Joint Compression and Restoration of Documents with Bleed-through

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

Abstract

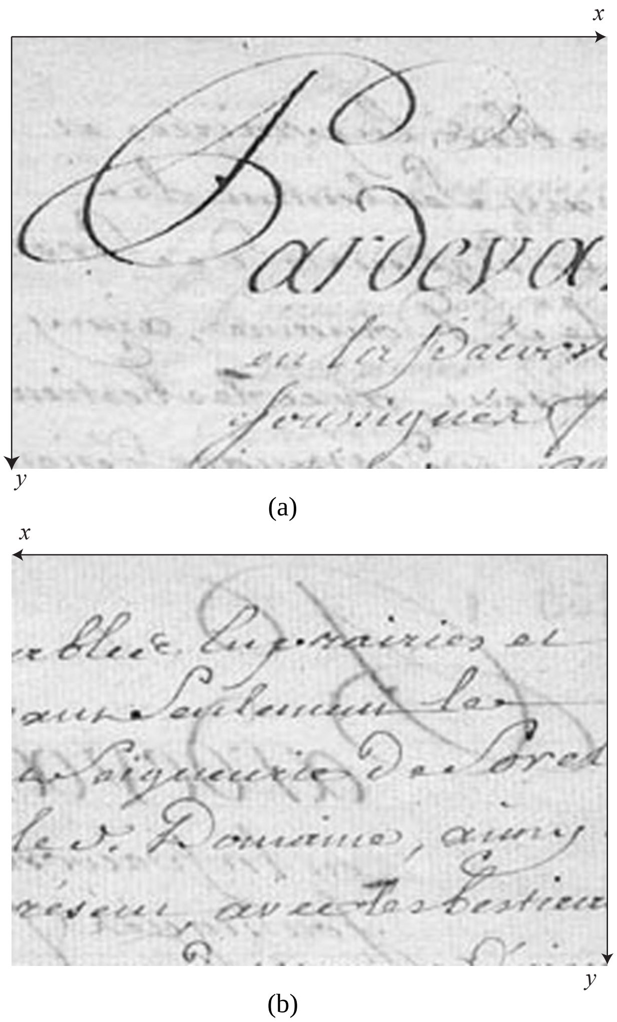
This paper presents research on the digital restoration of scanned two-sided documents suffering from bleed-through, and the joint compression of the original document and its bleed-through corrected version. It is often required to have easy and efficient access to both the original document and the restored version. The method works simultaneously on both the recto and the verso sides, and requires four steps: i) registration, ii) segmentation, iii) inpainting and iv) compression. The first step involves the registration of the recto and the flipped verso so that bleed-though on one side will be aligned with the original information (called foreground) on the other side. An optimization method based on an affine transformation is used for this step. The second step requires segmentation of each side into four regions: 'foreground only', 'background only', 'bleed-through only' and 'mixed bleed-through and foreground'. Then, the areas identified as 'bleed-through only' are replaced with an estimate of the background using an inpainting technique. Finally the two-sided image is compressed for efficient storage and transmission. Each side is first compressed using any standard efficient document compression algorithm such as JPEG 2000. Then the segmentation information identifying the region 'bleed-through only' on each side is compressed using a standard bilevel compression algorithm such as JBIG2. The information required to represent the inpainted sections must also be transmitted.

Introduction

Bleed-through occurs on documents that are written on both sides of the sheet of paper. The ink from the reverse side of the document may have seeped through the paper over time, or it may simply show through the paper. Such bleed-through can significantly impair the readability of the document and also cause visual fatigue for the reader. Thus, there is great interest in digital techniques to reduce the visibility of such bleed-through. Fig. 1 shows an extract of the recto and verso sides of a typical eighteenth century document exhibiting significant bleed-through. Since the darkness of the bleed-through may be comparable to, or even greater than, that of some parts of the desired recto writing, a simple thresholding operation cannot be used to remove bleed-though. However, by processing both sides of the document simultaneously, it is possible to identify regions that are mainly bleed-though, and replace them by an estimate of the background. Since the algorithms for bleed-through reduction may be complex, it would be preferable to implement them at the site of the digital archive. Then a remote user could access either the bleed-through corrected version (for easier reading) or the original uncorrected version (for authenticity). However, a researcher may want access to both versions. Since they are very similar, it is not necessary to independently download both versions, as shown in this paper. Additional information can be sent to generate the corrected image from the original uncorrected image with little extra cost.

