Reduction of Bleed-through Effect in Images of Chinese Bank Items

Bingyu Chi
Graduate School at Shenzhen
Tsinghua University
Shenzhen,China
AmandaChi.XC@gmail.com

Youbin Chen
Graduate School at Shenzhen
Tsinghua University
Shenzhen,China
chenyb@sz.tsinghua.edu.cn

***Abstract*—Because of the existence of possible carbon and seals, it's quite often that images of financial documents such as Chinese bank checks are suffered from bleed-through effects, which will affect the performance of automatic financial document processing such as seal verification and OCR.**

This paper presents an effective algorithm to deal with bleed-through effects existing in the images of financial documents. Double-sided images scanned simultaneously are used as inputs, and the bleed-through effect is detected and removed after the registration of the recto and verso side images. There are two major aspects of contribution in our work. First, our algorithm can deal with images with complex background from real-life financial documents while most other algorithms only deal with images with simple background. Second, we combine the fast ICA algorithm with Gatos' local adaptive thresholding algorithm [1] to deal with the bleed-through effects. Experiments show that our proposed algorithm is very promising.

***Keywords*-Bleed-through;Show-through;Binarization;Fast ICA;Document image processing;**

I. Introduction

Bleed-through is a common problem in low-quality document images. Reducing this effect in scanned documents, either handwritten or machine-printed ones, has been a difficult problem in degraded document image analysis. There has been a lot of research work on this topic which can be divided into two categories.

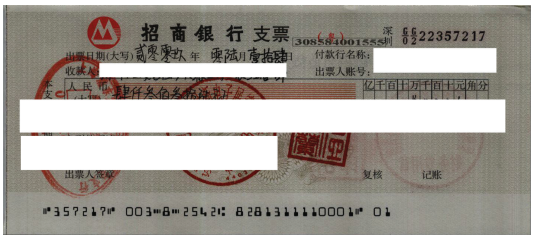
The first category is to process single-side documents only. Multi-stage thresholding is widely used. For example, in [2], the authors described several commonly used methods based on various kinds of entropy. Hysteresis thresholding was tried in [3]. Edge detection was used in [4]. The authors observed that the edges of the sipping strokes from the reverse side were not as sharp as those on the front side. So a thresholding operation applied after the extraction of edges might be helpful and their results proved this. All these methods can be used globally or locally and can solve the bleed-through effects efficiently but their results are not good enough because perfect thresholds can hardly be found or they may not exist actually.

The second category is to process double-sided documents. Sometimes the pixel brightness of those bleed-through effects is even darker than that of the real document pixels, it is impossible to remove the bleed-through effects completely with a simple single-side operation. When scanned images of both sides are available, the problem seems easier but not really. The first challenging task is the registration of the recto and verso images. Several papers are based on the hypothesis that two sides have already been registered. Some other research work provides a solution. The method in [5] was based on parameter optimization of an affine transformation, computed by minimizing the sum of the squared difference between corresponding pixels in the recto and verso images. According to their results, it did a perfect job. In [6], a fully automatic method which detected specific number of corresponding control points from manuscript document was presented. As this method dealt with manuscript documents only, it had limitation. The authors in [7] summarized some registration methods including correlation-based method, Fourier-Mellin method, and some methods mentioned above. Fourier-Mellin method was recommended. After registration, several algorithms were supplied to separate the verso and recto images. In [6], identification of bleed-through regions and replacing the bleed-through pixels with estimated background were carried out. The authors in [7] made use of decorrelation technique and the authors in [8] proposed a linear pattern overlapping model which allowed them to use blind source separation(BSS) techniques. ICA, PCA and whitening helped them to accomplish the goal. However, the approximation of linear model caused some unwanted effects which were observed from their results. BSS is a popular technique in solving this kind of problem. It was also used in [9]. Other methods such as Hidden Markov Random fields [10] and neural networks [11] were also used for bleed-through or show-through effect removal.

