REMOVAL OF BLEED-THROUGH EFFECT VIA NONNEGATIVE LEAST-CORRELATION

Ye Ai[†]*, Weifeng Li*[†]*, Tsung-Han Chan*[♯]*, and Qingmin Liao*[†]

[†] Department of Electronic Engineering / Graduate School at Shenzhen, Tsinghua University, China
[♯] Advanced Digital Sciences Center, Singapore

ABSTRACT

Bleed-through effect is one of the most common degradations in old documents, even in today's newspapers. This effect must be removed for hancing human and automatic readability. The two images scanned from the recto and verso pages of a document can be treated as a linear combination of clean text images from two sides. In this paper, we introduce a nonnegative least-correlation approach to demix bleed-through text images. Experiments have been conducted on both synthetic and real world images. In synthetic case, our method can recover the source text images exactly. Under real world conditions, our approach also performs well. In addition, our approach is computationally efficient and does not need any postprocessing task.

Index Terms— Bleed-through, linear mixing model, nonnegative blind source separation, nonnegative least-correlation.

1. INTRODUCTION

Ancient documents are widely used for studying of culture and other aspects. Access to these documents is often limited because of their fragile. The only way to read them is from the scanned digital images. Under this condition, the quality of their scanned images becomes important. However, bleed-through, one of the most common degradations in ancient documents, reduce the legibility significantly. This effect is the result of the seeping of ink from the reverse side of a paper, and interferes the main text on this side. Not only in ancient documents, but also in today's newspapers, we can see this effect and some similar degradation phenomena. To improve both human and automatic readability, bleed-through effect must be removed by image processing technique. Removing degradations from a scanned digital image is not a trivial task. There have been several studies on this problem. Some of them use single-side documents as their input. In [2], the authors used multistage thresholding to extract text in degraded documents. In [3], the authors described several commonly used methods based on various kinds of entropy. Hysteresis thresholding was used in [4]. Edge detection was used in [5].

For double side documents, In [6], the author developed a transmission model, and then the model was linearized using suitable transformations and simplifying approximations. Based on the linearized model, an adaptive linear filtering scheme is developed for the electronic removal of show-through. [7] and [8] used Principal Component Analysis (PCA) and Independent Component Analysis (ICA) method in a blind source separation frame to extract the main text from the other interference patterns. In [9], the authors proposed a linear model based on a blind source separation technique. Hidden Markov Random Fields [10], neural networks [11] and a new flow filed model [12] are also used.

Studies based on thresholding [2] always relies on the assumption that gray levels of the main text and the background are separable. However, unlike other thresholding task, gray levels of main text and reverse side's text in bleed-through documents are usually hard to distinguish and perfect thresholds can hardly be found, or they may not exist actually. Meanwhile, computational cost of statistic models [10] is very high, and separation result is not so good. As well, statistic method usually need other postprocessing tasks such as removing residual information and connecting the gaps.

In this paper, we propose a nonnegative least-correlation approach to recover two sides of bleed-through documents. We treat the two observation images (i.e. the scanned digital images) as a linear combination of two clean text images. Since we have no knowledge of clean sources or the mixing coefficients, this is the so-called blind source separation problem. Our model is linear and based on convex analysis framework [13]. The included angle of two observations represents the correlation between them. We design an optimal demixing matrix $\mathbf{W}$ to demix the bleed-through images and estimate the true text sources via minimizing the included angle within the intersection of observations' affine hull and nonnegative signal space. This approach has a close-form solution and is computationally efficient.

2. ALGORITHM

2.1. Problem Formulation

Consider the following linear mixing model in common blind source separation tasks:

$$\mathbf{x}[n] = \mathbf{A}\mathbf{s}[n], \quad n = 1, 2, \ldots, L, \tag{1}$$

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

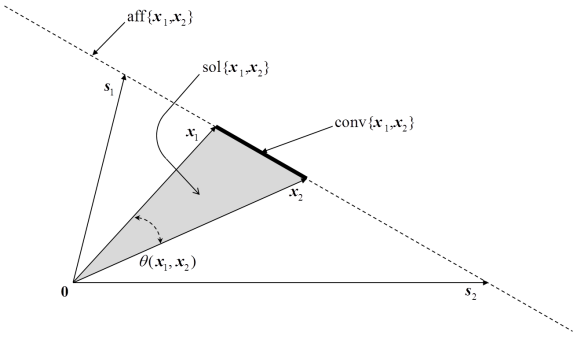

Fig. 1. Geometric illustration of the text sources $\boldsymbol{s}_1, \boldsymbol{s}_2$ and the observations $\boldsymbol{x}_1, \boldsymbol{x}_2$.

