

A Fully Automatic System for Restoration of Historical Document Images

Jie Wang

School of Computing
13 Computing Drive
National University of Singapore

Michael S. Brown

School of Computing
13 Computing Drive
National University of Singapore

Chew Lim Tan

School of Computing
13 Computing Drive
National University of Singapore

Abstract

Historical document images are subject to intrinsic distortions such as background noise and bleed-through interference due to aging and extrinsic distortions such as displacement, uneven surfaces introduced during image acquisition procedure. In this paper, we propose a fully automatic restoration framework that corrects bleed-through distortion on double-sided handwritten historical document images. First, the two sides of a document are registered with corresponding control points which are selected by inspecting the images' gradient maps and minimizing a predefined dissimilarity measure. The established correspondences are refined by median filters and consistency checking. Piecewise linear mapping function is chosen to represent the spatial relationship between the two images. Based on the estimated transform model, backward re-sampling strategy and bi-cubic spline interpolation are adopted to obtain final registered images. Once the two sides of a page have been registered, enhancement/smearing feature images are extracted and iterative wavelet decomposition/construction is performed to restore the degraded images. Experiments on the real documents from the National Archives of Singapore demonstrate a completely automatic framework to the restoration of historical document images.

Introduction

In this digital age, historical documents in the achieves are usually preserved in the form of scanned images or microfilms to enable long-term preservation and convenient access. Due to aging, these documents are subject to different types of degradation such as background noise, optical blur, low resolution, etc. In particular, double-sided handwritten historical documents additionally suffer from bleed-through distortion caused by the ink seeping from the reverse side of the page. Apart from the above intrinsic degradation of the document, the captured images are further suffered from extrinsic distortion introduced during image acquisition procedure. For instance, imprecise positioning of pages and different scale setting during scanning procedure may cause translation, rotation and scale distortion. Moreover, uneven

surfaces (such as near the spine area of a book) in the document lead to more complicated local distortion. To facilitate human perception and subsequent document analysis tasks (recognition, indexing, retrieval etc.), these distortions need to be detected and corrected.

Related Work

Several approaches to the correction of bleed-through distortion on historical document images have been proposed. In general, they can be categorized into either non-registration based approach or registration based approach, depending on whether the verso image of a page is available and precisely registered to the recto image. In non-registration based approaches, the two sides of a page are treated independently and processed separately. Representative methods include multistage thresholding techniques (Leedham et al. 2002) to extract clean text from degraded document images, local adaptive filters (Franke and Koppen 2001) to separate forensic handwritten text from background and noise-based thresholding techniques (Don 2001) to extract main text. In common, these approaches rely on the assumption that the foreground text layer and the background layer are separable.

On the contrary, registration based methods require both sides of a page being available and precisely registered. Using information from both sides, various classification methods can be used to extract foreground texts from degraded images. Tonazzini et al. (Tonazzini, Bedini, and Salerno 2004; Tonazzini, Salerno, and Bedini 2007) classify the separation of foreground texts and bleed-through interferences as a blind source separation (BSS) problem and approach it with independent component analysis (ICA). Tan et al. (Tan, Cao, and Shen 2002) propose the restoration method that iteratively enhances the foreground strokes and smears the interfering strokes by wavelet decomposition/construction.

It is noticed that the accuracy of these methods heavily depends on the precision of registration procedure. Perfect registration of recto images and verso images however is difficult, therefore in most registration based approaches including the above two, manual registration is conducted. As manual registration is time consuming and unable to deal with locally distorted images, fully automatic registration and bleed-through correction systems are required by the achieves. Wang et al. (Wang and Tan 2001) register the two

images using Murtagh’s point pattern matching method and restore documents with an orientation-constrained canny edge detector. In our previous work (Wang, Brown, and Tan 2008), template matching is applied to a particular sub-window and then each block to register the document and a heuristic based classifier is designed to restore images.

