Reduction of bleed-through in scanned manuscript documents

Eric Dubois and Anita Pathak
School of Information Technology and Engineering,
University of Ottawa, Ottawa, ON Canada K1N 6N5

Abstract

Many old manuscript documents were written on both sides of the paper, and the bleed-through from one side of the document to the other increases the difficulty in reading or deciphering the information on the page. This paper presents techniques for reducing such bleed-through distortion using techniques of digital image processing. Both sides of the document are scanned, maintaining full spatial and amplitude resolution (8 bits/sample). The bleed-through is reduced by processing both sides of the document simultaneously. First the verso side is ßipped from left to right, and then the recto and flipped verso images are registered. This registration is necessary since it is impossible to perfectly align the front and back when scanning the document, and the scanner may not be perfectly uniform. We used a six-parameter affine transformation to register the two sides, determining the parameters using an optimization method. Once the two sides have been registered, areas consisting primarily of bleed-through are identified and replaced by the background color or intensity. The method has been tested on a number of documents, including documents we generated under controlled conditions and some original manuscripts; the readability of documents with heavy bleed-through has been greatly improved by this method.

1. Introduction

Many documents written or printed on both sides of the page suffer from bleed-through which can significantly impair the readability of the document. Fig. 1 shows an extract of the corresponding portions of the front (recto) and rear (verso) of a typical eighteenth century manuscript document, where the bleed-through clearly makes the task of reading the document more challenging and fatiguing. There is thus great interest in removing this bleed-through using digital image processing techniques. Since the darkness of some of the bleed-through is comparable to the darkness of some of the desired writing, a simple thresholding operation will not be successful in removing the bleed-through. However, by processing both sides of the document together, it is possible to identify regions of the image that are due to bleed-through and replace them with an estimate of the background. Techniques of this type are reported in [1, 2] for reducing show-through in scanned documents; the basic idea is presented in [1] and a restoration technique using adaptive filtering is presented in [2]. In order to adequately remove bleed-through, the recto and left-right flipped verso images must be registered; this did not receive much attention in [1, 2]. This paper presents a method to carry out this registration along with a proposed method to reduce the bleed-through. Section 2 presents the general formulation of the problem and describes the registration and bleed-through removal algorithms. Section 3 gives experimental results with a test document, followed by conclusions in Section 4.

2. Bleed-through Removal Algorithm

2.1. Assumptions

In this paper, we assume that the original document consists of some type of paper on which ink has been applied to both sides, either through writing or printing. Ink may simply show through from one side to the other, or it may have actually "bled" through to the other side. Both sides of the document are digitized in order to apply the bleed-through removal algorithm. The sampled recto and verso images are denoted $f_r(x, y)$ and $f_v(x, y)$ respectively, where the sample points (x, y) lie on a two-dimensional rectangular sampling structure $\mathcal{L}$. In this article, we assume that 8-bit gray-scale versions of the image of size pw by ph are acquired, which are normalized such that $0 \leq f_r(x, y) \leq 1$ and $0 \leq f_v(x, y) \leq 1$ with gray-level 0 corresponding to white and 1 corresponding to black. Color information may be helpful and will be addressed in future work.

We assume that there exist *ideal* recto and verso images representing the writing applied to the front and the back of the paper, denoted $f_{wr}(x, y)$ and $f_{wv}(x, y)$ respectively; these are zero where there is no writing. Similarly, we assume that there is an ideal background $f_{br}(x, y)$ and $f_{bv}(x, y)$ corresponding to the image of the paper without writing. The measured recto image combines the back-

Figure 1: Extracts of recto and verso sides of a document with significant bleed-through.

ground, the ideal recto image, and the left-right flipped ideal verso image in some way:

$$f_r(x,y) = \mathcal{C}(f_{br}(x,y), f_{wr}(x,y), \mathcal{R}f_{wv}(x,y)) \quad (1)$$

where $\mathcal{R}$ is a linear operator that performs a left-right ßip of an image:

$$g = \mathcal{R}f: \qquad g(x,y) = f(pw - x, y). \quad (2)$$

A simple additive model would be

$$f_r(x,y) = f_{br}(x,y) + f_{wr}(x,y) + \alpha\mathcal{R}f_{wv}(x,y) \quad (3)$$

where α represents attenuation of the verso writing in the bleed-through process. This model is probably reasonably accurate except where f_{wr} and $\mathcal{R}f_{wv}$ overlap. Another possible model is

$$f_r(x,y) = \max(f_{br}(x,y), f_{wr}(x,y), \alpha\mathcal{R}f_{wv}(x,y)). \quad (4)$$

