2009 10th International Conference on Document Analysis and Recognition

Manuscript bleed-through removal via hysteresis thresholding

Rolando Estrada and Carlo Tomasi
Department of Computer Science, Duke University
Durham, NC, USA
restrada@cs.duke.edu, tomasi@cs.duke.edu

Abstract

Many types of degradation can render ancient manuscripts very hard to read. In bleed-through, the text from the reverse, or verso, side of a page seeps through into the front, or recto. In this paper, we propose hysteresis thresholding to greatly reduce bleed-through. Thresholding alone cannot properly separate ink and bleed-through because the ranges of intensities for the two classes overlap. Hysteresis thresholding overcomes this limitation via the two steps of thresholding and ink regrowth. In order to provide quantitative measures of the effectiveness of this approach, we constructed a novel dataset which features bleed-through and has available ground truth. We evaluated our method and a number of previously proposed approaches on ink pixel precision and recall. Hysteresis thresholding significantly improves over existing methods.

1. Introduction

Ancient manuscripts held in libraries and academic centers throughout the world provide an invaluable window into the culture of days past. However, many of these documents are subject to progressive decay, due to chemical breakdown, humidity and other factors, that may render them unreadable. One particularly pervasive problem is the phenomenon of *bleed-through* in which the ink from the reverse, or *verso*, side of the page seeps through into the front, or *recto*, side, producing interference with the main text that can range from faint to severe. Bleed-through has the primary effect of obscuring the original ink and thus makes reading, as well as automatic character recognition, more difficult. Since it becomes part of the manuscript itself, bleed-through will remain in any standard digitalization of the document. Examples of bleed-through are shown in Figures 1 and 4. In this paper, we propose hysteresis thresholding as an effective method of bleed-through removal in ancient manuscripts. By combining thresholding with ink regrowth through a spanning tree of connected pix-

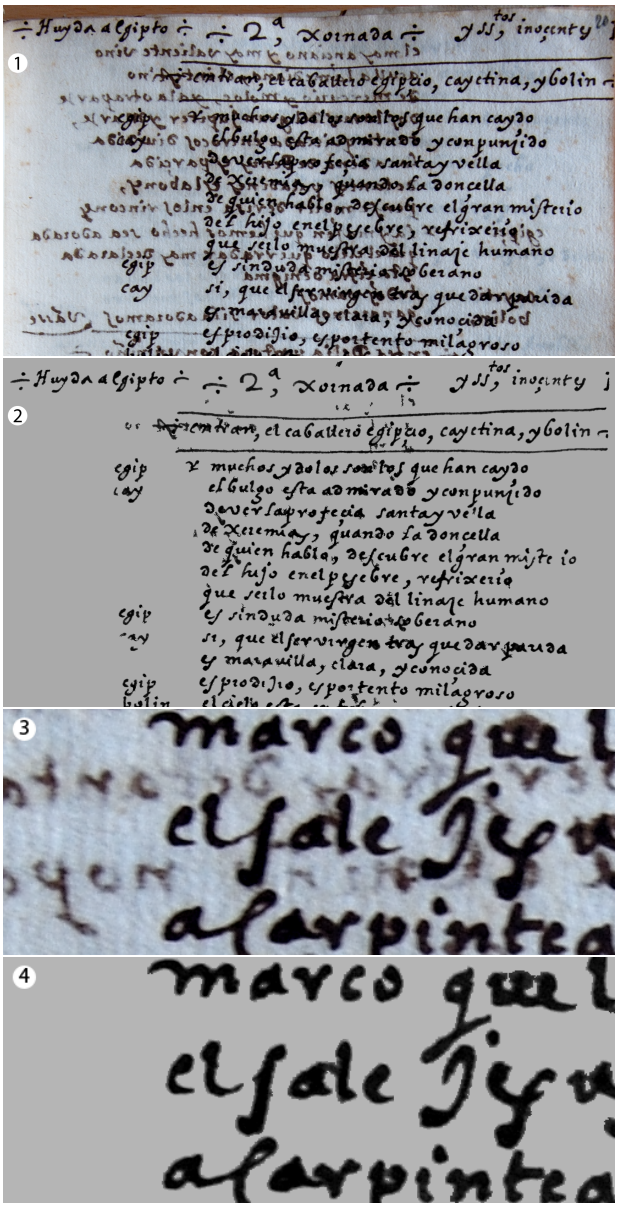

Figure 1. 1 & 3: Manuscripts with bleed-through. 2 & 4: After hysteresis thresholding