Our Contribution

In this paper, we propose a fully automatic restoration framework that corrects bleed-through distortion on doubled-sided historical document images. First corresponding control points (CCPs) are selected by inspecting the gradient maps and matched by a searching strategy minimizing the predefined dissimilarity measure. Median filters and modified consistency checking are applied to refine the established correspondences. In this work, we use piecewise linear functions to approximate the geometric deformation between the recto image and the verso image of a page. Registered images are then obtained by transforming the flipped verso image using estimated mapping functions and interpolating holes and overlaps using bi-cubic spline interpolator. Once the two images of a document are registered, prominent foreground strokes are extracted and guide the wavelet decomposition/construction procedure in enhancing/smearing foreground/bleed-through strokes. A general work flow of the restoration framework is shown in Figure 1. Compared with our previous work, this framework provides a more general, precise and stable method for the registration of the two sides of a page. The bleed-through correction method is also more sophisticated to deal with various degraded document images.

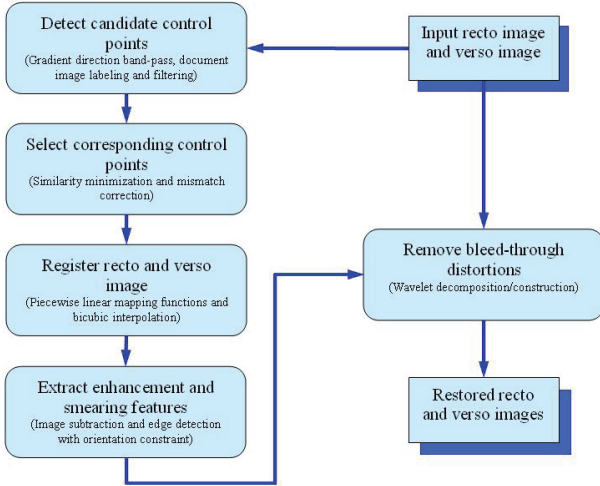


Figure 1: A general work flow of the restoration framework that corrects bleed-through distortion.

Document Image Registration

Registering the two sides of a page before restoration is crucial because imperfect image capturing procedure and complicated surface structures of historical documents lead to

spatial difference between the two images. Meanwhile this task is difficult for the following reasons: (1) The transform models between the two images and the degradation on the documents are unknown and variously complicated. (2) The matching counterparts are significantly different in terms of intensity and sharpness. (3) Sub-pixel registration accuracy is required while the severe background noise on historical documents significantly affects the registration accuracy. In view of all these difficulties, in this work, we use CCPs based approach to register the recto image and the verso image of a historical document. We assume that the input images are of the same spatial resolution and the two images of a document have been identified and paired. Before any subsequent processing, the verso image is flipped horizontally or vertically with respect to the document’s layout. For simplicity and with no loss of generality, the flipped verso image is referred to as target image and the recto image as reference image.

Candidate Points Selection

It has been observed that foreground strokes slanting at particular angles are much stronger than other foreground strokes and therefore have more chance to seep through the page. Meanwhile, according to our experience on manually selecting CCPs for historical document registration, points without foreground texts and bleed-through interference overlapping are much easier to be located and precisely matched. Based on the above two observations, we select corresponding control points from locations where gradient orientations lie in particular range and where only one side of the document has foreground texts.

First, the gradient maps and the labeled images are computed for both the reference image and the target image. Then the gradient magnitude map and the gradient orientation map for candidate points on the reference image and the target image are calculated as:

$$\begin{cases} \theta_{can} = \theta_{xy} \otimes (\theta_{xy} > \theta_l) \otimes (\theta_{xy} < \theta_h) \otimes B_{xy} \\ M_{can} = M_{xy} \otimes (\theta_{xy} > \theta_l) \otimes (\theta_{xy} < \theta_h) \otimes B_{xy} \end{cases} \quad (1)$$

where θ_{xy} , M_{xy} represent the gradient magnitude map and the gradient orientation map of original images and θ_{can} , M_{can} are the gradient magnitude map and the gradient orientation map of the identified candidate control points. θ_l , θ_h are the parameters thresholding the gradient maps and $\otimes$ represents Hadamard product. In the first phase of this procedure, θ_l and θ_h are set as -1.8 and -0.5 respectively. Figure 2 demonstrates this selection procedure with a sample document.