None of the above methods is perfect in dealing with the bleed-through effects existing in color Chinese bank item images. In this paper, we will focus on the bleed-through effects in double-sided Chinese bank item images (recto and verso) scanned simultaneously as shown in Fig.1. First, a simple method to register the recto and verso images will be carried out. Then, a multi-stage binarization approach and a fast ICA algorithm to reduce the bleed-through effects will be done. Last, the bleed-through pixels will be replaced with estimated background.

978-0-7695-4774-9/12 $26.00 © 2012 IEEE
DOI 10.1109/ICFHR.2012.260

The rest of the paper is organized as follows. Section 2 will present the workflow of our proposed algorithm and will describe the implementation details. In section 3, experimental results will be shown. Finally, our conclusion and future work will be presented in section 4.



(a) Recto side

(b) Verso side

Figure 1. An example of low-quality Chinese bank check images with double-side scanned simultaneously (for security reason, some parts are hidden).

(a) Recto side (b) Verso side

Figure 3. An example of bleed-through effect caused by seals

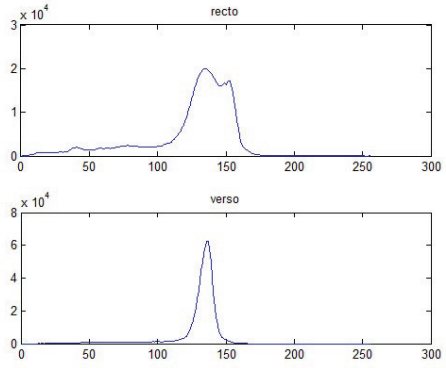


Figure 4. Histograms of gray-scale images of typical Chinese bank items

II. Implementation Details

The workflow of our proposed algorithm is shown in Fig.2. The input color image is converted to grayscale and the verso side is flipped horizontally in order to simplify the process. For convenience, Fig.3 is used as an example to explain our proposed algorithm.

A. Registration

Since both sides of a document are scanned simultaneously, only translation exists between the recto and verso images. This greatly simplifies the registration. There are two steps in this procedure. One is the the rough binarization and the other is the parameter optimization of the translation transformation.

In order to do the registration easily, the same regions in both the recto and verso images should be found. A rough binarization for both sides is carried out. In the binary image,

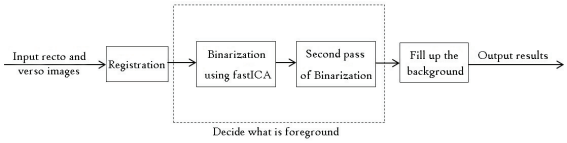


Figure 2. The flow chart of our proposed algorithm

we expect that the value '1' corresponds to foreground pixels and possible pixels of bleed-through effect from the reverse side. A simple threshold could help. First, the gray level histograms of recto and verso images are computed. Different kinds of Chinese bank items have different distribution in histogram, but almost all of them have an obvious peak representing most of the text information as shown in Fig.4. This peak represents the most possible pixels of the bleed-through effects from one side to the other. Assume the gray value corresponding to gray histogram peak is p, we use $k \cdot p$ as threshold to binarize the image, where k is a scale parameter. In our experiments k is 0.8. An example of rough binarization is shown in Fig.5(a).

Assume the binary image of recto is R_B, and binary image of verso is V_B. R_B and V_B have some common regions, which might be the part suffered from bleed-through effects. Since the recto and verso are scanned simultaneously, we can register two images by solving the following optimization function:

$$(a, b) = \arg\min \sum |R_B(x, y) - V_B(x + a, y + b)|$$

Then we can get registered recto (R_R) and verso (V_R):

$$\begin{cases} R_R(x, y) = R(x, y) \\ V_R(x, y) = V(x + a, y + b) \end{cases}$$

(a) Rough binarization

(b) Result of fastICA

(c) Decision after fastICA

(d) Final decision

Figure 5. Intermediate processes of our proposed algorithm

B. First pass of Binarization

This part is crucial in our proposed algorithm. The information from the reverse side is used in the first pass of binarization. A basic decision on what are foreground pixels in the two sides is decided by this step. Both the recto and verso images are preprocessed first. After that the common regions of the two images are extracted. Each image is converted into a one-dimensional vector. These two vectors are considered as linear combination of two independent vectors and de-correlation is applied with them. In particular, fast ICA algorithm is used and it is a part of the contribution of our proposed algorithm in this paper.