Several approaches to bleed-through and show-through reduction have been investigated. Several authors have used techniques involving one side of the document only ([1], [2], [3]) using various features that distinguish bleed-through from foreground writing. While these certainly perform better than simple thresholding, there is no way to unambiguously differentiate foreground from bleed-through without comparing both sides of the document. Knox [4] and Sharma [5] assume that the distortion is due to show-through and that the impairment when scanning such a document can be well modelled by the properties of the physical scanning process. Then, the show-through can be cancelled using adaptive filtering techniques [5]. However, this model does not apply to the case of bleed-through. Tan et al. [6] propose a method whereby the two sides of the document are combined into a single composite image for further processing. A manual registration of the two sides is used, and then a wavelet method is applied to enhance foreground strokes while smearing interfering strokes (bleed-through). A modified Canny edge detector is then used to identify and segment out foreground strokes.

In our previous work (Dubois and Pathak [7]), we proposed a method that does not highly depend on a model of



(a)

(b)

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

the bleed-through process. In this approach, the recto and verso images are first registered, areas assumed to correspond to bleed-through are detected, and then replaced by an estimate of the background. We follow this approach in this paper with some improvements that will be described. Then, since the bleed-through areas are detected, their locations can be transmitted in the form of a bi-level image, along with the values to be used to estimate the background, to generate the restored image from the original image at the receiver.

Bleed-through removal

Problem formulation

This paper follows the formulation of [7] which is summarized here for completeness, with some minor modifications to the notation. Let $f_r^c(x, y)$ and $f_v^c(x, y)$ be the original continuous-space recto and verso images, defined in the same coordinate frame. The origin is located in the upper left corner of the recto image, and thus it is located in the upper right corner of the verso image, with the x-axis going from right to left on the verso as illustrated in Fig. 1. We assume that there exist *ideal* recto and verso images representing the writing applied to the front and back of the paper, denoted $f_{wr}^c(x, y)$ and $f_{wv}^c(x, y)$ respectively, assumed to be zero where there is no writing; we refer to these as foreground. Similarly, we assume that there is an ideal background for each of the recto and the verso, denoted $f_{br}^c(x, y)$ and $f_{bv}^c(x, y)$, corresponding to the image of the paper without any writing. The resulting recto image with bleed-through is formed by combining the recto background, the recto foreground and the verso foreground (which causes the bleed-through) according to some function $\mathcal{C}$, and similarly for the verso image:

$$f_r^c(x, y) = \mathcal{C}(f_{br}^c(x, y), f_{wr}^c(x, y), f_{wv}^c(x, y)), \quad (1)$$
$$f_v^c(x, y) = \mathcal{C}(f_{bv}^c(x, y), f_{wv}^c(x, y), f_{wr}^c(x, y)). \quad (2)$$

Some possible models for the combining function $\mathcal{C}$ are presented in [7]. The goal of our restoration is to estimate the recto and verso images without bleed-through, i.e.,

$$\hat{f}_r^c(x, y) = \mathcal{C}(f_{br}^c(x, y), f_{wr}^c(x, y), 0), \quad (3)$$
$$\hat{f}_v^c(x, y) = \mathcal{C}(f_{bv}^c(x, y), f_{wv}^c(x, y), 0). \quad (4)$$

The recto and verso sides of the document are scanned to obtain sampled versions. In general, the document must be turned over to scan the second side, so the sampled images will not line up properly. We model the geometric transformation between the two sides by an affine transformation $\mathcal{A}_{\boldsymbol{p}}$ with parameter vector $\boldsymbol{p}$; this transformation can model a shift, rotation and even some skew. We thus obtain the two sampled images

$$f_r[m, n] = f_r^c(mX, nX), \quad (5)$$
$$f_v^\dagger[m, n] = (\mathcal{A}_p f_v^c)(mX, nX), \quad (6)$$
$$0 \le m \le M-1, 0 \le n \le N-1,$$

where $\boldsymbol{p} = (p_{11}, p_{12}, p_{13}, p_{21}, p_{22}, p_{23})$, X is the sample spacing, and

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

The superscript † is used to denote that the sampled verso image is not aligned with the recto image. Note that for typical conditions, $\mathcal{A}_p$ is invertible, and the inverse is also an affine transformation $\mathcal{A}_t$ [7].

Algorithm

The proposed algorithm consists of three main steps: (i) registration: alignment of the sampled recto and verso images; (ii) segmentation: detection of the regions designated as bleed-through; (iii) in-painting: replacement of the regions designated as bleed-through by an estimate of the background. A brief description of our implementation of these steps follows.

