A Framework for Reducing Ink-Bleed in Old Documents

Yi Huang[†] Michael S. Brown[‡] Dong Xu[†]

[†]Nanyang Technological University; School of Computer Engineering; Republic of Singapore
[‡]National University of Singapore; School of Computing; Republic of Singapore *

Abstract

We describe a novel application framework to reduce the effects of ink-bleed in old documents. This task is treated as a classification problem where training-data is used to compute per-pixel likelihoods for use in a dual-layer Markov Random Field (MRF) that simultaneously labels image pixels of the front and back of a document as either foreground, background, or ink-bleed, while maintaining the integrity of foreground strokes. Our approach obtains better results than previous work without the need for assumptions about ink-bleed intensities or extensive parameter tuning. Our overall framework is detailed, including front and back image alignment, training-data collection, and the MRF formulation with associated likelihoods and intra- and inter-layer cost computations.

1. Introduction

Ink-bleed is a serious problem commonly found in aging handwritten documents. Ink-bleed occurs when ink written on one side of a page penetrates the paper to become visible on the opposite side. The severity and characteristics of ink-bleed is related to a variety of factors including the ink's chemical makeup, the paper's physical and chemical construction, the amount of ink applied and the paper's thickness (both spatially varying), the document's age, and the amount of humidity in the environment housing the documents. Figure 1-(a) shows examples of ink-bleed exhibiting various levels of severity and intensity characteristics from four different documents. The documents are from the same archival collection dating from 1820-1850.

The obvious drawback of ink-bleed is the reduction in the document's legibility. The motivation of our work is to provide a practical approach to reduce ink-bleed interference in imaged documents in order to improve legibility as shown in Figure 1-(b). We also strive for a solution that is applicable in a real-world setting that considers the wide-range of ink-bleed diversity as well as practical concerns, such as the abilities of the end users.

*Email: brown@comp.nus.edu.sg; {hu0005yi|dongxu}@ntu.edu.sg

(a) (b)

Figure 1. (a) Closeup of image regions from different handwritten documents, circa 1820-1850, suffering ink-bleed. (b) Our goal is to retain the original foreground strokes to improve legibility.

[Contribution] This paper describes a novel application framework to reduce ink-bleed from images of the front and back of a document. The problem is treated as one of classification where image pixels are labeled as either foreground, ink-bleed, or background. This pixel labeling is aided by a dual-layer MRF with smoothness cost designed to reduce noise while maintaining foreground strokes in regions where foreground and ink-bleed overlap. Our approach operates on a wide-range of ink-bleed and does not require assumptions about the ink-bleed intensity or extensive parameter tuning. All necessary components needed for this application are presented, including front and back image alignment, collection of training-data, and the dual-layer MRF setup with associated likelihoods and intra- and inter-layer cost computations.

The remainder of this paper is organized as follows: section 2 discusses related work; section 3 provides an overview of our framework; section 4 details the data-cost computations and dual-layered MRF formulation; section 5 presents results including comparisons against other approaches; section 6 provides a short discussion and summary.

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

2. Related Work

There is surprisingly little previous work targeting complex ink-bleed. This is likely due to the difficulty in obtaining access to older handwritten materials that are housed under tight regulation. Much of the previous work that does exist focuses on relatively simple ink-bleed that can be removed using variations of local or global thresholding (e.g. [1, 6, 10]).

Drira et al [5] presented a recent approach targeting complex ink-bleed that uses Principal Component Analysis (PCA) to first reduce the dimensionality of an RGB input image. Pixels then are classified as foreground and background by iteratively clustering the PCA data into two groups via an adaptive threshold. Tonazzini et al [14] targeted complex ink-bleed using blind signal separation via Independent Component Analysis which linearly decomposes an RGB image into three signals assumed to be foreground, background, and ink-bleed. Wolf [16] recently extended this idea using non-linear blind separation via an MRF framework that could separate the foreground and ink-bleed from either RGB or grayscale images.

Thresholding and source separation approaches produce good results when the ink-bleed and foreground have clearly distinguishable graylevel intensities or RGB signatures. Both techniques, however, suffer when the ink-bleed and foreground have similar intensities as shown in some of the examples in figure 1-(a). Thresholding techniques make a further assumption that the ink-bleed intensity is *always* lighter than the foreground, an often invalid assumption. In addition, these techniques use a single image only which provides limited information.

