PRINTER FORENSICS BASED ON PAGE DOCUMENT'S GEOMETRIC DISTORTION

Yubao Wu[*], *Xiangwei Kong*[*], *Xin'gang You*[+], *Yiping Guo*[*]

[*]School of Electronic and Information Engineering, Dalian University of Technology, Dalian, 116024, China

[+]Beijing Institute of Electronic Technology and Application, Beijing, 100091, China

ABSTRACT

A printed document can provide intrinsic features of the printer so as to distinguish which printer it comes from. But how to extract the intrinsic features is critical in printer forensics. In this paper, the page document's geometric distortion is extracted as the intrinsic features, and a printer forensics method based on the distortion is proposed. Firstly projective transformation model is used to model the geometric distortion. After the feature point set of the model is extracted, the model's parameters considered as the geometric distortion features can be estimated, and then the model's error pattern can be obtained. During the process, the least squares method is used to estimate the model's parameters, and SVM technique is used for classification. The effectiveness of the model's parameters in the printer forensics is demonstrated by experimental results.

Index Terms— Printer forensics, intrinsic features, geometric distortion, projective transformation

1. INTRODUCTION

Due to the wide use of printed and scanned documents, there are an increasing number of cases related to forged documents. An important issue which concerning printed documents in many applications is related to the integrity of these printed documents. In recent years, a non-destructive passive printer forensics technology which can be used to identify the authenticity of documents has been developed. To address such issues, how to link a given printed document to its source printer has recently received attention since 2002 [1-6].

Oliver et al. [1] introduced several print quality metrics including line width, raggedness and over spray, dot roundness, perimeter and number of satellite drops. Edward Delp and his colleagues exploited printer banding artifacts, Mikkilineni et al. [2] extracted graylevel cooccurrence features from the printed letter "e". Cyril Murie et al. [3] employed invariant moments for printer forensics. Shen Linjie et al. [4] studied the "banding" noise in the character image through Gaussian filter, and then the statistic features denoting the characteristics of each printer are extracted using image quality measures from the noise. Thomas Breuel et al. [5] experimented on the graylevel features which are based on the general textures and description of edges, and gave detailed experimental results. Farid et al. [6] used PCA method to model the degradation in a document caused by printing, and the resulting printer profile was then used to detect the source of document according to the printer device.

The methods mentioned above all extracted features from one connected component or local character's area in the document image, while this paper extracted features from the whole document image. We found that when comparing with ideal images, page document images have geometric distortion. In this paper we model the geometric distortion by projective transformation, and solve the parameters of the model by the least squares method. The resulting parameters of the model are then used to decide the link between document and printer.

The rest of this paper is organized as follows. Section 2 introduces the diagram of the proposed method, and Section 3 describes the method in detail. Section 4 shows experimental results on the forensics of 10 printers.

2. THE EXPERIMENTAL DIAGRAM OF PRINTER FORENSICS METHOD BASED ON PAGE DOCUMENT'S GEOMETRIC DISTORTION

In this section, we first discuss the geometric distortion of a document introduced by printing process, and then describe the diagram of printer forensics method based on page document's geometric distortion.

2.1. The geometric distortion of printed document

On ideal circumstances, the rows in one page of the document are strictly parallel. But actually, for printed document the slope of the rows changes regularly along the printing direction. In some printers, the row slope becomes smaller gradually from top to bottom of one page; while some becomes larger. This phenomenon is called the geometric distortion of printed document. It may be caused by the imperfection and differences of the paper feeding mechanism of different printers.

This geometric distortion will turn ideal parallel lines into intersecting lines, so the projective transformation could be applied to model the distortion. Further analysis of the modeling of geometric distortion will produce intrinsic features to be used in printer forensics.

2.2. The diagram of proposed method

The block diagram of the printer forensics method based on page document's geometric distortion is shown in Figure 1. We need get the ideal image and the document image from the word document page to built projective transformation model. On the one hand, print a general Word document page, and then scan the document at 600 dpi resolution into a document image, which contains geometric distortion imported by the printing process; on the other hand, convert the Word document page into PDF file, and then save it as Tiff format's image at the same resolution, which is considered as the ideal image.