where $\mathbf{x}[n] = (x_1[n], x_2[n])^T$ is the n-th data point of the given two observations, $\mathbf{A} = [a_{ij}]_{2\times 2} \in \mathbb{R}^{2\times 2}$ is an unknown mixing matrix, $\mathbf{s}[n] = (s_1[n], s_2[n])^T$ is the nth source point consisting of two sources, and L is the number of pixels in each observation and each source.

In our model, $\boldsymbol{x}_i = (x_i[1], x_i[2], \ldots, x_i[L])^T$, $i = 1, 2$, is the grayscale image obtained from the recto and verso pages of a document, respectively. Suppose that they have been registered. We consider these two observations as a linear combination of clean text image $\boldsymbol{s}_i = (s_i[1], s_i[2], \ldots, s_i[L])^T$, $i = 1, 2$. Our goal is to find a demixing matrix $\boldsymbol{W}$ such that

$$\mathbf{y}[n] = \mathbf{W}\mathbf{x}[n] = \mathbf{W}\mathbf{A}\mathbf{s}[n] = \mathbf{P}\mathbf{s}[n], \tag{2}$$

where $\mathbf{y}[n] = (y_1[n], y_2[n])^T$ is the n-th extracted source point and $\mathbf{P} = \mathbf{W}\mathbf{A} \in \mathbb{R}^{2\times 2}$ is a permutation matrix. After that we have the extracted source vector $\boldsymbol{s}_i = (s_i[1], s_i[2], \ldots, s_i[L])^T$, $i = 1, 2$. If we design $\mathbf{W}$ properly, we will reduce the bleed-through effect.

2.2. Nonnegative Least-correlation Demixing

Let us introduce some basic sets in convex analysis which are really important in following approach. The convex hull, the affine hull and the solid region of two vectors is defined in [13] as

$$\begin{aligned}
\text{conv}\{\boldsymbol{z}_1, \boldsymbol{z}_2\} &= \{\boldsymbol{z} \mid \boldsymbol{z} = \textstyle\sum_{i=1}^2 \alpha_i \boldsymbol{z}_i, \sum_{i=1}^2 \alpha_i = 1, \alpha_i \in \mathbb{R}_+\},\\
\text{aff}\{\boldsymbol{z}_1, \boldsymbol{z}_2\} &= \{\boldsymbol{z} \mid \boldsymbol{z} = \textstyle\sum_{i=1}^2 \alpha_i \boldsymbol{z}_i, \sum_{i=1}^2 \alpha_i = 1, \alpha_i \in \mathbb{R}\},\\
\text{sol}\{\boldsymbol{z}_1, \boldsymbol{z}_2\} &= \{\boldsymbol{z} \mid \boldsymbol{z} = \textstyle\sum_{i=1}^2 \alpha_i \boldsymbol{z}_i, \sum_{i=1}^2 \alpha_i \le 1, \alpha_i \in \mathbb{R}_+\}.
\end{aligned}$$

Geometric illustration of the text sources $\boldsymbol{s}_1, \boldsymbol{s}_2$ and the observations $\boldsymbol{x}_1, \boldsymbol{x}_2$ is in Fig. 1.The four assumptions in [1] are adopted:

- *A1*) For each i, $\boldsymbol{s}_i \in \mathbb{R}_+^L$;
- *A2*) $\mathbf{A} \in \mathbb{R}_+^{2\times 2}$;
- *A3*) rank($\mathbf{A}$) = 2;
- *A4*) Each row sum of $\mathbf{A}$ equals one.