One obvious strategy to obtain more information is to use images from both the front and back side of a document. Sharma [9] demonstrated a successful two-image 'show through' reduction approach for use in xerox imaging. Show-through, however, assumes global bleeding between the front and back images where ink-bleed typically varies spatially making it more difficult to model. The most significant two-image approach targeting ink-bleed are the wavelet-based approaches introduced by Tan et al [12] and Wang et al [15]. These techniques first globally align the front and back images from which an initial classification of the foreground and ink-bleed strokes is made using the magnitude of the image difference. Iterative filtering of the wavelet coefficients is used to dampen ink-bleed while sharpening foreground pixels. While this technique produces good results, six parameters must be tuned *per example*, including thresholds for the difference-image, dampening and sharpening coefficients, the number of wavelet scale-levels, and the number of iterations.

Our approach is unique from previous work in several distinct ways. First, while previous work performs well for examples that meet their assumptions, our approach operates on a larger range of inputs without explicit thresholding or extensive parameter tuning. Our approach also simultaneously corrects both the front and back images, where other two-images approaches ([15, 12, 9]) process each side individually. Lastly, we present a complete framework, including overlooked components such as the need for local-alignment of the front and back images, as well as an easy way to collect training-data.

3. Framework Overview

Our overall framework is discussed, including front and back image alignment, the feature used for classification, and training-data collection. Details to the labeling MRF are given in section 4. A brief description of our application's usage is presented first.

3.1. Application Usage

This work is done in partnership with the National Archives of Singapore which houses hundreds of volumes of governmental ledgers, circa 1820-1850, that suffer from ink-bleed. Many of these ledgers have been imaged to grayscale microfilm while others are imaged upon request. Researchers of these ledgers are most often lawyers and legal aides who still rely on these documents in legal disputes. Only digital images of the original materials are made available to users. Our application serves as a post-processing tool to help make the documents more legible. While most users are computer-literate they have little to no background in computer vision or image processing.

3.2. Step-by-step Procedure

Our framework starts with two high-resolution images ($\sim$2K$\times$3K) of the two sides of a page. Images obtained from microfilm are grayscale, while others are RGB scans of the original material. The pages in these volumes are typically bound and as a result are not completely pressed flat when imaged. This non-planar imaging compounded with small 3D surface variations that are typical of these older documents makes it impossible to align the front and back images with a single global transform. While 'flattening' techniques can remove these 3D surface variations (e.g. [4, 8, 13]), such approaches require additional 3D scanning equipment not available in mainstream imaging setups and cannot be used on existing microfilmed documents. As a result, a local alignment procedure in addition to global alignment is needed.

Considering our input images our overall procedure is as follows: 1) image alignment with local refinement; 2) training-data collection via minimal user-assistance; 3) pixels labeling using the dual-layer MRF framework; 4) output image generation.

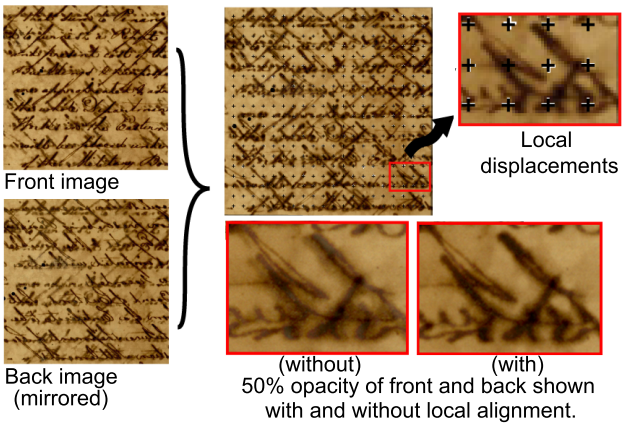

Figure 2. Local displacements computed between the front and back images. A zoomed inset shows front regions (black-crosses) matched to their corresponding locations in the back image (white-crosses). Overlapped image regions *with* and *without* local alignment are shown with 50% opacity. 'Ghosting', visible from misalignment, is removed with the local alignment procedure.

3.3. Image Alignment with Local Refinement

While input images are already coarsely aligned they still require an initial global alignment. To do this, the back image is mirrored and the location of the maximum correlation score of the front and mirrored back image over a $[-20, 20]$ pixel range in the horizontal and vertical direction is then taken be the global displacement. For our examples, a translation alignment suffices, however, a full affine alignment could be easily incorporated.