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.88

els, our method allows us to eliminate a considerable percentage of bleed-through while retaining most, if not all, of the original ink. Our method belongs to the blind family of bleed-through removal techniques, since it does not require *verso* information. While, in accordance with other works in the field, we present visual examples of the effects of our method on real ancient manuscripts, we also showcase a novel dataset production scheme that allows us to quantitatively test our method against previous approaches.

2. Previous efforts

Work directly tackling bleed-through itself is mostly confined to the present decade. The main dichotomy in the field is between blind (or single-sided) approaches and non-blind (or double-sided) ones. In the first type, only the *recto* page is used, while the second variety incorporates the *verso* into its processing. Clearly, since the data used by blind approaches is a subset of that employed by non-blind ones, the latter should output better results. However, although in principle *verso* information might be very useful, in practice there are two main complications that limit the applicability of non-blind methods. The first is that the *verso* has to actually be available, which is not always the case. The second is that both sides have to be registered correctly, which can be a non-trivial affair due to differing document skews, page warping, scaling issues, varying illumination and other factors. Therefore, we can't do away with blind methods, since they expand the number of documents in which bleed-through can be removed. Examples of non-blind approaches include wavelet analysis [12], symmetric orthogonalization [13], and intensity difference classification [4][5][15]. Also, there has been some non-blind work focused specifically on ancient music documents [3][1].

Concentrating on blind methods, some of the more prominent efforts include directional wavelets by Wang et al. [16], symmetric orthogonalization by Tonazzini et al. [14], and recursive unsupervised classification by Fadoua et al. [6]. The first is based on the difference in slant between ink and bleed-through strokes. Specifically, they note in their data that ink strokes tend towards $45°$ and bleed-through ones towards $135°$. This difference can be captured by the detail coefficients of a directional wavelet decomposition. Bleed-through is then reduced by iteratively enhancing detail in one orientation and suppressing it in the other. Symmetric orthogonalization is related to independent component analysis (ICA) and views the problem as as blind source separation (BSS), where the *recto* is a mixture of the original ink and paper with the bleed-through interference. Using the RGB components as three different views of the page, and assuming certain statistical properties for the data, the sources and mixing matrix are inferred. Recursive classification, on the other hand, is based on iteratively rotating the RGB space via principal component analysis (PCA) to decorrelate and reduce the data. Then, k-means is applied to separate the data into two classes; the darker set is classified as true ink and the lighter set is discarded. The procedure can then be iteratively applied.

Finally, we can note that bleed-through removal is closely related to document binarization. Binarization seeks to classify all pixels in the image as either foreground (black) or background (white). These classes ideally correspond to ink and paper for manuscripts, but the presence of bleed-through means that a naïve binarization will likely misclassify a lot of this class as foreground. Generally then, bleed-through removal can function as a crucial step in a comprehensive manuscript binarization scheme. Works that focus on document binarization, such as [10][9][11] or [7], have mostly dealt with other document degradations, such as stains and variable illumination, but their methods can eliminate some light bleed-through as well.

3. Problem formulation

As we saw in the introduction, bleed-through is caused by the seepage of ink from the *verso*, through the sheet of paper, into the *recto*. Explicitly or not, all bleed-through methods provide a classification of the pixels in the input image(s) such that, ideally, ink and bleed-through end up in separate classes, with the former being preserved and the latter scrubbed. Obviously, the *recto* and *verso* roles are reversible and the same method will apply to both sides. In general, we can view the (digitized) *recto* $R(x, y)$ and *verso* $V(x, y)$ pages as consisting of three types of pixels: [*ink*, *bleed-through*, *paper*]. We then want to produce a clean *recto* $C(x, y)$ where the bleed-through pixels have been reclassified as paper. For later comparison, we can also think of the original clean *recto* $T(x, y)$ as the target that we wish to approximate with C.