The pre-processing process is applied to both ideal and document image, and the feature point set is gotten respectively. The parameters can be estimated from the feature point set of ideal and document image by the least squares method. Part of the parameters is used by the SVM classification for printer forensics. And error pattern can be obtained from the ideal and rectified document image.

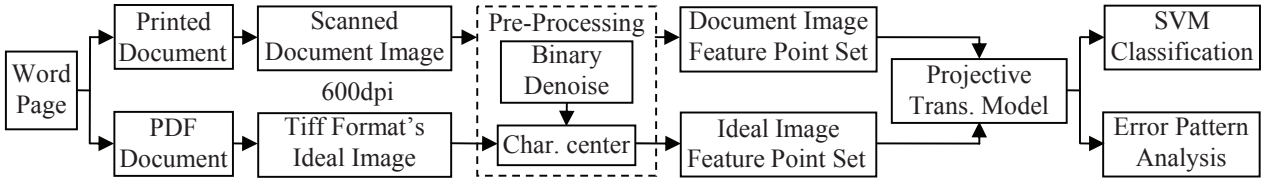

Fig.1. Block diagram of printer forensics method based on page document's geometric distortion

3. THE ESTIMATION OF PROJECTIVE TRANSFORMATION MODEL

In this section, a printed document needs pre-processing process to get individual character and the feature point set. Then built a projective transformation from the feature point set and get the parameters by least squares method.

3.1. Pre-processing

The purpose of pre-processing is to get the coordinates of the center of every character in ideal and document image and the feature point set.

Gray-level document image needs to be converted into binary image based on threshold, and those connected components in the binary image are considered as satellite drops and eliminated whose area is less than a certain number of pixels. The ideal image itself is binary without noise.

The denoised binary image gotten above is then segmented to extract character images. Because the document image is composed of Chinese characters row by row and the skew of the document image can be ignored. Therefore, row projection approach is used to segment each row of characters. For each row, take the column projection approach to segment each character.

The boundary and center can be extracted during the character segmentation, as the example shown in Figure 2. Figure 2(a) shows one character image in document image, Figure 2(b) shows its denoised binary image segmented from whole document image with red center and blue boundary marker, and Figure 2(c) from the ideal image at corresponding location. The centers of the two character images are matched and collected into the feature point set.

Fig.2 (a) Fig.2 (b) Fig.2 (c)

Fig.2 (a) a character in document image, Fig.2 (b) its binary image with center and boundary marker, and Fig.2 (c) the character in ideal image at corresponding location

3.2. Projective transformation Model

The projective transformation model is established to describe page document's geometric distortion.

The form of two-dimensional projective transformation mapping coordinates of point (x_1, y_1) to coordinates of point (x_2, y_2) is described in equation (1):

$$\begin{cases} x_2 = \dfrac{m_0 x_1 + m_1 y_1 + m_2}{m_6 x_1 + m_7 y_1 + 1} + e_x \\ y_2 = \dfrac{m_3 x_1 + m_4 y_1 + m_5}{m_6 x_1 + m_7 y_1 + 1} + e_y \end{cases} \tag{1}$$

Where, m_0, m_4 are the scaling coefficients, m_2, m_5 are the translation coefficients, m_6, m_7 are the coefficients which represent the degree of parallel lines distorted into intersecting lines in x and y direction respectively, m_1, m_3 are the rotation coefficients, and e_x, e_y are the residual error in x and y direction respectively. Assume that the noise $(e_x, e_y)^{\mathrm{T}}$ is Gaussian distributed $N(0, \sigma^2 \boldsymbol{I})$, then the least squares solution is also the maximum likelihood estimate.