Local displacements are computed by dividing the front image into local windows (60×60 for our examples). Correlation is performed between each front image window with its corresponding back location over a $[-10, 10]$ pixel range in the horizontal and vertical direction. The location of the maximum correlation score for each window is taken to be the local displacement. If the maximum score is below a minimum threshold, it is assumed that there is no local change. Using the local displacements, thin-plate-spline (TPS) interpolation [2] is used to warp the back page to align with the front. Figure 2 shows this local-alignment procedure. The front and back image regions are shown overlapped with 50% opacity. Ghosting from misalignment is visible when local alignment is not performed; this is removed after TPS warping.

3.4. Ratio Feature

A good feature is crucial for classification. Given the aligned images, a *ratio feature* is defined as $\rho_p = \frac{C_p}{C_{p'}}$, where C_p and $C_{p'}$ are the intensities of front image pixel p and corresponding back image pixel p' respectively. The back image feature, $\rho_{p'}$, is the reciprocal of the front feature.

This ratio feature saliently captures the difference in the various front-back configurations, including situations where the ink-bleed intensity is darker or lighter than the opposite page's foreground pixels. Difficulties can arise when the ink-bleed and foreground have similar intensities, resulting in a ratio close to the value 1, a value that can also occur in features where both front and back pixels are background or foreground. Inter-layer costs in our dual-layer MRF will be used to disambiguate this situation.

While this is a simple feature, it proved to be the best discriminant over other all available information including feature from pixel intensity (including RGB when available), pixel differences, and combinations of these that also included ratio.

3.5. Training Data Collection

Due to the ink-bleed diversity, training-data needs to be obtained for each image pair. This requires user assistance which is kept to a minimum using a two step procedure. The user first draws simple color-coded strokes on the front and back images, labeling a few examples of foreground, background, and ink-bleed.

The initial user labeled training samples are too sparse for practical use and unbalanced in terms of number of labels, with generally many more background examples provided than foreground and ink-bleed. To enlarge the training-sets, a K-NN classifier based on the ratio feature's distance from the sparse user-labeled data is used to label the entire image, where K is the square root of the size of the user-labeled data. Pixel-wise confidence scores are computed as discussed in Eq (2) in Section 4.1. The 10% most confident pixels for each class are selected to define the enlarged training-sets. Figure 3 shows this training-data collection procedure.

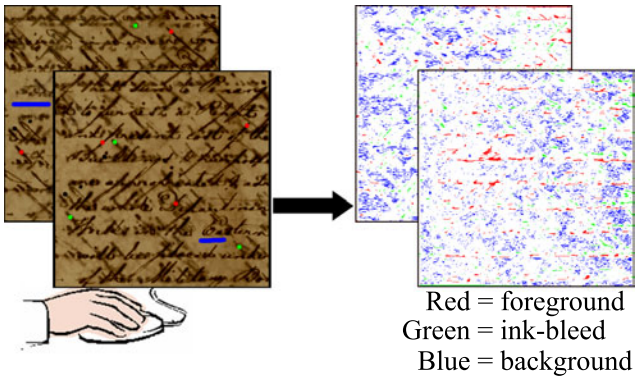

Figure 3. Initial training-data is provided via minimal user markup in the form of color strokes or points drawn over foreground (red), ink-bleed (green) and background (blue) examples in the front and back images (markup is enlarged for clarity). The training-data is enlarged using highly confident pixels labeled via the initial user markup.

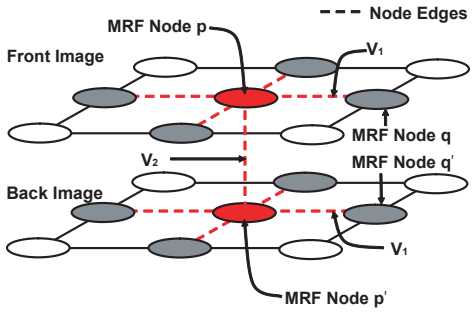

Figure 4. Our MRF network with associated nodes and edges.