A grayscale pixel can be viewed as a vector (x, y, g) in $\mathbf{R}^3$, where x and y are the spatial components and g is the pixel's intensity. Note that in this paper, higher or greater intensity means darker. The distribution of pixels in this space determines the difficulty of the classification problem. We can identify three levels of difficulty: (1) Classes are linearly separable (globally or piecewise) based on intensity. (2) Classes are not fully intensity separable, but are spatially distinct; there is always at least one paper pixel between an ink and a bleed-through pixel. (3) Classes overlap both in intensity and location. Achieving separation for each class requires a different principle. Simple thresholding, whether global, t, or locally adaptive (piecewise), $t(x, y)$, can only succeed for type 1 manuscripts, which satisfy condition (1). To tackle type 2, we propose hysteresis thresholding. For type 3, we need to impose additional regrowth restrictions on this method.

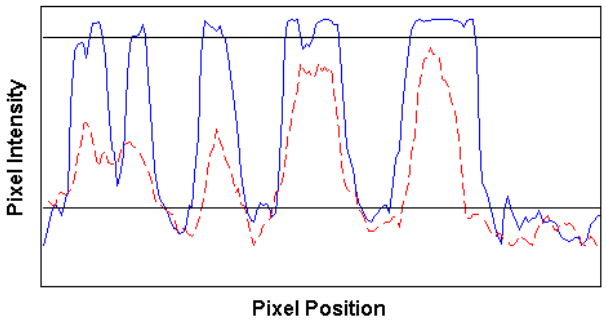

Figure 2. Scan line of ink (solid) and mirrored resulting bleed-through (dashed) intensities

Figure 3. Left to right: Original, High threshold (H),Low threshold (L), Reconstructed (C)

Basically no handwritten document will be of type 1 due to the mechanics of writing with pens or quills. In handwriting, the ink does not stop at a sharp cutoff, but gradually fades into the paper, which accounts for most of the intensity overlap between ink and bleed-through. In more detail, the vast majority of light ink pixels are located at the edges of characters, while the character interiors are usually significantly darker. In Figure 2, the solid line indicates a scan line along a real ink word and the dashed one the (mirrored) bleed-through that appeared on the other side as a result of it. Here, dark pixels are "high" and light pixels are "low". Clearly bleed-through is generally lighter than the ink that generated it and also fades into the paper. So, we can set a threshold above which only (part of) the ink remains (top horizontal line in Figure 2) and a lower threshold below which we can consider everything paper (lower line). So, while ink and bleed-through have similar minimum intensities, their maximum intensities are usually quite different.

More importantly, although many ink pixels are lighter than some of the darker bleed-through ones, these lighter ink pixels are nevertheless usually connected to much darker ink pixels that rise above almost all bleed-through intensities. Therefore, if we can identify the most prominent peaks, from them we can also obtain the lighter ink pixels connected to them, while avoiding bleed-through pixels of comparable intensities. This is essentially the principle behind the hysteresis thresholding used in Canny's edge detector [2] for small ridge suppression.

4. Hysteresis thresholding

If R is our grayscale input image, then given intensity thresholds t_h, t_l, we can construct two sets of pixels, $H = \{(x,y)|R(x,y) \geq t_h\}$ and $L = \{(x,y)|R(x,y) \geq t_l\}$, which can be used directly to produce binary masks such that all pixels above the threshold are "on" and all others are off. For a given image, t_h marks the level at which non-ink pixels are negligible, while t_l indicates the point below which no more ink pixels are expected. Both these values can be either global, t, or locally adaptive, $t(x,y)$. Empirically, the use of Otsu's threshold [10] for t_h and the median pixel intensity for t_l tends to produce good results. With these two sets, we now wish to obtain C, such that $H \subset C \subset L$. In short, H is too restrictive and removes too many ink pixels, while L is too permissive and keeps too many bleed-through ones. This implies that our target set C consists of $C = H \cup L'$, where $L' \subset L$. Now, all we need is a procedure or criterion for choosing L' out of L.