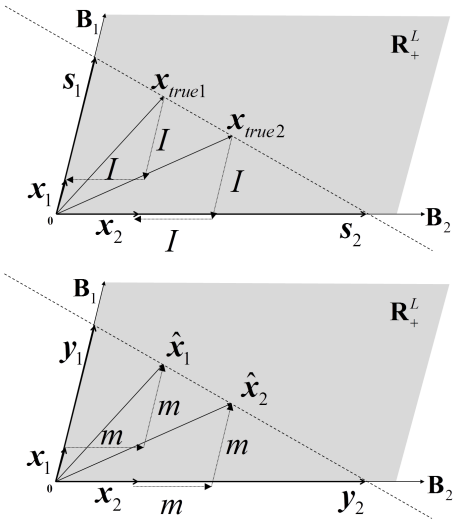

Fig. 2. Geometric illustration of the situation with interference.

These assumptions are easilly satisfied in our problem. Under these assumptions, we can prove that

$$\begin{aligned}
\text{conv}\{\boldsymbol{x}_1, \boldsymbol{x}_2\} &\subseteq \text{conv}\{\boldsymbol{s}_1, \boldsymbol{s}_2\},\\
\text{sol}\{\boldsymbol{x}_1, \boldsymbol{x}_2\} &\subseteq \text{sol}\{\boldsymbol{s}_1, \boldsymbol{s}_2\}.
\end{aligned}$$

The correlation coefficient between observations $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ is known as

$$0 < \rho(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{\boldsymbol{x}_2^T \boldsymbol{x}_2}{\|\boldsymbol{x}_1\| \cdot \|\boldsymbol{x}_2\|} = \cos(\theta(\boldsymbol{x}_1, \boldsymbol{x}_2)) \le 1. \tag{3}$$

Then $\rho(\boldsymbol{x}_1, \boldsymbol{x}_2)$, the correlation between observations, is never smaller than $\rho(\boldsymbol{s}_1, \boldsymbol{s}_2)$, that between true sources, for all $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \text{conv}\{\boldsymbol{s}_1, \boldsymbol{s}_2\}$. But demixing matrix $\mathbf{W} \notin \mathbb{R}_+^{2\times 2}$ and the extracted sources $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are in the affine hull of observations, not the convex hull. Thus $\mathbf{W}$ can be designed such that $\rho(\boldsymbol{y}_1, \boldsymbol{y}_2) < \rho(\boldsymbol{x}_1, \boldsymbol{x}_2)$. Moreover, if we choose $\mathbf{W}$ more wisely, $\rho(\boldsymbol{y}_1, \boldsymbol{y}_2) = \rho(\boldsymbol{s}_1, \boldsymbol{s}_2)$ can be achieved. Under this motivation, we consider the correlation coefficient between extracted sources $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ as our criterion. We can design $\mathbf{W}$ by minimizing the correlation coefficient,

$$\begin{aligned}
&\min_{\mathbf{W}\in\mathbb{R}^{2\times 2}} \rho(\boldsymbol{y}_1, \boldsymbol{y}_2),\\
s.t.\quad &\boldsymbol{y}_i = \sum_{j=1}^{2} w_{ij}\boldsymbol{x}_j \succeq 0, i = 1, 2, \mathbf{W}\mathbf{1}_2 = \mathbf{1}_2.
\end{aligned} \tag{4}$$

Then the removal of bleed-through convert to an optimization problem, and this optimization problem has a closed-form solution [1]:

$$
\begin{aligned}
w_{11}^{\star} &= \max_{n} \left\{ \frac{-x_2[n]}{x_1[n]-x_2[n]} \mid x_1[n] > x_2[n], \forall n \right\}, \\
w_{21}^{\star} &= \min_{n} \left\{ \frac{-x_2[n]}{x_1[n]-x_2[n]} \mid x_1[n] < x_2[n], \forall n \right\}, \\
w_{12}^{\star} &= 1 - w_{11}^{\star}, \\
w_{22}^{\star} &= 1 - w_{21}^{\star}.
\end{aligned} \tag{5}
$$

Actually the process of our method in absence of strong interference is quite brief. Given the two side bleed-through document, we compute the demixing matrix $\mathbf{W}$ using Eq. (5) and then demix the observations to obtain the text sources. We will illustrate how to handel the situation with strong interferences.