Equation (1) is adjusted as follows [7]:

$$\begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} x_1 & y_1 & 1 & 0 & 0 & 0 & -x_1x_2 & -y_1x_2 \\ 0 & 0 & 0 & x_1 & y_1 & 1 & -x_1y_2 & -y_1y_2 \end{bmatrix} \boldsymbol{M} + \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \quad (2)$$

Where, $\boldsymbol{M} = \begin{bmatrix} m_0 & m_1 & m_2 & m_3 & m_4 & m_5 & m_6 & m_7 \end{bmatrix}^{\mathrm{T}}$. Assume that the noise $(\alpha, \beta)^{\mathrm{T}}$ follows Gaussian distribution.

The feature point set contains center of each character, and all the feature points form an over-determined equation. Assume that there are n feature points, and then the over-determined equation can be expressed as follows [7]:

$$\begin{bmatrix} x_{21} \\ y_{21} \\ \cdots \\ x_{2n} \\ y_{2n} \end{bmatrix} = \begin{bmatrix} x_{11} & y_{11} & 1 & 0 & 0 & 0 & -x_{11}x_{21} & -y_{11}x_{21} \\ 0 & 0 & 0 & x_{11} & y_{11} & 1 & -x_{11}y_{21} & -y_{11}y_{21} \\ \cdots\cdots & \cdots\cdots\cdots & & \cdots & & & \cdots & \\ x_{1n} & y_{1n} & 1 & 0 & 0 & 0 & -x_{1n}x_{2n} & -y_{1n}x_{2n} \\ 0 & 0 & 0 & x_{1n} & y_{1n} & 1 & -x_{1n}y_{2n} & -y_{1n}y_{2n} \end{bmatrix} \boldsymbol{M} + \begin{bmatrix} \alpha_1 \\ \beta_1 \\ \cdots \\ \alpha_n \\ \beta_n \end{bmatrix} \quad (3)$$

Expressed as the form of compact matrix:

$$\boldsymbol{b} = \boldsymbol{A}\boldsymbol{M} + \boldsymbol{\varphi} \quad (4)$$

Where, the size of $\boldsymbol{b}$ is $2n \times 1$, of $\boldsymbol{A}$ is $2n \times 8$, and of $\boldsymbol{\varphi}$ is $2n \times 1$.

3.3. The solution of the over-determined equation

The over-determined equation can be solved in a least squares sense.

$$\hat{\boldsymbol{M}} = \arg\min_{\boldsymbol{M}} \left\| \boldsymbol{A}\boldsymbol{M} - \boldsymbol{b} \right\| \quad (5)$$

Where, $\hat{\boldsymbol{M}} = \begin{bmatrix} \hat{m}_0 & \hat{m}_1 & \hat{m}_2 & \hat{m}_3 & \hat{m}_4 & \hat{m}_5 & \hat{m}_6 & \hat{m}_7 \end{bmatrix}^{\mathrm{T}}$.

The method of singular value decomposition (SVD) [8] is applied to solve the least squares problem. The solution can be obtained by equation (6).

$$\hat{\boldsymbol{M}} = \boldsymbol{A}^{+}\boldsymbol{b} = \boldsymbol{V}\boldsymbol{\Sigma}^{+}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{b} \quad (6)$$

Where, $\boldsymbol{A}^{+}$ is the pseudo-inverse of $\boldsymbol{A}$, and $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}$ is the SVD of $\boldsymbol{A}$.

The residual error vector sequence $\hat{e}_i = \hat{e}_{xi} + j\hat{e}_{yi}$, $i = 1, \cdots, n$ can be calculated by equation (7).

$$\begin{cases} \hat{e}_{xi} = \dfrac{\hat{m}_0 x_{1i} + \hat{m}_1 y_{1i} + \hat{m}_2}{\hat{m}_6 x_{1i} + \hat{m}_7 y_{1i} + 1} - x_{2i} \\ \hat{e}_{yi} = \dfrac{\hat{m}_3 x_{1i} + \hat{m}_4 y_{1i} + \hat{m}_5}{\hat{m}_6 x_{1i} + \hat{m}_7 y_{1i} + 1} - y_{2i} \end{cases} \quad (7)$$