As we can see in Figures 1 and 3, the spatial distribution of pixel intensities is not uniformly random, but is heavily clustered with lighter and lighter pixels surrounding dark peaks. Therefore, viewing the pixels as nodes in a graph, we can establish connectivity as our main criterion for choosing L'. In other words, we can begin from every pixel in H and create a spanning tree from all its neighboring pixels that are part of L (and satisfy some additional criteria discussed below). These connected pixels will then form part of L'. Clearly, pixels in L that are not connected, either directly or through a chain of connected L pixels, to any pixel in H will not be included, so pixels of the same intensity in L might enjoy different fates. L', then, is simply the set of all pixels in L that are part of some spanning tree emanating from a pixel in H. Figure 3 illustrates the four sets for two particular characters, while Figure 1 presents the effect of our method on a larger section of a manuscript.

As an additional preprocessing criterion, we can establish a minimum cluster size for elements in H, below the size of the smallest expected characters. This will remove small, usually misclassified specks.

If the only tree growth stopping criterion is t_l, we can already effectively deal with type 2 manuscripts; however, type 3 will be troublesome since in them bleed-through and ink can be quite intermixed. In this case, we can better avoid regrowing bleed-through by establishing three constraints on the candidate pixel to be added to the spanning tree. First, the gradient between the incoming pixel and the connecting pixel already in the tree cannot exceed an additional threshold t_g since very sharp drops generally indicate meaningful boundaries. Second, the gradient cannot

change sign. This situation indicates that the incoming pixel is darker, so it is more likely part of a different character. Finally, we establish a maximum spanning tree branch length, since characters don't extend indefinitely. These three parameters are set, currently by trial-and-error, based on the intensity slopes at the edges of (a sample set of) characters.

Before delving into the experiments, it can be instructive to analyze why, despite the effectiveness that we will show for this method, hysteresis thresholding alone cannot, in general, completely remove all instances of bleed-through. Its optimality is bounded by the correctness of its thresholds, so poor thresholds, global or adaptive, will produce poor results. Thus, very dark bleed-through over t_h may be misclassified as ink and survive, while faint ink regions where all the pixels are rather bright may all fall below t_h and be removed as bleed-through. However, these limitations are not restricted to our method, by any means, and, as the experiments will show, they pose an essentially fundamental roadblock for bleed-through removal methods, particularly blind ones.

5. Experimental results

In the field's short lifespan, determining the efficacy of new bleed-through methods has been generally more qualitative than quantitative. Presumably, the main issue has been the lack of ground truth for ancient manuscripts since the clean recto, T, has been lost to time. When efforts have been made to obtain some form of ground truth [12][16], they have involved laborious manual, and hence subjective, segmentation of the test images. Therefore, previous efforts have mostly focused on visually presenting how their method improves on the original ancient manuscripts, as we have done in the previous section. However, the lack of common datasets has precluded the possibility of comparing different methods in an objective manner. We thus propose a simple, but effective procedure for producing useful datasets, with their associated ground truth. While the datasets used in this paper are primarily focused on featuring bleed-through effects, other document degradations, such as stains, can be incorporated into our basic methodology. The procedure in question consists of: (1) Write on one side of a porous sheet with heavy ink that bleeds through instantly; (2) Photograph both sides, making sure to minimize misalignment; (3) Write on the reverse side; (4) Photograph both sides again. We now have four different images, with images 1, 3 and 4 corresponding to T, R and V. In short, we can use image 3 as input into the different methods, while keeping image 1 as our coveted ground truth. Given that our method is blind, in this paper we only compare it to other blind methods, although our datasets include the *verso* as well, so non-blind methods could be tested on them. Although this procedure involves manual production, since it requires human writers, automated methods cannot produce realistic enough results.

