Laser Printer Source Forensics for Arbitrary Chinese Characters

Xiangwei Kong[1], Xin'gang You[1,2], Bo Wang[1], Shize Shang[1] and Linjie Shen[1]
[1]Information Security Research Center, Dalian University of Technology, Dalian, Liaoning, China
[2] Beijing Institute of Electronic Technology and Application, Beijing, China

Abstract - *Identifying the source of the printed documents and tracing the printed documents efficiently is a major challenge in lossless document inspecting. As many methods are not independent of the character content, in this paper, a new printer identification method based on the intrinsic printing features of arbitrary Chinese character is proposed. The distinguishing features include the 24-D wavelet and 7-D noise statistical features which denote the intrinsic printing properties can be extracted from arbitrary Chinese character. It is especially useful when there are few characters in the inspected document The results of experiment showed that the accuracy of character identification is higher than other related prior methods and the printers of the same brand and model can be identified effectively.*

Keywords: Laser printer forensics, Chinese character, intrinsic printing features, document inspection

1 Introduction

Laser printer is widely used in office by governments, companies and individuals. Printed documents are often presented as court evidence, which comes with an important issue concerning document inspecting. It is important to identify the source of the printed document and to trace the printing device.

There are mainly three ways for printer forensics using common electronic equipments such as computers and scanners. One is put forward by Electronic Frontier Foundation(EFF) US using encrypted yellow dots watermark contained in documents printed by some laser printers as forensic evidence[1]. But the announced brands and models of printers are very limited. The other is active forensic method by making use of digital watermarks which requires pre-embedded watermarks into the documents. Another is passive forensic technology only based on printed documents. One of the passive methods is non-destructive to the printed documents and it only needs a scanner and a following identification algorithm for printer forensics. The research group led by Professor J. Allebach and E.Delp in Purdue University has made great progress[2,3,4]. They extracted 1-D signal of the printing direction from characters, reduced feature dimensions using PCA and classified printers by mixed Gaussian model and tree classification. Mikkilineni et al. [2] extracted 22-D graylevel co-occurrence features from each printed letter "e", then used 5-NN as classifier. 9 out of 10 printers of different models are correctly identified. The method in [4] made some improvements and used SVM as classifier. The identifying accuracy of 10 printers achieved 100%, and the character classification accuracy was improved to 93.2%. Farid et al. [5] used PCA to model the degradation of a document caused by printing, and the resulting printer profile was then used to detect the source.

The methods in [2,3,4] can identify the right printer model only based on specific English letter "e". The result is better, the more letter "e". While the classification accuracy will decrease significantly when there are few letters "e" in the testing printed document.

In real cases, the characters in a printed file usually exist in few lines including not enough special characters. We want to find effective forensics method for arbitrary character so as to better forensics work in only a few characters.

In this paper, we proposed a laser printer forensic method based on intrinsic printing features of arbitrary Chinese character instead of a specific English letter. The results of experiment showed that the method can also identify laser printers of different brands, different models of the same brand and even the same brand and model. Our method is independent of character content and special character, which makes it especially useful when there are few characters or no trained specific character in the testing document.

The rest of this paper is organized as follows. Section 2 gives a brief introduction to the intrinsic property of laser printers and the framework of the proposed printer forensics method. Section 3 describes extracted features used in this method. Section 4 introduces the classifier design and Section 5 reports experimental results. Finally, Section 6 gives the conclusion.

2 Intrinsic property of laser printers and printer forensics method framework

In this section, we first discussed the plausibility of printer forensics using the intrinsic features and introduced the method framework, then explained the training and testing samples used in the experiment.

2.1 Intrinsic property of printers

Different printer manufacturers have different printing processes and use different hardware components, which result in distinguishing intrinsic property of printers.

On one hand, every printer manufacturer has unique software processing technology, such as Resolution Enhancing technology(REt) owned by HP, which produces printing quality differences among printers of different brands and even of different models from the same brand.

On the other hand, hardware components of printers also play an important part in the printing quality. An inevitable "Banding" artifact appears during the printing process, which is caused by the non-uniform scan line spacing due to variations in the optical photoconductor (OPC) drum velocity. Although a lot of research has been done to reduce it, the uniqueness of banding for every printer can be used in printer forensics.