CCPs Detection

Once candidate points on the target image and the reference image are identified, the correspondences between them are established by optimizing a dissimilarity measure which is predefined as:

$$C(x, y, x', y') = w_i(I_{xy} - I'_{x'y'}) + w_m(M_{xy} - M'_{x'y'}) + w_\theta(\theta_{xy} - \theta'_{x'y'}) + w_d\sqrt{(x - x')^2 + (y - y')^2}$$

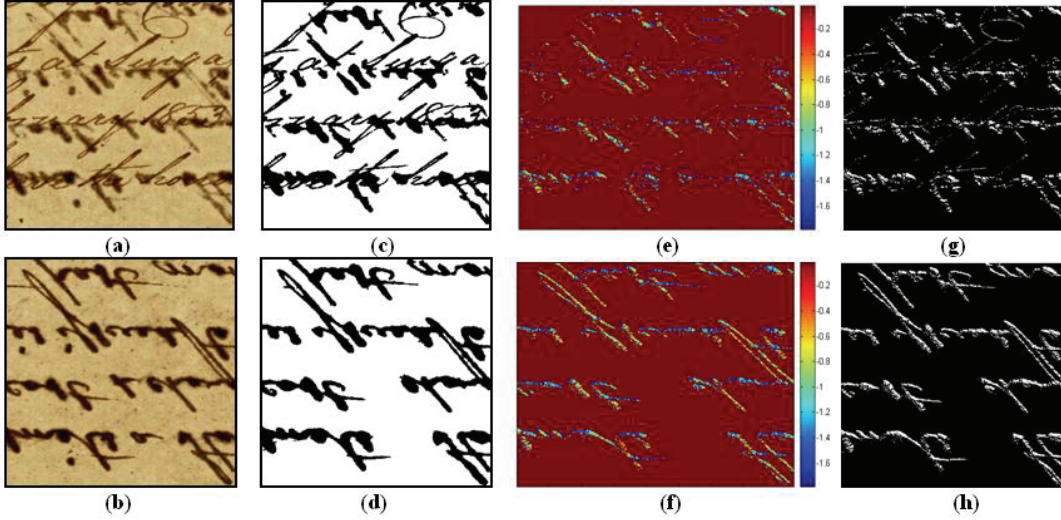


Figure 2: (a-b) original recto image and horizontally flipped verso image (cropped from 1628*2480); (c-d) labeled recto image and verso image; (e-f) band-pass filtered gradient orientation maps of the recto image and the verso image; (g-h) identified candidate control points in the recto image and the verso image.

where I_{xy} , M_{xy} , θ_{xy} are the intensity, gradient magnitude, and gradient orientation of the candidate point (x, y) in the target image respectively; and $I'_{x'y'}$, $M'_{x'y'}$, $\theta'_{x'y'}$ are the corresponding values for the candidate point (x', y') in the reference image. w_i , w_m , w_θ and w_d are the weights that specify the relative importance of the intensity, gradient magnitude, gradient orientation, and displacement in the determination of the correspondences. The optimization procedure is performed in the way that for each candidate point (x', y') on the reference image, searching over a rectangular window on the target image to find the most similar candidate point. The search window has a size of $r * r$ and is centered at the point (x, y) which is exactly beneath (x', y') .

As we have discussed, perfect matching is difficult because of the intrinsic differences between the matching counterparts thus mismatches are expected to occur. As this error diminishes the registration accuracy and even leads to the failure of registration, it should always be avoided or at least alleviated (Zitova and Flusser 2003). For this purpose, we make use of median filters to correct non-collectively occurring mismatches (Goshtasby, Turner, and Ackerman 1992). The median filters can be represented as:

$$\begin{cases} dx_i = \mathbf{Median}(x_i - x'_i) \\ dy_i = \mathbf{Median}(y_i - y'_i) \end{cases} \quad (2)$$

To further correct collective mismatches and improve the quality of the detected CCPs, the reference image and the target image are switched to conduct consistency checking. The parameters θ_l and θ_h are also switched to 0.5 and 1.8. Figure 3 shows some detected CCPs. For clarity, only part of CCPs with high similarity are shown.