Table 1. Bleed-through removal

Method	Recall	Precision	F-measure
Hyst_best	**92.6($\pm$2.5)**	**86($\pm$7)**	**89.2**
Hyst_otsu	**95($\pm$1.5)**	**78.9($\pm$7.5)**	**86.2**
Wang	92.7($\pm$1.9)	74.2($\pm$6.6)	82.4
Otsu	89.1($\pm$1.2)	75.9($\pm$6.1)	82
ICA	86.8($\pm$3.5)	75.4($\pm$6.7)	80.7
Tonazzini	80.5($\pm$8.5)	66.6($\pm$8.7)	72.9
Fadoua	60.8($\pm$6.6)	85.5($\pm$7.1)	71.1
Sauvola	100($\pm$0)	28.7($\pm$11.5)	44.6

For our experiments, we used two datasets: (1) the first chapter of Miguel de Cervantes' *Don Quixote* (ten pages) handwritten by Estrada and (2) a two poem set (16 pages), with William Wordsworth's *Daffodils* on the *recto* and John Donne's *Death Be Not Proud* on the *verso*. The poems were written by 8 different Duke University students and feature an array of calligraphic styles. All images were taken on a Nikon D200 camera with a 28mm lens. The original 2592×3872 RGB images were cropped and resized to 1213×1563 to allow some of the more memory intensive methods to process the images. To compensate for minor physical misalignments, we registered the [T,R] pairs via rigid translation followed by cross-correlation. Figure 4 shows examples of both datasets.

The algorithms being tested are listed in Table 2. We have included the three blind bleed-through removal methods mentioned earlier plus a few additional approaches: Otsu and Sauvola [11] are examples of document binarization methods, with the former being global and the latter adaptive, while ICA [14] represents a general statistical technique. For our method, Hyst_otsu uses Otsu's threshold as t_h, while Hyst_best uses a slightly more corrosive threshold that achieves better bleed-through removal. Except for the Tonazzini RGB [14], Otsu and ICA methods, all the algorithms were implemented by the authors.

The metrics employed for these experiments are the familiar precision, recall and F-measure, as defined in [7], measured for ink pixels. Recall thus measures how much ink is preserved and $1-$precision indicates how much bleed-through still remains. Presented as percentages, these values are listed in Table 1. For each output image, pixels were classified as ink or paper based on intensity. For each method, the values in Table 1 represent the ink/paper separation that maximized their F-measure.

As Table 1 shows, obtaining complete bleed-through removal while retaining all the original ink is not an easy task. Essentially methods tend to be either too corrosive or too permissive: in the first case, their recall suffers; in the second case, precision becomes an issue. Due to significant

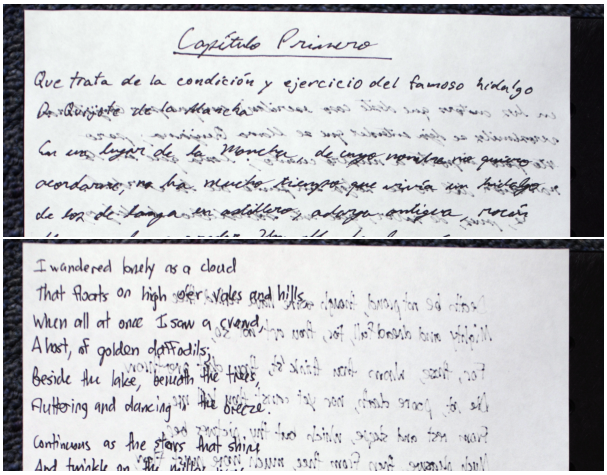

Figure 4. Dataset examples

bleed-through removal while still respecting the original ink, our method clearly outperforms previous algorithms. While small datasets should be interpreted with care [8], our results so far are very promising.