The measured verso image is obtained in a similar fashion. However, since the two sides are scanned in separate operations, the two scanning rasters will not be aligned; they will differ by some offset, rotation and possible skew. The ideal measured verso image with perfect registration is given by

$$f_v^I(x,y) = \mathcal{C}(f_{bv}(x,y), f_{wv}(x,y), \mathcal{R}f_{wr}(x,y)) \quad (5)$$

where the coordinate systems of the recto and verso are perfectly aligned. However, the actual measured verso image is

$$f_v(x,y) = \mathcal{A}_p f_v^I(x,y) \quad (6)$$

where $\mathcal{A}_p$ is a linear operator that models the geometric distortion between the two scanning lattices. In our work, we have assumed that $\mathcal{A}_p$ is an affine transformation specified by six parameters $\boldsymbol{p} = (p_{11}, p_{12}, p_{13}, p_{21}, p_{22}, p_{23})$ defined by

$$g = \mathcal{A}_p f: \quad g(x,y) = f(p_{11}x + p_{12}y + p_{13}, p_{21}x + p_{22}y + p_{23}). \quad (7)$$

2.2. Problem formulation

With these assumptions, the bleed-through removal problem can be stated as follows: given the sampled recto and verso images f_r and f_v, estimate the restored images

$$\hat{f}_r(x,y) = \mathcal{C}(f_{br}(x,y), f_{wr}(x,y), 0) \quad (8)$$

$$\hat{f}_v(x,y) = \mathcal{C}(f_{bv}(x,y), f_{wv}(x,y), 0). \quad (9)$$

The problem can be broken down into two steps:

1. Estimate f_v^I using f_v and f_r (registration).
2. Estimate $\hat{f}_r$ and $\hat{f}_v$ from f_r and $\hat{f}_v^I$ (restoration).

2.3. Registration

The registration problem is to estimate the ideal verso image $f_v^I(x,y)$ of Eq. 5 using the observed recto and verso images $f_r(x,y)$ and $f_v(x,y)$. In our case, registration would involve shifts of at most a few millimeters, rotations of at most a few degrees, and possibly some mild geometric distortion, which is well suited to the affine model. From Eq. 6, we see that $f_v^I = \mathcal{A}_p^{-1} f_v$. If the matrix

$$\begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix}$$

is non-singular, then the inverse operator $\mathcal{A}_p^{-1}$ exists and is also an affine transformation operator, say $\mathcal{A}_s$, where the

parameters of $\mathcal{A}_s$ are given by

$$\begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix}^{-1} \tag{10}$$

$$\begin{bmatrix} s_{13} \\ s_{23} \end{bmatrix} = - \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix}^{-1} \begin{bmatrix} p_{13} \\ p_{23} \end{bmatrix} \tag{11}$$

Thus it is equivalent to estimate the parameter vector $\boldsymbol{s}$ of the inverse affine transformation $\mathcal{A}_s = \mathcal{A}_p^{-1}$.

Assume the simple linear combining model for $\mathcal{C}$. Then

$$f_r = f_{br} + f_{wr} + \alpha \mathcal{R} f_{wv} \tag{12}$$

$$f_v^I = f_{bv} + f_{wv} + \alpha \mathcal{R} f_{wr} \tag{13}$$

$$f_v = \mathcal{A}_p f_{bv} + \mathcal{A}_p f_{wv} + \alpha \mathcal{A}_p \mathcal{R} f_{wr} \tag{14}$$