Mapping Function Estimation

In order to register two or more images, the type of transform model between images must first be chosen based on the

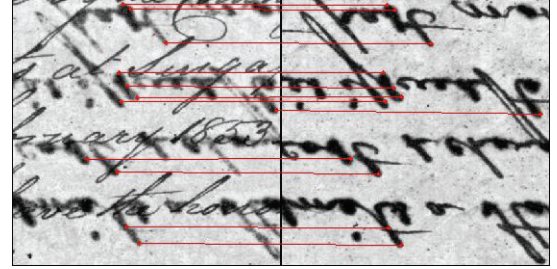


Figure 3: Illustration of the detected CCPs from sample images in Figure 2.

priori knowledge about the image acquisition process and the expected image degradation. In this work, we deploy Goshtasby's piecewise linear mapping function (Goshtasby, Stockman, and Page 1986) to approximate the spatial relationship between the recto image and the verso image of a page. The local function can be shown as:

$$\frac{1}{N} \sum_{j=1}^m \sum_{i=1}^{n_j} \left\{ [X_i - (a_0 + a_1x_i + a_2y_i)]^2 + [Y_i - (b_0 + b_1x_i + b_2y_i)]^2 \right\}^{1/2}$$

where n_j is the number of control points under patch j . With this method, the target image is first triangulated with respect to the distribution of the detected CCPs and linear mapping function is estimated for each triangulated patch according to the detected CCPs located in this patch. Although this function is not yet perfect for the representation of geometric distortion between the two sides of historical documents, it is to some extent sensitive to the warping structures near the spine area of a book.

Re-sampling and Interpolation

Using the mapping functions estimated in the above procedure, the registered target image is constructed. As transformed coordinates could be fractional, discretization and rounding are inevitable which may lead to holes and overlaps on the reconstructed target image. To address this, cubic interpolation is adopted as:

$$P(x, y) = \sum_{i=0}^3 \sum_{j=0}^3 a_{ij} x^i y^j \quad (3)$$

where a_{ij} are 16 coefficients which can be determined by the values and derivatives of the functions at the four corners of the square centering at (x, y) .

Image Restoration

Once the two images of a document are registered, we adopt the wavelet-based technique proposed in (Tan, Cao, and Shen 2002) to restore document images. This approach can be described as follows:

- Compute the foreground overlay image $a(x, y)$ for image $f(x, y)$ by:

$$a(x, y) = \text{flip}(\text{invert}(b(x, y))) + f(x, y) \quad (4)$$

where $b(x, y)$ is the corresponding reverse side image.

- Weaken the suspected bleed-through interference on the foreground overlay image by scaling it with a nonlinear transform:

$$\text{curve}(x) = 2^k - 1 - ((2^k - 1)^2 - x^2)^{0.5} \quad (5)$$

- Detect the foreground strokes on the foreground overlay image using a modified canny edge detector with orientation filters and constraints (Cao et al. 2000; Tan et al. 2000) to form the “enhancement feature image”, $E(x, y)$.
- Conduct the above steps on $b(x, y)$ to get the “smearing feature image”, $S(x, y)$.
- Decompose an original image into wavelet domain while retaining the size of the image as:

$$wf(x, y) = \{C_j(x, y), D_j^k(x, y), j = 0, \dots, k = 1, 2, 3\}$$

where $C_j(x, y)$ is the wavelet approximation coefficient and $D_j^k(x, y)$, ($k = 1, 2, 3$) are the wavelet detail coefficients at scale j of the wavelet decomposition.

- Modify the wavelet detail coefficients by referring to $E(x, y)$ and $S(x, y)$:

$$D_j^k(x, y) = \begin{cases} e_j^k D_j^k(x, y) & \text{if } E(x, y) == 1 \\ s_j^k D_j^k(x, y) & \text{if } S(x, y) == 1 \end{cases} \quad (6)$$

where $j = 0, \dots, J; k = 1, 2, 3$.