2.3. Anti-interference Factor

Noise is an inescapable problem in linear model. Since the extracted source in our method is the intersection of affine hull aff$\{\boldsymbol{s}_1, \boldsymbol{s}_2\}$ and nonnegative signal space $\mathbb{R}_+^L$, if the pure-source samples exist, i.e, there exists at least one index set$\{l_1, l_2\}$ such that $\mathbf{s}[l_i] = s_i[l_i]\mathbf{e}_i, i = 1, 2$, the true source will lie in the boundry of nonnegative signal space. In this situation we can recover the source images exactly [1]. This condition is easy to be satisfied when there is no strong interference such as bad illumination. However, in noisy case, observation vector may be pulled to the coordinate plane of nonnegative space $\mathbb{R}_+^L$ not where they are actually supposed to be. In our assumptions, the signal lying in the coordinate plane is considered to be the true source, thus our approach will fail under this condition.

Now we focus on this worst situation. Our method relies on the correlation between signals. Suppose the effect of interference is only darkening the whole observation signals after mixing while remaining the structure of signals. We make this assumption because this interference will right lead to the proximity of observation signal and the boundry of nonnegative space $\mathbb{R}_+^L$. When the observation signal intersect any coordinate plane in $\mathbb{R}^L$, our method would treat it as true source signal while they are actually not, and thus we fail to estimate the actual text image. To solve this interference-caused problem, we introduce an anti-interference factor m. When our approach in ideal situation is invalid, we shift each pixel of observation vector via adding a proper m such that the observation signal will get out of the boundry plane, and our approach can keep on decorrelating the observations to estimate the true sources. Fig. 2 demonstrates this situation, where $\mathbb{R}_+^L$ is the nonnegative signal space, $\mathbf{B}_1$ and $\mathbf{B}_2$ are the boundry plane of $\mathbb{R}_+^L$. The true sources $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ are mixed under unknown coefficients to obtain the true mixing results $\boldsymbol{x}_{true1}$ and $\boldsymbol{x}_{true2}$. However, the unknown interference I produces the actual observations $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. Our anti-interference factor m pushes the actual observations out of the boundry plane and thus we obtain the estimated sources $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$. This factor should not be too large. If it is, the correlation between extracted sources will be eliminated exceedingly. The spatial crossing of texts from two sides will be removed. For this reason, in following experiments of real world bleed-through documents, we will use a smaller m.

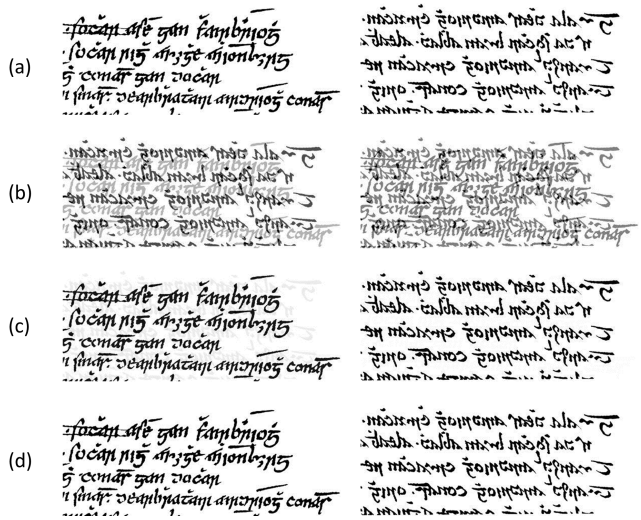

Fig. 3. Results of different approachs on synthetic images. (a) groundtruth images, (b) observation images, (c) ICA's result, (d) our result.

2.4. Related to Other Works

We use a linear mixing model and blind source separation technique which is unusual in bleed-through problem. In [7], [8] and [9], the authors solve this bleed-through problem using ICA. Differ from them, we do not assume the independence of source signals. We use convex analysis to obtain a criterion of designing demixing matrix $\mathbf{W}$. Even though the source signals do not lie in the boundry plane of $\mathbb{R}_+^L$, our approach can decorrelate bleed-through image and obtain clean source text images as well.

3. EXPERIMENTS

3.1. Database

We use the bleed-through database of Irish Script On Screen Project (www.isos.dias.ie). This database consists of a set of 25 recto and verso sample grayscale image pairs, taken from larger manuscript images, with varied degrees of bleed-through. The verso side of each image pair has been flipped horizontally and registered to the recto side. It also contains the groudtruth images.