First, the mean of all the gray-scale pixel values in each of the two images is calculated. In order to get the same mean value of the two images, the gray-scale value of every pixel in one of the two images should be scaled. From this process, R_A and V_A are got accordingly. Second, the AND operation is applied with the rough binarization results R_B and V_B to generate a mask by extracting pixels which are "1" on both sides. The pixels in this mask represent the positions of all the pixels possibly suffered from bleed-through effects. Third, we multiply R_A and V_A by this mask to get R_C and V_C. Treating them as two vectors r_c and v_c , we can apply fast ICA algorithm with them, which can solve the 2×2 matrix A and get two independent vectors r_i and v_i as follows.

$$\begin{pmatrix} r_c \\ v_c \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} r_i \\ v_i \end{pmatrix}$$

At this stage, the more possible a pixel belongs to the original foreground of one side, the bigger is the number representing that pixel on the very side in r_i or v_i . This is shown in Fig.5(b). Thus, a threshold is needed to decide which pixel belongs to the foreground on each side. We apply a method similar to Otsu thresholding on two vectors r_i and v_i respectively. Combining them with the pixels which are not suffered from bleed-through effects, we can get the decision as shown in Fig.5(c).

C. Second pass of Binarization

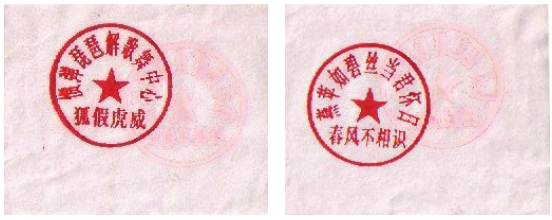
Since noise could exist after the first pass binarization, the second pass of binarization is necessary. An adaptive binary approach by B.Gatos in [1] is used and it is proved to be very effective. It makes full use of the background information and it is complimentary to the first pass binarization. The combination of fast ICA algorithm and the local adaptive thresholding such as Gatos algorithm is the second part of the contribution of our proposed algorithm in this paper. Our experiments show that this method is quite effective in dealing with bleed-through effects in real-life financial document images with complex background.

Similar to Gatos algorithm, background surface is estimated. The image is binarized according to the difference between the estimated background surface and its original image. During the background estimation, a pixel of possible foreground is replaced with the average gray value of background pixels around it. The background surface estimation $B(x, y)$ is generated by traversing all the possible foreground pixels. After that, the second pass of binarization result $T(x, y)$ is given by the following function according to the estimated background $B(x, y)$ and the original image $I(x, y)$:

$$T(x,y) = \begin{cases} 1 & \text{if } B(x,y) - I(x,y) > d(B(x,y)), \\ 0 & \text{otherwise.} \end{cases}$$

where $d(B(x, y))$ is a function of $B(x, y)$ and decided by the gray level of $B(x, y)$. The brighter the estimated background $B(x, y)$ is, the larger $d(B(x, y))$ will be.

After processing recto and verso images separately with this method, we can get the final binary image, where '1' indicates the final decision of the foreground, as shown in Fig.5(d).



(a) Recto side (b) Verso side

Figure 6. Final Result

D. Filling up background

Those pixels of bleed-through effects, i.e. the pixels are '1' in rough binary image but '0' in final binary image, are replaced with estimated background. The same approach as estimating the background surface in the second pass of binarization is used here. For each pixel of the bleed-through effect, it is replaced with the average gray-scale value of the estimate background around it. Then the final result after bleed-through removal is generated as shown in Fig.6. To deal with color images, the RGB channels are processed respectively.