4. MRF formulation

After training-data collection, each image pixel is labeled as one of three classes: *Foreground*, *Ink-bleed* and *Background* denoted as $\{\mathcal{F}, \mathcal{I}, \mathcal{B}\}$. This task is formulated as a discrete labeling MRF where each pixel, p is assigned a label l_p, where $l_p \in \{\mathcal{F}, \mathcal{I}, \mathcal{B}\}$ (see [7] for details to MRF formulations). The optimal label assignment is found by minimizing the following energy terms:

$$E = E_d + \lambda E_s, \tag{1}$$

where E_d represents the data-cost energy associated with the likelihoods of assigning a l_p to each pixel and E_s is a smoothness energy based on the MRF's prior cost for assigning neighboring pixels different label values. The scalar weight λ is set to one in our work. While this energy function is standard for all MRFs, the associated likelihood (data cost) and the prior (smoothing cost) are unique for each problem. Details of E_d are given in section 4.1.

Our smoothness term E_s is composed of intra-layer edge costs, $V_1(l_p, l_q)$, that computes the cost of assigning neighboring pixels the labels l_p and l_q and inter-layer edge costs (between layer), $V_2(l_p, l_{p'})$, that computes the cost of assigning a label combination to pixel p and its corresponding pixel on the opposite layer p'. Intra-layer edges are set for both the front and back image, thus we also have edges $V_1(l_{p'}, l_{q'})$ as shown in figure 4. Intra-layer edge costs are designed to encourage consistent labels based on feature and color similarity. The inter-layer edge costs are designed to avoid invalid label configurations and aid in resolving regions with overlapping ink. This dual-layer combination proves significantly more effective at maintaining foreground strokes compared with using the intra-layer alone. Details to E_s are given in section 4.2.

4.1. Data Cost Energy E_d

The data-cost E_d is defined for both the front and back image. Only the front is described here for example. The 1-dimensional ratio feature is normalized to be zero centered with standard deviation of one before the following procedure. For speedup, the dense training-data is clustered by K-means, with cluster centers of each class represented as $\{\rho_i^{\mathcal{F}}\}|_{i=1}^{L}$, $\{\rho_j^{\mathcal{I}}\}|_{j=1}^{M}$ and $\{\rho_k^{\mathcal{B}}\}|_{k=1}^{N}$. While choosing the optimal number of cluster centers is an open problem, we set $L = M = N$ as 10% of the size of the smallest training-set.

For each pixel p, we compute the Euclidean distances (L2-norm) between ρ_p and all the $L + M + N$ cluster centers and then select the top-K closest centers where K is set as $\sqrt{L + M + N}$. The top-K centers are denoted as $\{\rho_m\}|_{m=1}^{K}$ and are further divided into three index sets $\pi^{\mathcal{F}}$,$\pi^{\mathcal{I}}$ and $\pi^{\mathcal{B}}$ according to their labels. The distance between ρ_p and the m-th cluster center ρ_m is computed by $d_{pm} = ||\rho_p - \rho_m||$. We also denote d_p^2 as the mean squared distance to the top-K centers. The similarity of pixel p to each class is defined as:

$$\begin{aligned} S_{\mathcal{F}} &= \sum_{m \in \pi^{\mathcal{F}}} \exp(-d_{pm}^2/d_p^2) \\ S_{\mathcal{I}} &= \sum_{m \in \pi^{\mathcal{I}}} \exp(-d_{pm}^2/d_p^2) \\ S_{\mathcal{B}} &= \sum_{m \in \pi^{\mathcal{B}}} \exp(-d_{pm}^2/d_p^2). \end{aligned} \tag{2}$$

The data-cost term, E_d, for each label is defined as:

$$\begin{aligned} E_d(l_p = \mathcal{F}) &= \frac{S_{\mathcal{I}} + S_{\mathcal{B}}}{2 \times (S_{\mathcal{F}} + S_{\mathcal{I}} + S_{\mathcal{B}})} \\ E_d(l_p = \mathcal{I}) &= \frac{S_{\mathcal{F}} + S_{\mathcal{B}}}{2 \times (S_{\mathcal{F}} + S_{\mathcal{I}} + S_{\mathcal{B}})} \\ E_d(l_p = \mathcal{B}) &= \frac{S_{\mathcal{F}} + S_{\mathcal{I}}}{2 \times (S_{\mathcal{F}} + S_{\mathcal{I}} + S_{\mathcal{B}})}. \end{aligned} \tag{3}$$

Eq (3) results in E_d ranging between zero and one, and $E_d(l_p = \mathcal{F}) + E_d(l_p = \mathcal{I}) + E_d(l_p = \mathcal{B}) = 1$.