Registration

As shown in [7], under a simple additive combining model, the sampled verso can be aligned with the recto using the affine transformation with parameter $\hat{\boldsymbol{t}}$ determined by the optimization

$$\hat{\boldsymbol{t}} = \arg\min_{\boldsymbol{t}} \sum_m \sum_n (f_r[m,n] - (\mathcal{A}_t f_v^\dagger)[m,n])^2. \quad (8)$$

The sum may run over a subset of the pixels in the sampled images, including leaving a border. Note that $\mathcal{A}_t f_v^\dagger$ involves values of $f_v^\dagger$ that are not on the original sampling structure; we interpolated these using bicubic interpolation. The optimization was generally done in stages. First the shift parameters (t_{13}, t_{23}) were estimated using a direct search optimization procedure. Then all six parameters were refined using the optimization function `fminunc` under MATLAB. See [8] for more details of the implementation. Then, the registered verso image is denoted

$$f_v[m,n] = (\mathcal{A}_{\hat{t}} f_v^\dagger)[m,n]. \quad (9)$$

Segmentation

Based on the model of Eq. 1, 2, we can identify four different types of regions on each of the recto and verso sides of the document. These four regions, illustrated in Fig. 2, are defined as follows for the recto side, with corresponding definitions for the verso side:

1. R1: Foreground only, $f_{wr} > 0$, $f_{wv} = 0$.
2. R2: Bleed-through only, $f_{wr} = 0$, $f_{wv} > 0$.
3. R3: Background, $f_{wr} = 0$, $f_{wv} = 0$.
4. R4: Foreground and bleed-through overlap, $f_{wr} > 0$, $f_{wv} > 0$.

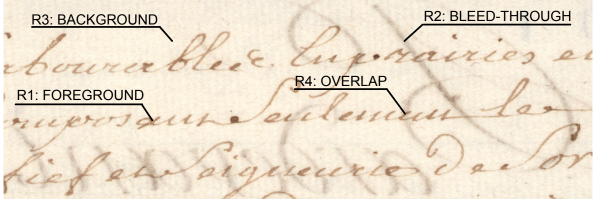


Figure 2: Illustration of the four types of regions in a document.

Our goal is to segment each side of the document into these four regions, although the most important one is region R2; the pixels in this region will be replaced by an estimate of the background. It is important that pixels in regions R1 and R4 not be misclassified as region R2, as this will cause portions of the recto writing to be erased. It is not serious if small portions of region R3, the background, are misclassified as region R2. In the method of [7], a pixel is classified as belonging to region R2 if f_v exceeds a certain threshold, and if f_v is sufficiently darker than f_r according to a second threshold. This simple scheme was found to introduce too many misclassification errors, so a more elaborate algorithm was devised, as follows.

A few considerations went into the selection of the new segmentation scheme. One observation is that the bleed-through may be somewhat diffused as it has seeped through the paper and thus occupy a larger area on the reverse side, so that a pixel-by-pixel comparison between recto and verso may not work well. Thus, local windows have been introduced into the algorithm. Roughly speaking, bright areas of the document can be considered to be background. Areas that are dark on the side of interest and not on the reverse side can be considered to be foreground. Areas that are dark on both sides, but more so on the reverse side can be considered to be bleed-through, while if they are dark on both sides but similar, they may be regions of overlap.

Thus, in a first step, pixels with values larger than a first threshold were considered to belong to the background. In our tests, this threshold T_b was chosen to be 90% of the location of the peak value of the image histogram. It is not serious if some of the background points are missed. The second step is to identify points considered to be foreground. In this case, a local window was used, and the darkest point in the window was considered to be representative of the point at the center of the window. Let

$$W_P = \left\{(k,l) | -\frac{P-1}{2} \leq k,l \leq \frac{P-1}{2}\right\} \quad (10)$$

where P is the window size, assumed to be odd. Then, the following filtered versions of the recto and verso images are defined:

$$r_{\min}[m,n] = \min_{(k,l)\in W_P} f_r[m-k, n-l], \quad (11)$$

$$v_{\min}[m,n] = \min_{(k,l)\in W_P} f_v[m-k, n-l]. \quad (12)$$

If the filtered recto is darker than the filtered verso, the pixel is declared to be foreground. In fact, we biassed this decision slightly. If $r_{\min}[m,n] \leq \alpha v_{\min}[m,n]$, then the pixel at location (m,n) is declared to be foreground. Tests on numerous documents led us to choose $P = 5$ and $\alpha = 1.2$. The remaining pixels were considered to be *possible* bleed-through. A final classification step assigned these pixels to either regions R2 or R4. It was assumed that in areas of bleed-through alone, the recto and verso images should be highly correlated. Thus, we define $c[m, n]$ to be the cross-correlation of f_r and f_v over a $Q \times Q$ square window centered at (m, n). The final decision is to assign the pixel at (m, n) to region R2 if $c[m, n] \geq T_c$ and to region R4 otherwise. Again, through experimentation, values $T_c = 0.5$ and $Q = 15$ were found to give good results. Fig. 3 shows the segmentation of a portion of the image of Fig. 1(b) using the algorithm and parameter values described above. Once again, the key property required is that the region detected as R2 not encroach into the regions R1 and R4 containing foreground; it may extend into the background region R1 with little serious consequence.