6. Conclusions & future work

We have presented a novel method for dealing with bleed-through in ancient manuscripts through the two steps of thresholding and ink regrowth. By employing a hysteresis approach, we are able to more effectively classify ambiguous pixels as either ink or bleed-through than conventional thresholding methods. Also, thanks to our constructed datasets, we were able to show that our method can achieve significantly better quantitative values than previous efforts. For subsequent work, using our proposed dataset construction method, we hope to expand our testing to much larger and more diverse (color figures, stains, water damage, etc.) datasets, as well as test methods that incorporate *verso* data.

As we noted in our method description, the difficulty in identifying very dark bleed-through regions and very light characters is a constant theme for all the methods tested. Currently, two avenues of attack seem most promising: First, by incorporating *verso* data, we should be able to craft more accurate H and L' sets. More interestingly, since humans are often able to differentiate through character recognition alone, we can conceive of adding character analysis steps to better disambiguate borderline cases. Our future research will explore both of these possibilities.

7. Acknowledgments

The authors thank Prof. Margaret R. Greer, at the Duke Romance Studies Department, for providing the ancient manuscripts, as well as Dr. Tonazzini for her source code.

References

[1] J. A. Burgoyne, J. Devaney, L. Pugin, and I. Fujinaga. Enhanced bleedthrough correction for early music documents with recto-verso registration. In *Int. Conf. on Music Information Retrieval*, pages 407–412, 2008.

[2] J. Canny. A computational approach to edge detection. *IEEE Transactions on PAMI*, 8(6):679–698, November 1986.

[3] P. Castro, R. J. Alameda, and J. R. C. Pinto. Restoration of double-sided ancient music documents with bleed-through. In *Iberoamerican Congress on Patt. Rec.*, pages 940–949, Montreal,Canada, 2007.

[4] E. Dubois and P. Dano. Joint compression and restoration of documents with bleed-through. In *IS&T Archiving*, pages 170–174, 2005.

[5] E. Dubois and A. Pathak. Reduction of bleed-through in scanned manuscript documents. In *IS&T Image Processing, Image Quality, Image Capture Systems Conf.*, pages 177–180, Montreal,Canada, 2001.

[6] D. Fadoua, F. Bourgeois, and H. Emptoz. Restoring ink bleed-through degraded document images using a recursive unsupervised classification technique. In *DAS*, pages 38–49, 2006.

[7] B. Gatos, I. Pratikakis, and S. J. Perantonis. Efficient binarization of historical and degraded document images. In *DAS*, pages 447–454, 2008.

[8] M. Junker, R. Hoch, and A. Dengel. On the evaluation of document analysis components by recall, precision and accuracy. In *ICDAR*, pages 713–719, September 1999.

[9] W. Niblack. *An Introduction to Digital Image Processing*. Prentice-Hall, Englewood Cliffs, NJ, 1986.

[10] N. Otsu. A thresholding selection method from gray-level histogram. *IEEE Trans. on Systems, Man and Cybernetics*, 9:62–66, March 1979.

[11] J. Sauvola and M. Pietikainen. Adaptive document image binarization. *Pattern Recognition*, 33(2):225–236, 2000.

[12] C. Tan, R. Cao, and P. Shen. Restoration of archival documents using a wavelet technique. *IEEE Trans. on PAMI*, 24(10):1399–1404, October 2002.

[13] A. Tonazzini, E. Salerno, and L. Bedini. Fast correction of bleed-through distortion in grayscale documents by a blind source separation technique. *IJDAR*, 10(1):17–25, 2007.

[14] A. Tonazzini, E. Salerno, M. Mochi, and L. Bedini. Bleed-through removal from degraded documents using a color decorrelation method. In *DAS*, pages 229–240, 2004.

[15] J. Wang, M. S. Brown, and C. L. Tan. Accurate alignment of double-sided manuscripts for bleed-through removal. In *DAS*, pages 69–75, 2008.

[16] Q. Wang, T. Xia, L. Li, and C. L. Tan. Document image enhancement using directional wavelet. In *Int. Conf. on Comp. Vision and Patt. Rec.*, pages 16–22, Wisconsin, June 2003.