- Reconstruct the wavelet representation with the modified wavelet detail coefficients.
- Iteratively repeat the decomposition/construction procedure at most 15 times to get enhanced/smear gray images.
- Apply the same edge detector to the enhanced images to obtain final output images.

Experimental Results

The proposed framework has been tested with 56 (ie. 28 pairs of) real historical document images obtained from the National Archives of Singapore and shown encouraging results. These images are typically scanned at 150dpi and mostly with a size of 1800×2800 , a few of them with a larger size like 3000×4500 . Based on the severity of bleed-through distortion, we group these images into: slightly interfered (12), intermediately interfered (18) and heavily interfered (26).

To evaluate the performance of the proposed framework, visual assessment is first performed by the experts working at the archives. In particular, we compare the resultant images with those restored by other three methods (Tan et al. 2000; Wang and Tan 2001; Wang, Brown, and Tan 2008). Similar to the results shown in Figure 4, this approach produces better images for most tested documents.

To quantitatively assess our method, we selected 12 documents from the three groups (3, 3, 6 respectively) and manually calculated the “ground truth” for these images. In particular, we counted the following numbers: the total number of foreground words on the original document, denoted as N_{fgd} ; the number of words detected by restoration methods, denoted as N_{detect} ; the number of words correctly detected, denoted as $N_{correct}$. Among these values, detected words refer to all words appearing in the resultant image, which include fully detected foreground words and partially or fully detected interfering words. If some characters in a foreground word were lost, the whole word was considered as lost or undetected. If some characters or parts of words from the reverse side were picked up by the system, the number of detected words increased by one. However, only fully detected foreground words were counted in correctly detected words $N_{correct}$. When calculating these numbers, connected words were treated as separate words. We measured the performance of the proposed method by using the traditional document analysis metrics (Junker, Hoch, and Dengel 1999) defined as:

$$\text{Precision} = \frac{N_{correct}}{N_{detect}}, \quad \text{Recall} = \frac{N_{correct}}{N_{fgd}} \quad (7)$$

Table 1 shows the evaluation of recovery results for the 12 images used for testing. The average precision and recall of our approach are 96% and 94.7% respectively. As a comparison, the results from Wang’s method (Wang and Tan 2001) and our previous work (Wang, Brown, and Tan 2008) are also listed. As the method in (Wang, Brown, and Tan 2008) adopts an additional recovery procedure, its precision is a little higher than its actual case. However, as shown in the table, the new method still achieved higher correction precision and recall on most images. Moreover, experiments show that the framework achieves good results even on heavily degraded document images. One such example can be seen in Figure 5.

Besides, the framework is much more efficient than the other two approaches. In general, less than 1 minute is needed to process a typical-sized document. The fast process can be explained by the quick selection of CCPs. More-

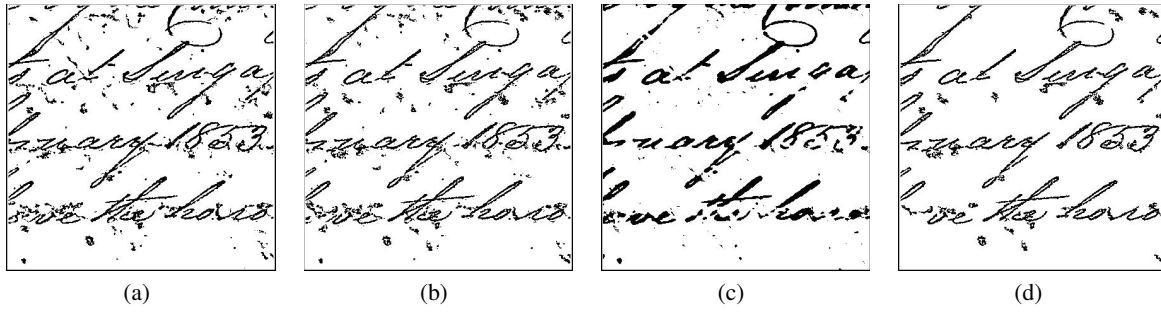


Figure 4: Restored images for Figure 2 (a) resulted from four methods. (a) result from method in (Tan et al. 2000); (b) result from WangQ’s method (Wang and Tan 2001); (c) result from method in (Wang, Brown, and Tan 2008); (d) result from this framework

