of vector $\mathbf{v}$. In equation (9), in those pixels of scanned images where front and backside images are nearly black (*i.e.* $x_1(i)$ and $x_2(i)$ have small values), the third term in right-hand side of this equation, *i.e.* $\begin{bmatrix} v(i) & v(i) \end{bmatrix}^T$, will have a large value and therefore will decrease more $s_1(i)$ and $s_2(i)$ in these kind of pixels (will cause them to become more black). However, if at least one of the front or backside scanned images be nearly white (have high pixel value), $e^{-\alpha(x_1^2(i)+x_2^2(i))}$ and consequently $\begin{bmatrix} v(i) & v(i) \end{bmatrix}^T$ will be small and therefore the third term will have no effect on $s_1(i)$ and $s_2(i)$. This means that in those pixels where the writings of two sides of the paper overlap with each other, recovered images will be darker compared to those images recovered by using (3). But in those pixels where the writings of two sides of the paper do not overlap with each other, recovered images will have no difference from those obtained in previous section.

Figures 2(a) and 2(b) again show the result of applying the algorithm proposed in previous section to Figs. 1(a) and 1(b) but this time by using equation (9) instead of equation (3) (note that in this case, the other steps such as initialization and projection are remained unchanged). Figure 2(c) shows the same part previously shown in Fig. 1(e) for having a better comparison. As this image indicates, degradation caused by using linear mixing model for show-through (shown in Fig. 1(e) by arrows) has been removed considerably.



(a) Scan of the front side of the paper. (b) Scan of the back side of the paper.

(c) recovered front side image. (d) recovered backside side image.

Fig. 3. Show-through cancelation on actual scanned images. (a) and (b) show the scanned input images applied to our method (obtained from *http://www.site.uottawa.ca/∼edubois/documents*). Output images are shown in (c) and (d) indicate that this scheme can remove show-through almost completely.

4 Experimental Results

In this section, we will demonstrate the advantages of the proposed algorithm by performing an experiment on real scanned images. In the following simulation, σ and μ are set to 22 and 10^{-6} respectively. A suitable value for w_1 and w_2 is 6×10^5 which is obtained by trial and error and this value is used in all of the simulations performed in this paper. In the following experiments, we ran our proposed algorithm in MATLAB 7.1 on a Windows Vista PC with Intel 2.10 GHz Core 2 Duo CPU and 3 GB RAM. The results of the simulation that we performed by using the presented method are shown in Figs. 3(c) and 3(d). The original registered images (Figs. 3(a) and 3(b)) are ancient document which are degraded by show-through severely. Improvement in document readability is obvious in Figs. 3(c) and 3(d). The size of the input images was 150×570. For these particular input images, the algorithm converged after about 55 seconds.

Like BSS techniques, in NMF methods, separation is achieved up to some indeterminacies: The order of the sources and their amplitudes remain unknown. Scaling indeterminacy of sources is a drawback of our proposed method because regularization parameters w_1 and w_2 depend on the amplitudes of sources $\mathbf{s}_1$ and $\mathbf{s}_2$ which are unknown. Fortunately after performing some experiments, we recognized that in our case, amplitudes of the recovered sources converge to approximately the same values for different input images. We are not sure but

maybe, this is because of the added regularizing terms or because the input images are bounded between 0 and 255. As a result, it seems that fix values can be used for w_1 and w_2.

5 Conclusion

In this paper, we used gradient based non-negative matrix factorization algorithm for removing show-through. By performing an experiment, we showed the disadvantage of using linear mixing model for show-through and therefore, we changed it into the nonlinear mixing model simply by modifying the cost function used in NMF method. Finally, we justified the effectiveness of this new method by performing an experiment on actual scanned images.

References

1. Nishida, H., Suzuki, T.: 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)
2. Wang, Q., Tan, C.L.: Matching of Double-sided Document Images to Remove Interference. In: IEEE CVPR 2001 (December 2001)
3. Tonazzini, A., Bianco, G., Salerno, E.: Registration and enhancement of double-sided degraded manuscripts acquired in multispectral modality. In: 10th International Conference on Document Analysis and Recognition, Spain, July 2009, pp. 546–550 (2009)
4. Tonazzini, A., Salerno, E., Bedini, L.: Fast Correction of Bleed-through Distortion in Grayscale Documents by a Blind Source Separation Technique. Int. Journal of Document Analysis IJDAR 10(1), 17–25 (2007)
5. Tonazzini, A., Salerno, E., Mochi, M., Bedini, L.: Bleed-Through Removal from Degraded Documents Using a Color Decorrelation Method. In: Marinai, S., Dengel, A.R. (eds.) DAS 2004. LNCS, vol. 3163, pp. 229–240. Springer, Heidelberg (2004)
6. Merrikh-Bayat, F., Babaie-Zadeh, M., Jutten, C.: A Nonlinear Blind Source Separation Solution for Removing the Show-through Effect in the Scanned Documents. In: 16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland (August 2008)
7. Almeida, M.S.C., Almeida, L.B.: Separating nonlinear image mixtures using a physical model trained with ICA. In: IEEE International Workshop on Machine Learning for Signal Processing, Maynooth, Ireland (2006)
8. Sharma, G.: Cancellation of Show-through in Duplex Scanning. In: Proc. IEEE Int. Conf. Image Processing, vol.2, pp. 609–612 (September 2000)
9. Almeida, L.B.: Separating a Real-life Nonlinear Image Mixture. Journal of Macine Learning Research 6, 1199–1232 (2005)
10. Almeida, M.S.C., Almeida, L.B.: Wavelet-based separation of nonlinear show-through and bleed-through image mixtures. Journal of Neurocomputing 72, 57–70 (2008)
11. Chu, M., Diele, F., Plemmons, R., Ragni, S.: Optimality, Computation and Interpretation of Non-negative Matrix Factorizations. Wake Forest University (2004)
12. Lin, C.J.: Projected Gradient Methods for Non-negative Matrix Factorization. Neural Computation 19, 2756–2779 (2007)