4.2. Smoothness term E_s

As previously stated, the prior term E_s is computed as edge costs within a layer and between layers giving rise to:

$$E_s = \sum_{p,q \in \mathcal{N}} V_1(l_p, l_q) + \sum_{p,p' \in \mathcal{M}} V_2(l_p, l_{p'}), \tag{4}$$

where $p, q \in \mathcal{N}$ are the within layer edges and $p, p' \in \mathcal{M}$ are the between layer edges. These two terms are weighted equally.

4.2.1 Intra-Layer Edge Costs

Intra-layer costs are based on the intensity difference or ratio difference between two intra-layer neighbors p and q. We define $d_{pq}^{\rho} = ||\rho_p - \rho_q||$ as the distance between p and q ratio feature. Similarly we define $d_{pq}^{c} = ||C_p - C_q||$ as the distance between p and q pixel intensity. We normalize d_{pq}^{ρ} and d_{pq}^{c} to range between zero and one. To impose a smoothness constraints in the intra-layer while preserving

the edges between different classes, the intra-layer cost is expressed as:

$$V_1(l_p, l_q) = \frac{1}{1 + (\xi_{pq})^2}, \quad (5)$$

where ξ_{pq} is defined in the following table:

l_p	l_q		
	Foreground	Ink-Bleed	Background
Foreground	∞	d^{ρ}_{pq}	d^{c}_{pq}
Ink-Bleed	d^{ρ}_{pq}	∞	d^{ρ}_{pq}
Background	d^{c}_{pq}	d^{ρ}_{pq}	∞

It is worthwhile to note that we use d^{c}_{pq} to define the intra-layer cost in *Foreground-Background* configuration because we observe that the intensity from the foreground and background pixels differ more than that of the ratio feature. In other configurations, we use the default ratio feature. Moreover, if the neighbors have the same label, we use zero cost to enforce the smoothness constraint (the three ∞ in the diagonal cells result in a zero cost).

4.2.2 Inter-Layer Edge Costs

Inter-layer costs, $V_2(l_p, l_{p'})$, are defined as:

l_p	$l_{p'}$		
	Foreground	Ink-Bleed	Background
Foreground	0	0	0
Ink-Bleed	0	∞	∞
Background	0	∞	2ω

In the above table, we have a conditional constraint for the *Background-Background* configuration. We set ω as 1 if ($C_p < C^1_{avg}$ and $C_{p'} < C^2_{avg}$), 0 otherwise, where C^1_{avg} and C^2_{avg} are the average intensities of the foreground pixels in the front and back images respectively. The background pixels are usually the brightest pixels in the whole document, thus we assume that both front and back pixel that have lower intensity (i.e. darker) are not likely to be a *Background-Background* configuration and a small weight of 2 is used as penalization. We use three *infinity* penalties in the above table because these three cases will never happen. If one pixel is labeled as ink-bleed in one side, the corresponding pixel in the opposite image can only be foreground. All other configurations are possible and are therefore assigned a zero cost.

4.3. Minimizing the Objective Function Energy

We use the Graph-cuts approach described by Boykov and Kolmogorov [3] to minimize our global objective function stated in Eq (1). The Middlebury's MRF code provided by [11] is modified to incorporate our dual-layer configuration. In all of our experiments, the optimization converges within 5-6 iterations.

5. Results

We compared our approach with the single-image adaptive thresholding approach [5][1], the front and back image wavelet-based approach [15], and a single-layer MRF based on our data-cost and intra-layer cost formulations. Markup varies per example, but generally consists of $5 - 15$ strokes or points drawn on both the front and back images. Processed images are roughly 2K×3K in resolution. Pixels labeled as foreground are shown with the input images intensity, all other pixels are set to the mean intensity of the background training-labels.

Figure 5 shows sub-regions from four examples that represent a reasonably diverse range of ink-bleed. Shown are the front and back input and our results, as well as a comparison of the other approaches on the front image only which are combined into a single image partitioned as follows: (top) single-image adaptive thresholding, (middle) single layer MRF and (bottom) two-image wavelet approach. Figure 6 shows a full-page example with comparisons of selected regions shown at the bottom. For all examples our approach provides subjectively the best results. The wavelet approach [15] produces comparable results in some cases but requires six-parameters to be tuned per example.