Table 1. Quantized evaluation of our method.

	Huang's method[14]	Moghaddam' method[12]	our method(m=0)	our method
D_r	7904.2	5570.6	5618.1	4652.6
D_v	8089.0	5671.3	5528.4	5032.8

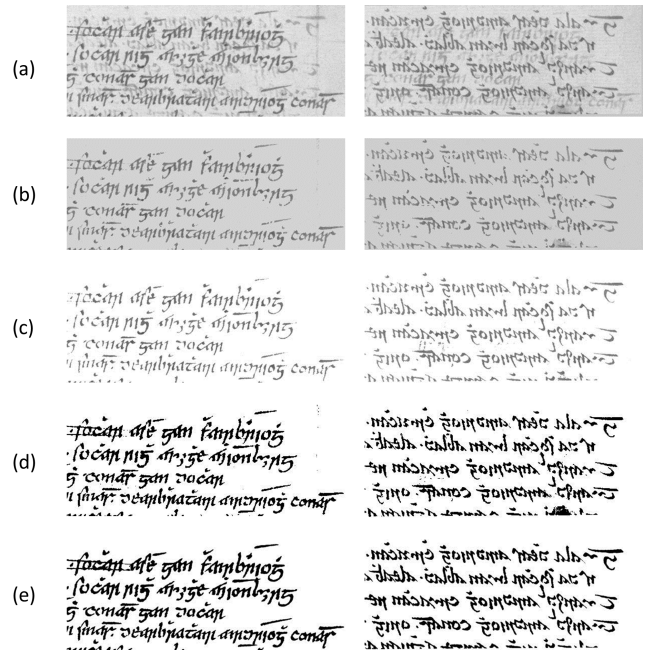

Fig. 4. Results of different approachs. (a) observation images suffer from bleed-through, (b) Huang's result [14], (c) Moghaddam's result [12], (d) our result, (e) groundtruth images.

3.2. Settings and Results

In this section, two kinds of experiment have been conducted. First, we test our method on synthetic images obtained from groundtruth in order to see whether we can recover the text images exactly or not in the ideal situation. We generate the first coefficient of each row in mixing matrix **A** from a uniform distribution between 0 and 1. Then we subtract this coefficient from 1 to obtain another two coefficients. After that, we use **A** and groundtruth images to synthesis observation images. Finally we demix the observations to obtain source images. Fig. 3 shows the result. Unlike ICA, our approach can restore source images exactly.

We next move to the real world case. The recto and verso images are tranformed into intensity vectors, and then our approach can be applied. An example of separation results using different methods is shown in Fig. 4. By using our proposed method, texts on one side has been separated from its reverse side without any prejudice on their crossing.

Some images suffer from a strong interference and our method can not yield a good result, as we mentioned above.

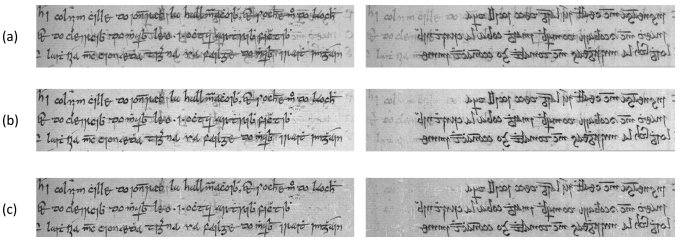

Fig. 5. Effect of the anti-interference factor m. (a) observation images suffer from bleed-through, (b) result images without m, (c) result images with $m = 40$.

Images in Fig. 5 show the effect of our anti-interference factor. Because of the noise-caused proximity to nonnegative space bound, extracted images in second row still contain slight bleed-through. Third row in figure is the separating result with $m = 40$, the bleed-through is further eliminated.