The registration problem amounts to aligning the writing in the flipped verso with the bleed-through in the recto, and the bleed-through in the flipped verso with the writing on the recto. Let $\mathcal{A}_q$ be a candidate or trial for the inverse transformation $\mathcal{A}_p^{-1}$. Then

$$\mathcal{A}_q f_v = \mathcal{A}_q \mathcal{A}_p f_{bv} + \mathcal{A}_q \mathcal{A}_p f_{wv} + \alpha \mathcal{A}_q \mathcal{A}_p \mathcal{R} f_{wr} \tag{15}$$

and so

$$\mathcal{R}\mathcal{A}_q f_v = \mathcal{R}\mathcal{A}_q \mathcal{A}_p f_{bv} + \mathcal{R}\mathcal{A}_q \mathcal{A}_p f_{wv} + \alpha \mathcal{R}\mathcal{A}_q \mathcal{A}_p \mathcal{R} f_{wr}. \tag{16}$$

The difference between the recto image and the flipped transformed verso is thus

$$\begin{aligned} f_r - \mathcal{R}\mathcal{A}_q f_v = (f_{br} - \mathcal{R}\mathcal{A}_q \mathcal{A}_p f_{bv}) + \\ (\mathcal{I} - \alpha \mathcal{R}\mathcal{A}_q \mathcal{A}_p \mathcal{R}) f_{wr} + (\alpha \mathcal{R} - \mathcal{R}\mathcal{A}_q \mathcal{A}_p) f_{wv}. \end{aligned} \tag{17}$$

It can be shown that under mild assumptions $\|f_r - \mathcal{R}\mathcal{A}_q f_v\|^2$ will be minimized when $\mathcal{A}_q = \mathcal{A}_p^{-1}$, where

$$\|g\|^2 = \sum_{(x,y) \in \mathcal{L}} (g(x,y))^2.$$

Then, the parameters of the inverse affine transformation operator can be estimated by solving the optimization problem

$$\widehat{\boldsymbol{s}} = \arg\min_{\boldsymbol{q}} \|f_r - \mathcal{R}\mathcal{A}_q f_v\|^2 \tag{18}$$

so that

$$\hat{f}_v^I = \mathcal{A}_{\hat{s}} f_v. \tag{19}$$

Note that $\mathcal{A}_q f_v$ involves computing values of f_v that are not on the original sampling structure $\mathcal{L}$. These samples must be computed with a suitable interpolation scheme such as bilinear or bicubic interpolation.

2.4. Restoration

Once we have registered the recto and the ßipped verso images, we must estimate $\hat{f}_r$ and $\hat{f}_v$ using f_r and $\hat{f}_v^I$. We discuss the procedure to estimate $\hat{f}_r$; the procedure to determine $\hat{f}_v$ is similar. Our approach is to identify the regions of the recto image corresponding to bleed-through only, and replace these areas with an estimate of the background.

The observed recto image can be partitioned into four regions characterized as follows:

A There is recto writing but no bleed-through. In this case $f_r(x,y)$ should have a high value and $\mathcal{R} f_v^I(x,y) \approx \alpha f_r(x,y)$.

B There is bleed-through but no recto writing. In this case $f_r(x,y) \approx \alpha \mathcal{R} f_v^I(x,y)$ where $\mathcal{R} f_v^I(x,y)$ is a high value.

C There is no recto writing and no bleed-through. Here both $f_r(x,y)$ and $\mathcal{R} f_v^I(x,y)$ are low values corresponding to background.

D There is both recto writing and bleed-through. In this case $f_r(x,y) \approx \mathcal{R} f_v^I(x,y)$ which are high values.

Based on these considerations, we can identify that a pixel (x,y) belongs to category B if $\mathcal{R}\hat{f}_v^I(x,y) > T$ AND $f_r(x,y)/\mathcal{R} f_v^I(x,y) < \alpha_0$, where T and α_0 are suitably chosen thresholds. If this condition is detected, the pixel is replaced with an estimate of the background intensity, which can be obtained using median filtering for example.

$$\hat{f}_r(x,y) = \begin{cases} \hat{f}_{br}(x,y) & \text{if } (x,y) \in B, \\ f_r(x,y) & \text{otherwise.} \end{cases} \tag{20}$$

3. Experimental Results

We show here some results on a simple test document shown in Fig. 2 exhibiting significant bleed-through. By superimposing the recto and the flipped verso on the same image in Fig. 3, we see clearly the lack of registration. After applying the registration technique described above, the recto and the flipped verso are well aligned, as shown in Fig. 4. Finally, using the restoration algorithm described above, we obtain the recto image with suppressed bleed-through shown in Fig. 5.