Quantitative results were obtained by counting the number of errors observed in the output images. Errors are considered any foreground word not detected correctly or any background/ink-bleed detected as foreground. Since ground-truth is not available, our quantitative results are still subjective as errors are decided by a human observer. However, in a best effort for fairness we found that our method had a precision accuracy of 85.96%, compared to 63.70% for [5], 71.94% for [15], and 75.00% for the single-layered MRF. Precision is defined as $(W - W^F)/(W + W^B)$, where W is the total number of foreground words, W^F is the number of incorrectly classified foreground words, and W^B are the number of stroke-size background or ink-bleed regions that were classified as foreground. Precision was computed for 10 full-page images with a total of 4896 words.

6. Discussion and Summary

Our results demonstrate the effectiveness of our approach for reducing ink-bleed. As with all supervised learning techniques training-data is needed. Our two step procedure for labeling data requires minimal user markup. While markup differs from input to input, only a dozen or so quickly drawn strokes are typically used per image. It is arguable that user-assisted markup is similar to setting parameters or thresholds, but we note that it is much easier for the domain user to specify image examples in lieu of tuning algorithmic parameters. Furthermore, in a long-term work-

[1]The PCA step of this algorithm is omitted for grayscale input images.

Figure 5. [Examples I-II are from microfilm; Examples III-IV are from RGB scans.] (a-b) Front and back with markup. The back markup is mirrored for clarity. (c-d) Output obtained using our dual-layered MRF approach. (e) Comparisons with adaptive thresholding [5], single layer MRF and wavelet [15] in top-middle-bottom format. Red lines separate the different results. Note that sub-regions are chosen to show some markup, however, the entire markup is not shown.

ing scenario it should be possible to use previously collected training-data on new pages with similar ink-bleed.

The documents targeted in our work have the same color ink throughout a page, however, our framework can be used on other types of documents exhibiting different background and foreground characteristics (e.g. multi-colored ink). Furthermore, different features for different document types and even classifiers (e.g. SVM) could be used in the likelihood computation with the same overall framework and dual-layer MRF setup remaining intact. Many old documents exhibit problems in addition to ink-bleed; e.g. the water-stains in Figure 5. Our classification approach is inherently more robust to these types of problems.

Lastly, we note that additional semantic information such as stroke direction or character characteristics were not used in our MRF. This was purposely done to circumvent tailoring our solution to the examples at hand. Incorporation of higher-level semantics into this framework undoubtedly deserves further consideration.

In summary, we have presented a novel framework for reducing ink-bleed. This framework provides a practical approach to ink-bleed removal that can target a wide range of examples and is suitable for use in a real-world setting.

Acknowledgements

We gratefully acknowledge the support of our colleagues from the National Archives of Singapore. This work was supported by the ASTAR SERC Grant No:0521010104.

References

[1] J. Bescos. Image processing algorithms for readability enhancement of old manuscripts. *Electronic Imaging*, 1:392–397, 1989.

[2] F. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Trans. PAMI*, 11(6):567–585, June 1989.

[3] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Trans. PAMI*, 26(9):1124–1137, 2004.

[4] M. Brown and W. B. Seales. Image restoration of arbitrarily warped document. *IEEE Trans. PAMI*, 26(10):1295–1306, 2004.

[5] F. Drira, F. L. Bourgeois, and H. Emptoz. Restoring ink bleed-through degraded document images using a recursive unsupervised classification technique. In *Document Analysis Systems (DAS'06)*, pages 38–49, 2006.

[6] F. C. M. et al. Toward on-line, worldwide access to vatican library materials. *IBM Journal of Research and Development*, 40(2):139–162, March 1996.

[7] S. Li. *Markov Random Field Modeling in Image Analysis (2nd Edition)*. Springer-Verlag, 2001.

[8] M. Pilu. Undoing paper curl distortion using applicable surfaces. In *CVPR'01*.

[9] G. Sharma. Show-through cancellation in scans of duplex printed documents. *IEEE Trans. on Image Processing*, 10(5):736–754, 2001.

[10] Z. Shi and V. Govindaraju. Historical document image enhancement using background light intensity normalization. In *ICPR'04*.