In the experiment, as a result of poor printing quality, some characters have greater detection center deviation, which causes outliers in the sequence $|\hat{e}_i|$. PaЙTa method selected as the outlier removing criteria is as follows:

For a sequence data $|\hat{e}_i|$, calculate arithmetic average $\bar{e} = (1/k)\sum_{i=1}^{k} |\hat{e}_i|$, residual error $v_i = |\hat{e}_i| - \bar{e}$, and the root mean square deviation $\sigma = \left(\sum v_i^2 / k - 1 \right)^{1/2}$. Discrimination based on the following criteria:

If $\left| |\hat{e}_i| - \bar{e} \right| > 3\sigma$, then $|\hat{e}_i|$ is considered as gross error and should be discarded; if $\left| |\hat{e}_i| - \bar{e} \right| \le 3\sigma$, then $|\hat{e}_i|$ is considered as normal data and should be retained.

If $|\hat{e}_i|$ is detected as the outlier, the corresponding pair of feature point should be removed from the ideal and document image feature point set. Then the parameter $\hat{\boldsymbol{M}}$ of the model is re-calculated. Repeat the process until that the feature point set no longer has outliers.

3.4. The proposed printer forensics method

The main steps of proposed printer forensics method summarize as follows: firstly, apply pre-processing process to the ideal and document image to get feature point set; secondly, establish the projective transformation model describing page document's geometric distortion; thirdly, employ SVD and PaЙTa method to solve the equation and get the model's parameters; finally, SVM classification [9] is used for printer forensics.

Considering that translation and rotation are inevitable during the process of printing and scanning, so the parameters m_1, m_2, m_3, m_5 can not represent the intrinsic features of printer; But scaling and the distortion that the parallel lines degrades into intersecting lines are stable, so the four parameters m_0, m_4, m_6, m_7 are selected to represent the intrinsic features of printer.

The selected features are classified by SVM classify-cation method for printer forensics. And the sequence of residual error vectors can be formed as matrix according to the location of their corresponding characters in the document page, which is called error pattern.

4. EXPERIMENTAL RESULTS

The printer models in the experiment are shown in Table 1. A total of 10 different printers composed of 5 kinds of models are selected. Each printer individual is assigned a label, as shown in Table 1. For example, label "03" represents a printer individual whose model is "Hp 1000".

Table 1. List of printer Models

Printer Model	Label
Hp 1000	01, 03, 04, 06
Hp 1020	05, 09
Hp 1320n	02, 10
Lenovo 2312P	07
Sumsang ML 1510	08

12 pages are printed from each printer, so there are total 120 pages printed from 10 printers. Each page contains neatly arranged 1496 Chinese characters (34 columns × 44 rows), which are randomly selected from the commonly used Chinese characters. The font is "small IV Song".

Each page is scanned by "Epson Pefection 1200" scanner at 600 dpi resolution and gray photo type.

1496 feature points can be extracted from each page document image, and matched with that from its ideal image. The document images are processed by the method proposed to get a vector of model parameters and an error pattern.

The distribution of parameters m_4 and m_7 extracted from document images of 10 printers is shown in Figure 3. As can be seen, the clustering and the inter-class separability are obvious. Parameter m_7 is close to 0, and m_4 close to 1, which demonstrates that the geometric distortion is minimal.

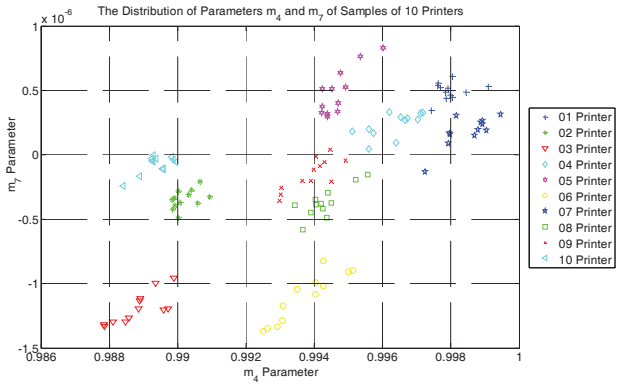

Fig.3. Distribution of m_4 and m_7 of samples from 10 printers

The 12 document images of each printer are divided into two parts equally, the first part training the SVM classifier, and second part testing. The test result shows that the correct classification rate of all the 10 printers is 100%.