2.2 Framework of the proposed printer forensics method

Figure 1 illustrates the framework of the proposed printer forensic method for arbitrary Chinese character which mainly involves in training and testing. Our framework is composed by the following procedures:

Step 1: Scan the training and testing printed documents and save them in computer as gray images in BMP format.

Step 2: Apply adaptive binary process to the document images, divide up single character from the images, and then extract features from each character image.

Step 3: Send the features of the training samples to the SVM classifier and obtain the optimal parameter model.

Step 4: Send the features of the testing samples to the SVM classifier and use the trained model to classify.

Step 5: Identify the correct printer of the testing documents by voting decision mechanism.

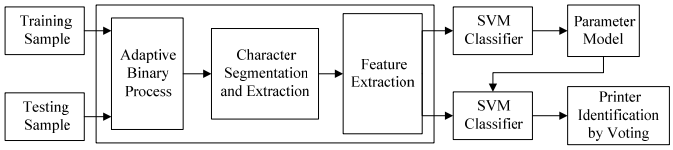

Fig.1. Framework of printer forensics method

2.3 Training and testing samples

We focused on the identification of limited Chinese characters. This is the biggest difference from other images printer forensic methods. Although the number of Chinese characters is finite, it's not necessary to use all of them as training samples. Therefore, we choose 3375 most frequently used Chinese characters provided by National Standard Coding GB-2312. It is reasonable for their usage rate is more than 99%. Testing samples are page documents including about 1300 Chinese characters chosen from frequently used Chinese characters randomly in experiments.

3 Feature extraction

We extract wavelet features and "banding" noise features to describe the difference of printing processing and the unique "banding" noise caused by hardware components in this section.

3.1 Wavelet features of character images

The printing features are only accessible from the printed area, but the printed area of a character is very limited, and the location and size of different Chinese characters' local texture are different. The wavelet transform can perform local analysis to images, and it involves in multi-scale image analysis, which makes it suitable for local texture analysis of character images.

In order to extract effective wavelet features, firstly the character image is decomposed using wavelet transform, and then features are extracted from the transformed image. In this paper, db8 wavelet is selected to perform 2-level wavelet decomposition. For the higher the decomposition level is, the more of printing attributes will lose, level 2 is suitable according to experiments. Assuming the character image is $I(i,j)$, 7 sub-images can be achieved from a 2-level wavelet transform. Among them, $D^{(1)}_{2^{-2}}f$, $D^{(2)}_{2^{-2}}f$, $D^{(3)}_{2^{-2}}f$, $D^{(1)}_{2^{-1}}f$, $D^{(2)}_{2^{-1}}f$ and $D^{(3)}_{2^{-1}}f$ are called detailed sub-images. $A_{2^{-2}}f$ is the low-frequency sub-image which includes most of the basic information. That information is greatly affected by different characters content, and has negative impact on printers' identification, so we excluded $A_{2^{-2}}f$ and only extracted statistical features from the other 6 detailed sub-images which contain many texture details.

One of the features is the mean value of each sub-image's wavelet coefficients defined as:

$$m = \frac{1}{N}\sum_{(i,j)\in R} v(i,j) \tag{1}$$

Where R denotes the printing area of one sub-image, N is the number of pixels in R, $v(i,j)$ is the pixel value at (i,j) in the sub-image.

The following three features are the standard variance, skewness and kurtosis of each sub-image's wavelet coefficients respectively:

$$\sigma = \sqrt{E\left(v(i,j)-m\right)^2} \tag{2}$$

$$s = \frac{E\left(v(i,j)-m\right)^3}{\sigma^3} \tag{3}$$

$$k = \frac{E\left(v(i,j) - m\right)^4}{\sigma^4} \tag{4}$$

Thus, there are 4 features for each sub-image, yielding a total number of $4 \times 6 = 24$ wavelet coefficient statistics.