Image number No. of words		1 61	2 121	3 117	4 189	5 106	6 243	7 170	8 135	9 206	10 103	11 158	12 196	Average
Precision (%)	Wang:2001	98	79	87	86	83	94	87	95	91	89	90	93	89.3
	Wang:2008	98	89	92	95	88	90	96	99	93	97	96	94	93.9
	New	97	94	98	94	95	97	96	98	91	98	97	98	96
Recall (%)	Wang:2001	95	86	100	79	83	81	87	84	84	76	92	84	85.9
	Wang:2008	97	92	100	88	87	92	89	95	91	85	89	93	91.5
	New	97	96	100	94	93	92	95	97	92	91	93	96	94.7

Table 1: The comparison of newly proposed framework with the approaches presented in (Wang and Tan 2001) and (Wang, Brown, and Tan 2008).

over, it scales well, which means the computational cost will not significantly increase when considerably large images are to be processed.

Conclusion

In this paper, we present a fully automatic restoration framework to correct bleed-through distortion on double-sided historical handwritten document images. The restoration is based on the accurate registration of the recto image and the verso image of a page. For the purpose of registration, corresponding control points are selected by inspecting the gradient maps and matched by minimizing the predefined dissimilarity metric. Consistency checking and median filters are applied to correct collectively and non-collectively occurring mismatches. The spatial distortion between the two images are approximated with piecewise linear mapping function and backward re-sampling and bi-cubic spline interpolation are deployed to construct the registered images. Once the document is registered, wavelet decomposition/construction is used to enhance foreground texts and smear bleed-through interference so that the document is well restored. Although the restoration results are not perfect yet, experiments on the real document images from the archives demonstrate encouraging results. Further improvements can be made by exploiting more precise transform models to represent the variously locally-distributed distortion on historical document images.

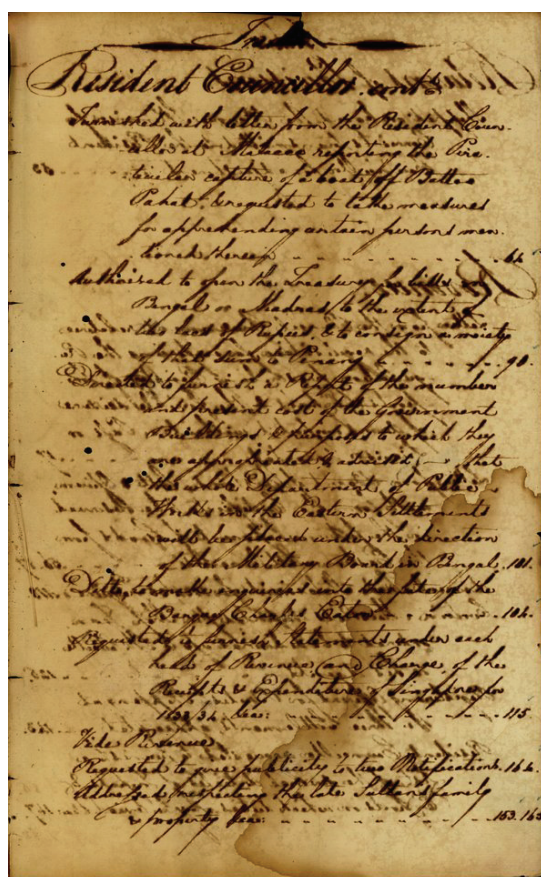
Acknowledgements

This research is supported by the Media Development Authority (MDA) of Singapore, under research grant R-252-