4. Conclusion

Effective methods for reduction of bleed-through can greatly facilitate the task of reading old documents for the historical researcher. They can also improve the performance of

compression schemes which need not waste bits on the undesired bleed-through components. We have demonstrated an effective method to reduce bleed-through in this paper by registering the recto and the flipped verso, and using a threshold-based test to replace bleed-through with a background level. In further work, we will incorporate color information, and combine the technique with document compression algorithms.

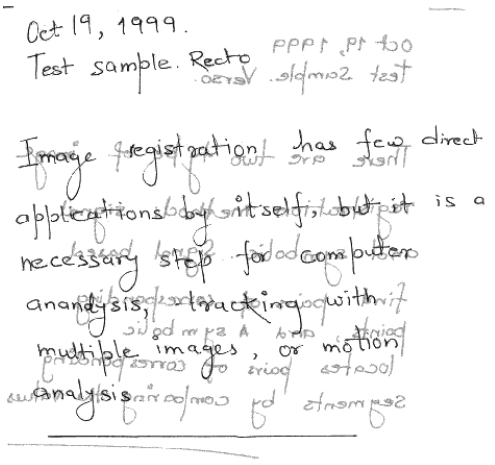

Figure 2: *Recto side of a test document with significant bleed-through.*

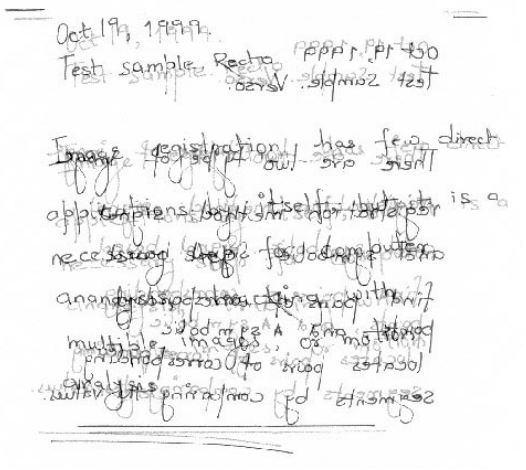

Figure 3: *Superimposed recto and verso sides of test document without registration.*

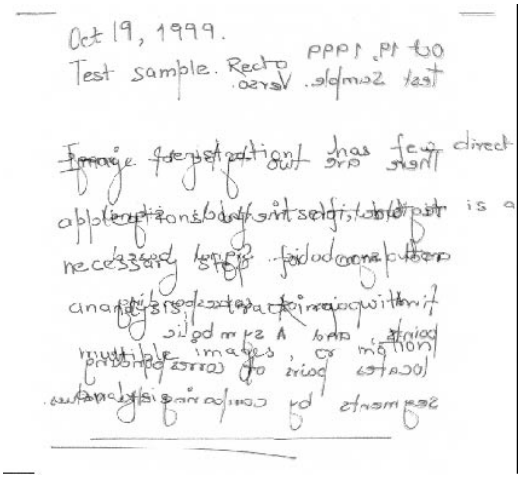

Figure 4: *Superimposed recto and verso sides of test document with registration.*

References

[1] K. Knox, "Show-through correction for two-sided documents." United States Patent 5,832,137, Nov. 1998.

[2] G. Sharma, "Cancellationof show-through in duplex scanning," in *Proc. IEEE Int. Conf. Image Processing*, vol. 2, pp. 609Ð612, Sept. 2000.

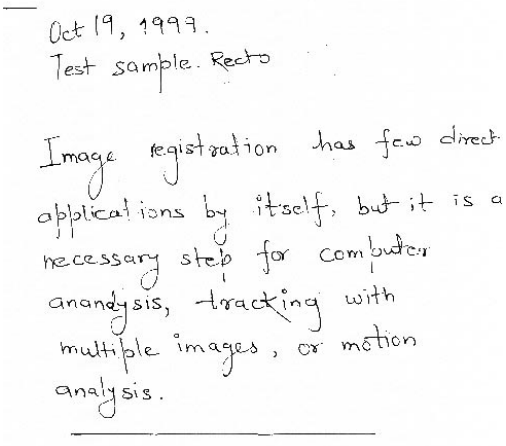

Figure 5: *Recto image with suppression of bleed-through.*