3.2 Noise features of character images

During the printing process of texts, noise could be brought into the character image by hardware components. Besides the noise introduced by characters' edge complexity, "banding" noise is the most obvious. The "banding" noise in the local areas can be considered as stochastic. The next problem is how to get the noise images and noise features of character images. We design a method to get the noise image from original image in Fig.2. During the general forensics processing, we just only get original image, then use Gaussian filter as filtering process carried on character images to obtain ideal estimated character image. The difference between them is the noise image. In order to extract noise features which are independent of character content, we choose three statistical features of the noise image as follows. The noise feature extraction process is shown in Fig.2

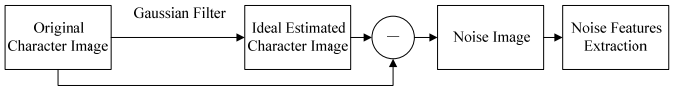

Fig.2. Noise feature extraction process

Let $I(i,j)$ denote the original image, $\hat{I}(i,j)$ denotes the Gaussian filtered image. M, N is the width and height of the image respectively. $M_i, i = 1, \cdots, 7$ denote 7 noise image features.

A. Minkowsky Measures: including mean absolute error, mean square error

$$M_\gamma = \left(\frac{1}{M \times N} \sum_{j=1}^{M} \sum_{i=1}^{N} \left| I(i,j) - \hat{I}(i,j) \right|^\gamma \right)^{1/\gamma} \tag{5}$$

M_1 means absolute error for γ=1, and M_2 means square error for γ=2.

B. Correlation Measures: including Czekanowski distance, Image Fidelity, Normalized Cross-Correlation which are defined respectively as follows:

$$M_3 = \frac{1}{M \times N} \sum_{j=1}^{M} \sum_{i=1}^{N} \left(1 - \frac{2 \times \min\left(I(i,j), \hat{I}(i,j) \right)}{I(i,j) + \hat{I}(i,j)} \right) \tag{6}$$

$$M_4 = 1 - \frac{\sum_{j=1}^{M} \sum_{i=1}^{N} \left(I(i,j) - \hat{I}(i,j) \right)}{\sum_{j=1}^{M} \sum_{i=1}^{N} I(i,j)^2} \tag{7}$$

$$M_5 = \frac{\sum_{j=1}^{M} \sum_{i=1}^{N} I(i,j) \hat{I}(i,j)}{\sum_{j=1}^{M} \sum_{i=1}^{N} I(i,j)^2} \tag{8}$$

C. Spectral Measures: including magnitude distortion and phase distortion measure which are defined as follows:

$$M_6 = \sum_{u=1}^{M} \sum_{v=1}^{N} \left| \left| \Gamma(u,v) \right| - \left| \hat{\Gamma}(u,v) \right| \right|^2 \tag{9}$$

$$M_7 = \sum_{u=1}^{M} \sum_{v=1}^{N} \left| \left| angle\left(\Gamma(u,v) \right) \right| - \left| angle\left(\hat{\Gamma}(u,v) \right) \right| \right|^2 \tag{10}$$

Where $\Gamma(u,v)$ and $\hat{\Gamma}(u,v)$ denote the Discrete Fourier Transform(DFT) of image $I(i,j)$ and image $\hat{I}(i,j)$ respectively. $angle(\bullet)$ is the angle calculation function.

4 Classifier design using SVM

Support Vector Machine(SVM) is a machine learning method based on statistical learning theory, which has been widely used in pattern recognition and artificial intelligence and proved to be useful tool for small sample classification. Considering limited training samples and the advantages of SVM, SVM is selected as the classifier for the classification of characters from different printers.

In our experiments, C-support vector classification with the non-linear RBF kernel is used[7]. RBF kernel is defined as:

$$K\left(x_i, x\right) = \exp\left(-\gamma \left\| x_i - x \right\|^2 \right) \tag{11}$$

where the appropriate parameter pair (C, γ) can be obtained by grid searching. The searching range for C is $\left\{ 2^{-5}, 2^{-4}, \cdots, 2^5 \right\}$, and $\left\{ 2^{-5}, 2^{-4}, \cdots, 2^3 \right\}$ for γ.

5 Experimental results

5.1 Experimental setup

In our experiments, we used 5 laser printers from 3 brands with higher market share which are HP, Epson and

Canon. In order to investigate the effectiveness of our approach to different brands, models and using time, two different Canon models and two printers of different using time from the same model are selected. Table 1 lists the parameters of these laser printers:

Table 1 Laser printers used in experiments

Printer Brand	HP	Epson	Canon	Canon	Canon
Printer Model	5500	C7000	5000	8500(1)	8500(2)
Label	1	2	3	4	5

5.2 Experimental results for a specific character

To evaluate the proposed method, we firstly performed experiments on specific Chinese character "的" which is used frequently in Chinese files. One page full of character "的" printed by each printer is used as the training sample, and another same page as test sample. Following the steps in Section 2 and using the features mentioned above, the results of the proposed printer forensics method can be obtained as shown in Table 2.