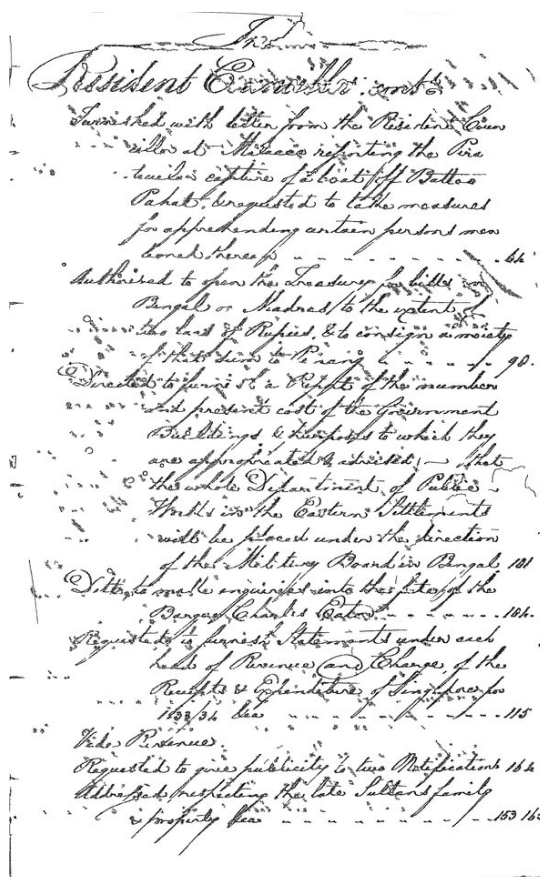
000-325-279. The authors would like to thank the National Archives of Singapore for the use of their archival documents.

References

- Cao, R.; Tan, C. L.; Wang, Q.; and Shen, P. 2000. Segmentation and analysis of double-sided handwritten archival documents. In *IAPR International Workshop on Document Analysis Systems (DAS'00)*, 147–158.
- Don, H. 2001. A noise attribute thresholding method for document image binarization. *International Journal on Document Analysis and Recognition* 4(2):131–138.
- Franke, K., and Koppen, M. 2001. A computer-based system to support forensic studies on handwritten documents. *International Journal on Document Analysis and Recognition* 3(4):218–231.
- Goshtasby, A.; Stockman, G.; and Page, C. 1986. A region-based approach to digital image registration with subpixel accuracy. *IEEE Transactions on Geoscience and Remote Sensing* 24:390–399.
- Goshtasby, A.; Turner, D.; and Ackerman, L. 1992. Matching of tomographic slices for interpolation. *IEEE Transactions on Medical Imaging* 11(4):507–516.
- Junker, M.; Hoch, R.; and Dengel, A. 1999. On the evaluation of document analysis components by recall. In *Precision and Accuracy, International Conference on Document Analysis and Recognition, India*, 713–716.



(a)



(b)

Figure 5: A historical document image with severe bleed-through distortion and background noise and the resultant image produced by our restoration system

Leedham, G.; Varma, S.; Patankar, A.; and Govindarayu, V. 2002. Separating text and background in degraded document images: A comparison of global thresholding techniques for multi-stage thresholding. In *Proceedings of the Eighth International Workshop on Frontiers in Handwriting Recognition (IWFHR'02)*, 244. Washington, DC, USA: IEEE Computer Society.

Tan, C.; Cao, R.; Shen, P.; Wang, Q.; Chee, J.; and Chang, J. 2000. Removal of interfering strokes in double-sided document images. *IEEE Workshop on Applications of Computer Vision* 0:16.

Tan, C. L.; Cao, R.; and Shen, P. 2002. Restoration of archival documents using a wavelet technique. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24(10):1399-1404.

Tonazzini, A.; Bedini, L.; and Salerno, E. 2004. Independent component analysis for document restoration. *Int. J.*

Doc. Anal. Recognit. 7(1):17-27.

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

Wang, Q., and Tan, C. L. 2001. Matching of double-sided document images to remove interference, pattern recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, CVPR 2001, 11-13.

Wang, J.; Brown, M.; and Tan, C. L. 2008. Accurate alignment of double-sided manuscripts for bleed-through removal. In *The Eighth IAPR International Workshop on Document Analysis Systems (DAS'08)*, 69-75. Los Alamitos, CA, USA: IEEE Computer Society.

Zitova, B., and Flusser, J. 2003. Image registration methods: a survey. *Image and Vision Computing* 21(11):977-1000.