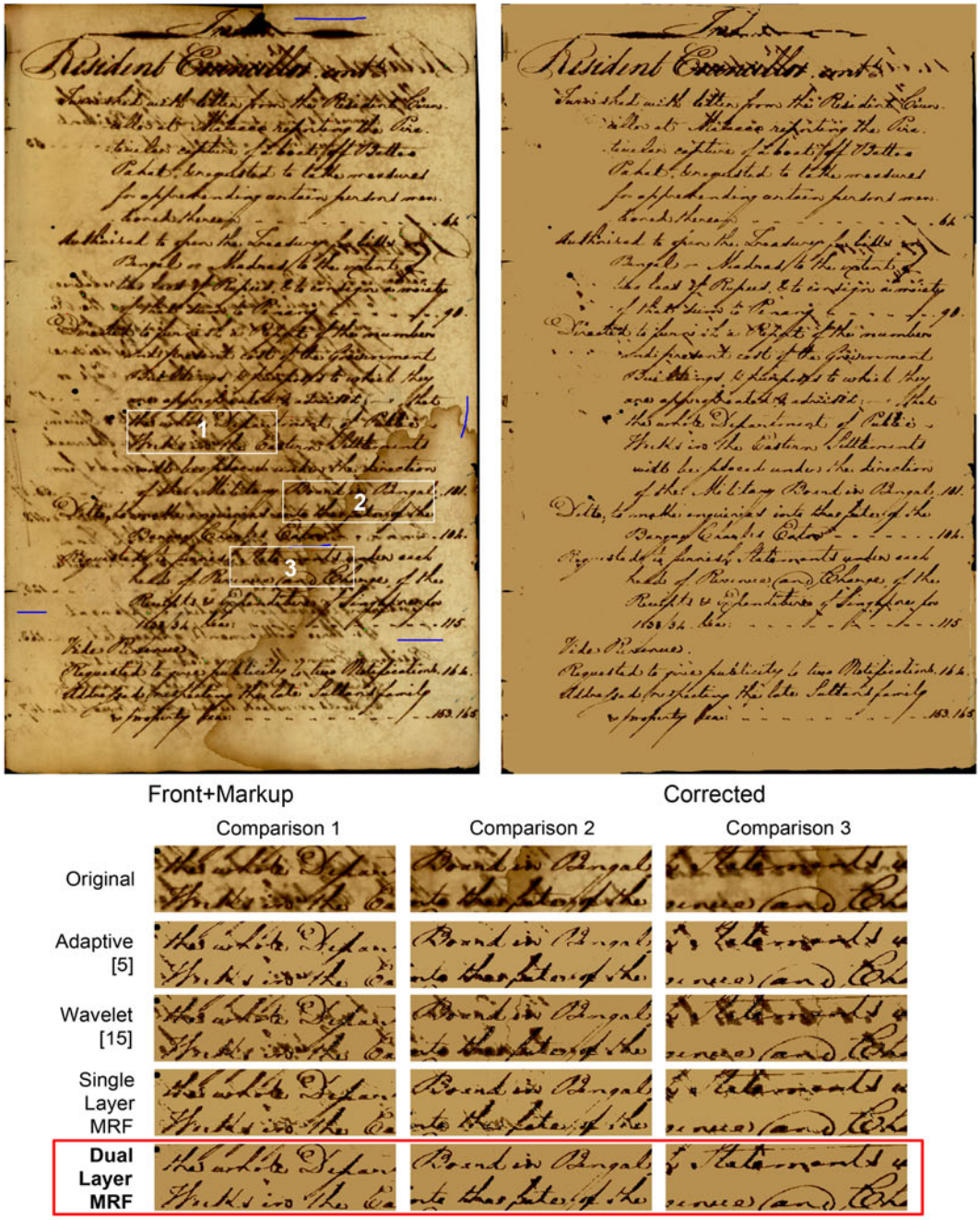

Figure 6. A full page example (RGB scan) with only the front is shown (top) with the complete user markup. Comparisons with other techniques are shown for three selected regions. Our dual-layer MRF approach produces the best results.

[11] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother. Comparative study of energy minimization methods for markov random fields. In *ECCV'06*.

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

[13] C. L. Tan, Z. Li, Z. Zhang, and T. Xia. Restoring warped document images through 3d shape modeling. *IEEE Trans. PAMI*, 28(2):195–208, 2006.

[14] A. Tonazzini, L. Bedini, and E. Salerno. Independent component analysis for document restortion. *International Journal on Document Analysis and Recognition*, 7:17–27, 2004.

[15] Q. Wang, T. Xia, L. Li, and C. Tan. Document image enhancement using directional wavelet. In *CVPR'03*.

[16] C. Wolf. Document ink bleed-through removal with two hidden markov random fields and a single observation field. In *Technical Report RR-LIRIS-2006-019*, 2006/2007.