Table 2 Experimental results of the proposed method for specific character "的"

Printers	1	2	3	4	5	Average
Accuracy(%)	99.93	92.45	79.95	91.16	92.51	91.20
Mis-classification(%)	0.07	7.55	20.05	8.84	7.49	8.80

5.3 Experimental results for arbitrary character

The experimental results of our proposed printer forensic method for arbitrary Chinese character are shown in Table 4. The average classification accuracy achieved 86.42%. Compared to it, the accuracy of the method proposed in [2] is only 62.31%, which means the graylevel co-occurrence method doesn't apply to printer forensics for arbitrary Chinese character.

However, for printer forensics, we are more concerned about which printer the testing document comes from, that is, the identification result of the page is the ultimate question instead of a single character. Assuming that only when more than 50% of characters in the testing page are correctly classified, the identification result of the page is considered to be correct. Of course the assumption is based on that there are enough characters in the page, thus the final identification result is convincible. All of the testing sample pages are correctly classified in this experiment.

As shown in Table 4, the proportion of confused characters between two printers is only 3.69%. Therefore we can conclude that the proposed features in the paper are distinguishing for each printer in a certain period, even for two printers of the same brand and model.

Table 4 Experimental results of the proposed method for arbitrary Chinese character

Input/ Output	1	2	3	4	5	Identificatin results
1	88.54	3.45	4.84	1.73	1.43	correct
2	0.56	84.19	6.80	3.35	5.10	correct
3	0.62	9.39	85.08	2.77	2.14	correct
4	0.91	6.99	5.27	85.89	0.94	correct
5	0.72	5.75	2.38	2.75	88.40	correct

6 Conclusion

A laser printer forensics method using intrinsic printing features of arbitrary Chinese character is proposed in this paper. Based on the software processing technology printers and hardware property of laser printers, we extracted statistical features which are independent of character content, and solved the low classification accuracy problem existing in [2,3,4] when there are few characters or no specific characters like "的" used in training in the testing document. Experiment results show that our proposed method is not only effective for printers of different brands and models, but also for two printers of the same brand and model.

However, it is only possible to be classified correctly with the pre-knowledge that the testing documents come from the training printer sets. Otherwise it will be mis-classified to one of the training printer set. More reasonable identification system is needed to solve this problem. Therefore, future work on the printing process is needed for the printer forensics method to be more practical and effective.

7 Acknowledgments

This work was supported by the National High Technology Research and Development Program of China (863 Program, No. 2008AA01Z418) and National Natural Science Foundation of China (No. 60971095).

8 References

[1] http://www.eff.org/issues/printers

[2] Mikkilineni A K, Chiang P J, Ali GN, et al. "Printer Identification Based on Graylevel Co-occurrence Features for Security and Forensic Applications". The SPIE International Conference on Security, Steganography, and Watermarking of Multimedia Contents VII. San Jose, CA, pp.430–440, 2005.

[3] Khanna1 N, Mikkilineni A K, George T.-C. Chiu, et al. "Survey of Scanner and Printer Forensics at Purdue University". The 2nd international workshop on Computational Forensics. Washington, DC, USA, pp.22–34, Aug 2008.

[4] Mikkilineni A K, Arslan O, Chiang P J et al. R. “Printer Forensics Using SVM Techniques”. International Conference on Digital Printing Technologies. Baltimore, MD, pp.223–226, Sep 2005.

[5] Eric Kee, Hany Farid. “Printer Profiling for Forensics and Ballistic”. The 10th ACM workshop on Multimedia and security, Oxford, United Kingdom, pp.3-10, Sep 2008.

[6] Avcibas I, Memon N, Sankur B. “Steganalysis Using Image Quality Metrics”. IEEE transactions on Image Processing, vol. 12, pp.221-229, Feb 2003.

[7] C.-C. Chang, C.-J. Lin. LIBSVM: a library for support vector machines. http://www.csie.ntu.edu.tw /~cjlin. 2007,6.