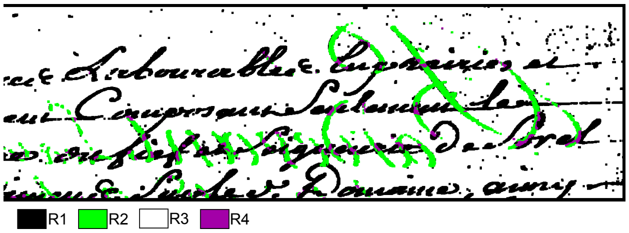


Figure 3: *Segmentation of a portion of the verso image of Fig. 1(b).*

Inpainting

Once the region R2 is identified, it should be replaced by an estimate of the background. Filling in the missing regions this way is known as inpainting [9]. Sophisticated techniques have been developed to accomplish this task. However, in this work so far, we have only applied relatively simple techniques, since the background we are trying to estimate is quite simple. We have filled in the regions corresponding to bleed-through with a fixed estimate of the background color, either globally or locally. The result of applying this simple scheme to the image of Fig. 1(b) is illustrated in Fig. 4.

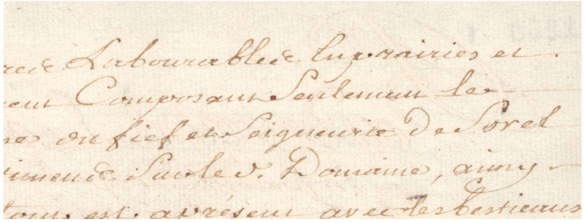
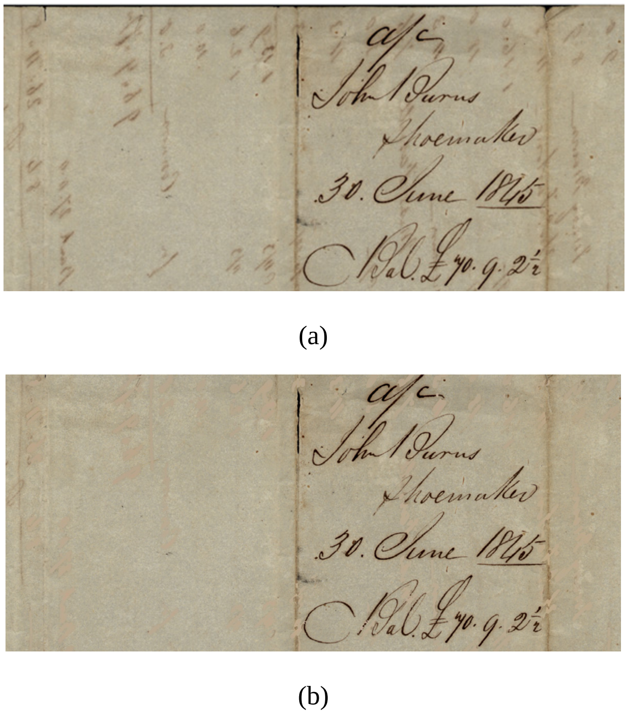


Figure 4: *Restored verso image of Fig. 1(b) using the simple inpainting scheme described above.*

Another example of an original document with bleed-through and the version corrected with this algorithm can be seen in Fig. 5.



(a)

(b)

Figure 5: *(a) Extract of the recto side of a document with significant bleed-through. (b) Corrected version.*

Compression scheme

The joint compression scheme we used is quite simple and is based on existing standards for continuous tone images and bi-level images. We first compress the original, uncorrected recto and verso images with a standard efficient compression scheme adapted to these kind of images, such as JPEG or JPEG 2000. We tried both of these and several others as well. For example, with a public domain implementation of JPEG 2000, we were able to compress the complete 2176 × 2662 image of Fig. 1 by a factor of 30:1 to give a file size of 4.6 Mbit and achieve a PSNR of 38.6 dB, resulting in a good quality reproduction. Then, rather than separately compressing and delivering the corrected image, we transmit the segmentation mask for region R2 as a bi-level image using a standard efficient compression scheme for such images. In our case, we used a public domain implementation of the JBIG standard, although the recently adopted JBIG2 would be more efficient. For example, the segmentation map determined for the verso image in Fig. 1(b) is shown in Fig. 6.