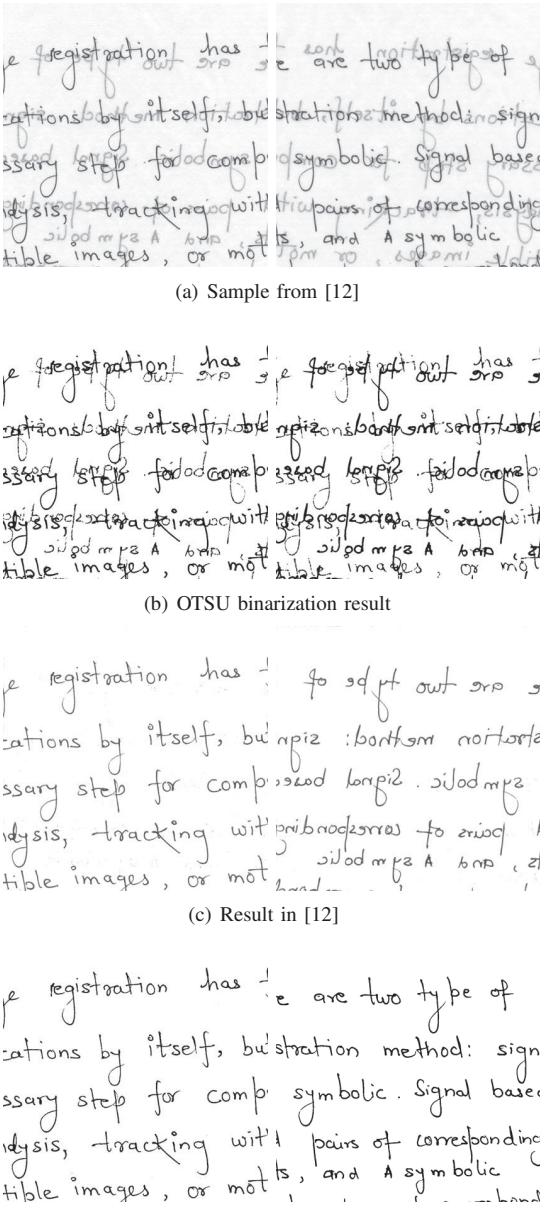
III. Experiments

Our proposed algorithm was tested on several kinds of sample images such as those suffered from bleed-through effect as shown in Fig.3 and some other document images of Chinese bank items with complex background. Some examples used in our experiments are shown in Fig.7 and Fig.8. Fig.7(a) was extracted from [12] and the results in [12] were shown for comparison. It can be seen that our result is better than that from [12] when processing simple degraded document images. Fig.8 shows a part of real Chinese bank items suffered from bleed-through effects caused by carbon. The handwritten digits in recto side can't be recognized after OTSU binarization but our algorithm can improve the image quality and make the digits recognizable.

Our proposed algorithm is helpful to OCR/ICR engines when dealing with the situations shown in Fig.7 and Fig.8. Also, we tested our proposed algorithm on a lot of images of Chinese bank items suffered from bleed-through effects caused by seals. The results are very similar to the one shown in Fig.6. So they are not presented here. This has demonstrated that our proposed algorithm is very promising to deal with the bleed-through effects.

IV. Conclusion and Future Work

In this paper, we proposed an algorithm to deal with the bleed-through effects in complicated financial document images such as Chinese bank items. The proposed algorithm includes the following procedures: registration, binarization using information from reverse side, binarization using background information, and filling up the background after the removal of bleed-through effects.