In order to evaluate the effect of our approach, we transform the groundtruth, our results, the results obtained by the method in [14], and the results obtained by the method in [12] to binary image by a simple Otsu thresholding. Then we calculate the Euclidean distance between results and groundtruth, i.e. the number of different pixels. Finally, mean value of the distance of 25 pairs will be taken:

$$\begin{aligned} D_r &= \frac{1}{25}\sum_{n=1}^{L}|y_r[n] - g_r[n]|, \\ D_v &= \frac{1}{25}\sum_{n=1}^{L}|y_v[n] - g_v[n]|, \end{aligned} \tag{6}$$

where $\boldsymbol{g}_r$ and $\boldsymbol{g}_v$ are the binary groundtruth images, $\boldsymbol{y}_r$ and $\boldsymbol{y}_v$ are estimated binary text source images and L is the number of pixels. The smaller the distance is, the closer the result is to groundtruth. We set a proper unitive m for all 25 pairs of bleed-through images. Table 1 shows the evaluation result. Our approach performs better than other methods.

4. CONCLUSIONS

We have presented a new blind source separation method on bleed-through removal problem. We treat the whole process as a linear mixing model and solve this problem by compute the optimal demix matrix **W** under the nonnegative least-correlation criterion. As a countermeasure of other strong interferences that may cause our model to fail, we introduce

an anti-interference factor m which has turned out to be successful. Experiments shows that our approach is effective in bleed-through situation and quantized evaluation has also been presented. Adaptive algorithm of the anti-interference factor m is left in future research.

5. REFERENCES

[1] Fa-Yu Wang, Chong-Yung Chi, Tsung-Han Chan, and Yue Wang, "Nonnegative least-correlated component analysis for separation of dependent sources by volume maximization," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 5, pp. 875–888, 2010.

[2] Graham Leedham, Saket Varma, Anish Patankar, and Venu Govindaraju, "Separating text and background in degraded document images-a comparison of global thresholding techniques for multi-stage thresholding," in *Frontiers in Handwriting Recognition, 2002. Proceedings. Eighth International Workshop on*. IEEE, 2002, pp. 244–249.

[3] João Marcelo Monte da Silva, Rafael Dueire Lins, Fernando Mário Junqueira Martins, and Rosita Wachenchauzer, "A new and efficient algorithm to binarize document images removing back-to-front interference.," *J. UCS*, vol. 14, no. 2, pp. 299–313, 2008.

[4] Rolando Estrada and Carlo Tomasi, "Manuscript bleed-through removal via hysteresis thresholding," in *Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on*. IEEE, 2009, pp. 753–757.

[5] Chew Lim Tan, Ruini Cao, Peiyi Shen, Qian Wang, Julia Chee, and Josephine Chang, "Removal of interfering strokes in double-sided document images," in *Applications of Computer Vision, 2000, Fifth IEEE Workshop on.* IEEE, 2000, pp. 16–21.

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

[7] Ivan Gerace, Francesco Cricco, and Anna Tonazzini, "An extended maximum likelihood approach for the robust blind separation of autocorrelated images from noisy mixtures," in *Independent Component Analysis and Blind Signal Separation*, pp. 954–961. Springer, 2004.

[8] Emanuele Salerno, Anna Tonazzini, and Luigi Bedini, "Digital image analysis to enhance underwritten text in the archimedes palimpsest," *International Journal of Document Analysis and Recognition (IJDAR)*, vol. 9, no. 2-4, pp. 79–87, 2007.

[9] Anna Tonazzini, 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.

[10] Christian Wolf, "Document ink bleed-through removal with two hidden markov random fields and a single observation field," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 3, pp. 431–447, 2010.

[11] Xiaowei Zhang, Jianming Lu, and Takashi Yahagi, "Blind separation methods for image show-through problem," in *Information Technology Applications in Biomedicine, 2007. ITAB 2007. 6th International Special Topic Conference on*. IEEE, 2007, pp. 255–258.

[12] Reza Farrahi Moghaddam and Mohamed Cheriet, "A variational approach to degraded document enhancement," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 8, pp. 1347–1361, 2010.

[13] Mokhtar S Bazaraa, Hanif D Sherali, and Chitharanjan Marakada Shetty, *Nonlinear programming: theory and algorithms*, John Wiley & Sons, 2013.

[14] Yi Huang, Michael S Brown, and Dong Xu, "User-assisted ink-bleed reduction," *Image Processing, IEEE Transactions on*, vol. 19, no. 10, pp. 2646–2658, 2010.