Then, this bilevel image is compressed with the JBIG lossless compression standard; for the image of Fig. 1, this data requires only 131 kbit compared to 4.6 Mbit to encode the entire image. All that needs to be transmitted in

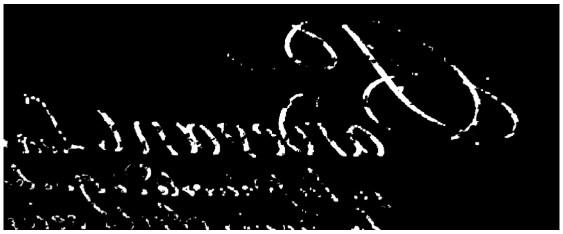


Figure 6: *Segmented region R2 for the verso image of Fig. 1(b).*

addition to that is the list of values to fill in the inpainted regions. In our previous example, this was only one value, which is negligible.

Conclusion

It is possible using the methods of this paper, and others as well, to significantly reduce the visibility of bleed-through and greatly improve the legibility of old documents suffering from this defect. For a researcher who wishes to remotely access both the original and the corrected version of the document, it is possible to send both with a data size only slightly larger than required to send only one version. This would of course require a special purpose display program on the user's computer. Such a program would allow the viewer to easily switch back and forth between the original and the corrected version of the document.

There still remains considerable work that can be done on this problem. The segmentation scheme presented here is quite empirical and could surely be improved using more sophisticated and rigorous statistical image processing techniques. Similarly, better inpainting algorithms can be used to further improve the quality of the restored documents.

Acknowledgement

This work was supported in part by a grant from the University of Ottawa Interfaculty Collaborative Research Initiatives program. The authors wish to thank René Paquet of Library and Archives Canada for providing the scanned documents used in this paper.

References

1. H. Nishida and T. Suzuki, A multiscale approach to restoring scanned color documents with show-through effects, in Proc. Seventh International Conference on Document Analysis and Recognition, vol. 1, pp. 584–588 (2003).
2. 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).
3. A. Tonazzini, E. Salerno, M. Mochi and L. Bedini, Bleed-through removal from degraded documents using a color decorrelation method, in S. Marinai and A. Dengel, editors, Proc. Document Analysis Systems VI: 6th International Workshop, Springer-Verlag GmbH, vol. 3163 of Lecture Notes in Computer Science, pp. 229–240 (2004).
4. K.T. Knox, Show-through correction for two-sided documents, United States Patent 5,832,137 (1998).
5. G. Sharma, Show-through cancellation in scans of duplex printed documents, IEEE Trans. Image Process., 10(5), 736 (2001).
6. C.L. Tan, R. Cao and P. Shen, Restoration of archival documents using a wavelet technique, IEEE Trans. Pattern Anal. Machine Intell., 24(19), 1399 (2002).
7. E. Dubois and A. Pathak, Reduction of bleed-through in scanned manuscript documents, in Proc. IS&T Image Processing, Image Quality, Image Capture Systems Conference (PICS), pp. 177–180 (2001).
8. P. Dano, Joint restoration and compression of document images with bleed-through distortion, Master's thesis, University of Ottawa, Ottawa, ON, Canada (2003).
9. M. Bertalmio, G. Sapiro, V. Caselles and C. Ballester, Image inpainting, in SIGGRAPH '00: Proceedings of the 27th annual conference on Computer graphics and interactive techniques, pp. 417–424 (2000).

Biography

Eric Dubois received the B.Eng. (hon.) degree and the M.Eng. degree from McGill University in 1972 and 1974, and the Ph.D. from the University of Toronto in 1978, all in electrical engineering. He was a professor in the INRS-Télécommunications center, University of Quebec, from 1977 to 1998. In 1998 he joined the School of Information Technology and Engineering at the University of Ottawa. His research has centered on the compression and processing of still and moving images, and in multidimensional digital signal processing.

Patrick Dano received the B.A.Sc. degree in Computer Engineering and the M.A.Sc. degree in Electrical Engineering from the University of Ottawa in 2000 and 2003. In between degrees, he spent time in industry working in various applications of signal processing (DSL communications). Currently, he is employed at the Canadian Imperial Bank of Commerce in the Technology sector.