(a) Sample from [12]

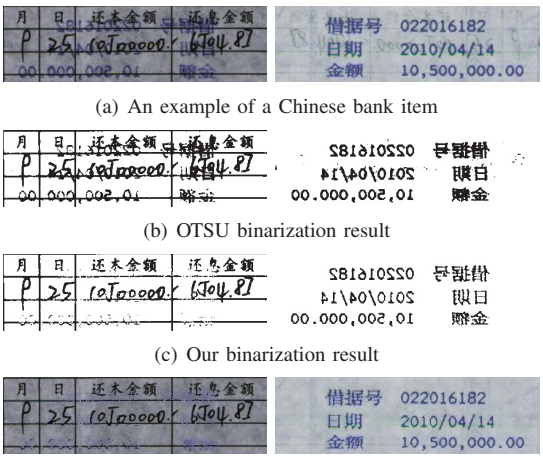
(b) OTSU binarization result

(c) Result in [12]

(d) Our result

Figure 7. Sample document image with simple background

There are two aspects of contribution in our work proposed in this paper. One is that our algorithm can deal with images with complex background from real-life Chinese financial documents such as checks while most other algorithms deal with images with simple background. The other is that we combine the fast ICA algorithm with a local adaptive thresholding algorithm called Gatos thresholding [1] to deal with the bleed-through effects. Our method is simple but effective. As it shows, our proposed algorithm performs very well. In the future, we will reduce the time complexity of the proposed algorithm and research on more



(a) An example of a Chinese bank item

(b) OTSU binarization result

(c) Our binarization result

(d) Final result

Figure 8. Real-life Chinese bank documents with complex background

universal methods.

REFERENCES

[1] B. Gatos, I. Pratikakis, and S. J. Perantonis, "Adaptive degraded document image binarization." *Pattern Recognition*, pp. 317–327, 2006.

[2] J. M. M. da Silva, R. D. Lins, F. M. J. Martins, and R. Wachenchauzer, "A new and efficient algorithm to binarize document images removing back-to-front interference." *J. UCS*, pp. 299–313, 2008.

[3] R. Estrada and C. Tomasi, "Manuscript bleed-through removal via hysteresis thresholding." in *ICDAR'09*, 2009, pp. 753–757.

[4] C. L. Tan, R. Cao, P. Shen, Q. Wang, J. Chee, and J. Chang, "Removal of interfering strokes in double-sided document images," in *IEEE Workshop on the Application of Computer Vision, WACV2000*, 2000, pp. 4–6.

[5] E. Dubois and A. Pathak, "Reduction of bleed-through in scanned manuscript documents." in *PICS'01*, 2001, pp. 177–180.

[6] J. Wang, M. S. Brown, and C. L. Tan, "Automatic corresponding control points selection for historical document image registration." in *ICDAR'09*, 2009, pp. 1176–1180.

[7] G. Bianco, F. Bruno, A. Tonazzini, E. Salerno, and E. Console, "Recto-verso registration, enhancement and segmentation of ancient documents," in *Virtual Systems and Multimedia, 2009. VSMM '09. 15th International Conference on*, sept. 2009, pp. 131 –136.

[8] A. Tonazzini, E. Salerno, and L. Bedini, "Fast correction of bleed-through distortion in grayscale documents by a blind source separation technique." *IJDAR*, pp. 17–25, 2007.

[9] B. Ophir and D. Malah, "Show-through cancellation in scanned images using blind source separation techniques." in *ICIP (3)'07*, 2007, pp. 233–236.

[10] C. Wolf, "Document ink bleed-through removal with two hidden markov random fields and a single observation field." *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 431–447, 2010.

[11] X. Zhang, J. Lu, and T. Yahagi, "Blind separation methods for image show-through problem," in *Information Technology Applications in Biomedicine, 2007. ITAB 2007. 6th International Special Topic Conference on*, nov. 2007, pp. 255 –258.

[12] R. F. Moghaddam and M. Cheriet, "A variational approach to degraded document enhancement." *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1347–1361, 2010.