Figure 4 shows two typical error patterns of page sampled from 01 and 05 printer respectively. Among them, blue arrow denotes the residual error vector between centers of character images in the ideal image and those in the rectified document image, and the contour with colormap "hot" represents the magnitude of the residual error vector.

The error pattern of 03 page from 01 printer has obvious vertical strips near twenty-second column as shown in Figure 4(a), and all patterns of 12 pages from 01 printer have that feature. But the horizontal strip appears near eleventh and thirty-eighth rows in the error pattern of 03 page from 05 printer as shown in Figure 4(b). Experiments show that comparing and analyzing the difference of the error patterns can help determine the source of document.

In the experiments, the effectiveness of the selected parameters feature is proved. At the same time, the error pattern also can assist finding the document's source printer.

5. CONCLUSIONS

The page document's geometric distortion assuredly exists, which can be modeled by projective transformation suitably. Some parameters of the model can be used for printer forensics. Future work will be focused on comparing the error pattern automatically for printer forensics and finding more suitable model. Thanks to the support of National 863 project of China (2008AA01Z418) for this paper.

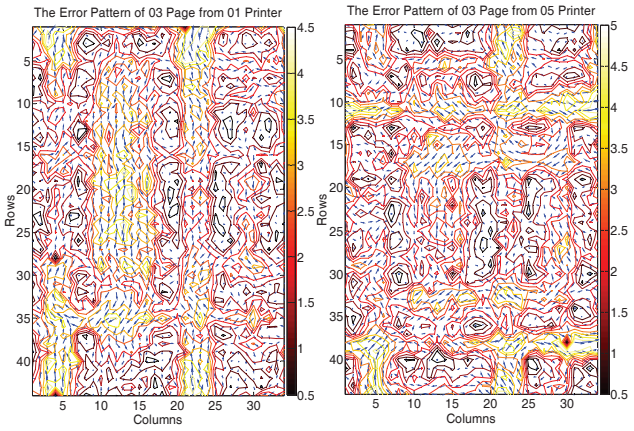

Fig.4. Error pattern of (a) 03 page from 01 printer and (b) 03 page from 05 printer. Blue arrow denotes the residual error vector, and contour with colormap "hot" denotes its magnitude

6. REFERENCES

[1] John Oliver, Joyce Chen. Use of signature analysis to discriminate digital printing technologies. In: IS&T's NIP18: 2002 International Conference on Digital Printing Technologies.

[2] Mikkilineni, A. K., Chiang, P.J., et al.: Printer identification based on graylevel co-occurrence features for security and forensic applications. In International Conference on Security, Steganography, and Watermarking of Multimedia, 2005.

[3] V. Talbot, P. Perrot, and C. Murie. Inkjet printing discrimination based on invariant moments. In International Conference on Digital Printing Technologies, 2006.

[4] Shen Linjie, Kong Xiangwei, You Xin'gang. Printer forensics based on character image quality measures. In Journal of Southeast University(Natural Science Edition), China, 2007, 37(1):92-95.

[5] Christian Schulze, Marco Schreyer, Armin Stahl, and Thomas Breuel. Evaluation of Graylevel-Features for Printing Technique Classification in High-Throughput Document Management Systems. 2008.

[6] Eric Kee, Hany Farid. Printer Profiling for Forensics and Ballistics. In Proceedings of the 10th ACM workshop on Multimedia and security, 2008.

[7] Tapas Kanungo. Document Degradation Models and a Methodology for Degradation Model Validation[D]. [PhD Thesis]. University of Washington, 1996.

[8] Gilbert S. Linear algebra and its applications. New York: Academic Press, 1976. 251-317.

[9] C.-W. Hsu, C.-C. Chang, C.-J. Lin. A Practical Guide to Support Vector Classification. http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf, 2005.