A Novel Multi-size Block Benford's Law Scheme for Printer Identification

Weina Jiang[1], Anthony T.S. Ho[1], Helen Treharne[1], and Yun Q. Shi[2]

[1] Dept. of Computing,
University of Surrey Guildford,
GU2 7XH, UK
w.jiang@surrey.ac.uk
[2] Dept. of Electrical and Computer Engineering
New Jersey Institute of Technology
Newwark, NJ 07102, USA

Abstract. Identifying the originating device for a given media, i.e. the type, brand, model and other characteristics of the device, is currently one of the important fields of digital forensics. This paper proposes a forensic technique based on the Benford's law to identify the printer's brand and model from the printed-and-scanned images at which the first digit probability distribution of multi-size block DCT coefficients are extracted that constitutes a feature vector as the input to support vector machine (SVM) classifier. The proposed technique is different from the traditional use of noise feature patterns appeared in the literature. It uses as few as nine numbers of forensic features representing each printer by leveraging properties of the Benford's law for printer identification. Experiments conducted over electrophotographic (EP) printers and deskjet printers achieve an average of 96.0% classification rate of identification for five distinct printer brands and an average of 94.0% classification rate for six diverse printer models out of those five brands.

Keywords: Digital forensics, printer identification, multi-size block based DCT coefficients, Benford's law, composite signature.

1 Introduction

Printed documents are used to disseminate important and valuable information for both personal and business usage. However, the tremendous growth of the digital era has also led to the ease of forgery of not only digital content but also printed documents of the digital data. Printed documents include legal contract, ID card, bank check etc. and its security concerns have been addressed by many researchers [1][2]. In addition to the security techniques, forensic tools can provide valuable information to law enforcement and forensic experts. Several forensic analysis techniques have been developed to identify the printer used for producing a given document [3][4][5]. The identification techniques make use of forensic characterization produced by devices, which is called *device signature* [6].

G. Qiu et al. (Eds.): PCM 2010, Part I, LNCS 6297, pp. 643–652, 2010.

The device signature is uniquely produced by sensors in printers, cameras, and scanners.

Two classes of device signatures have been investigated in the literature. Intrinsic signature means the noise feature which can be the artifacts due to optical, electrical, or mechanical distortion induced by the devices. In printer identification, banding [4] can be a good example for intrinsic signature detected in printed documents. Extrinsic signature can include features that tied to the modulation process by specific patterns such as halftoning, or watermarking that encoded with the device. In printed documents, these noise features are hardly isolated and also they may exhibit geometric distortion. The intrinsic features and the extrinsic features may compromise each other. In this paper, we define both intrinsic and extrinsic features as *composite signature*. In [7], geometric distortion signatures represent the composite feature incurred both by halftoning and eletrophotographic (EP) printer distortion. Such geometric distortion signatures do exhibit a high correlation with corresponding printer signatures and a low correlation with other printer signatures.

Although the intrinsic signature based on Photo-Response Non-Uniformity (PRNU) by Lukas et al. [8] and Chen et al. [9] established a good model for forensics, it does not provide an accurate detection rate to device identification such as cameras identification [10][1] achieving approximately 90.8% classification rate and printers identification [5][4] achieving approximately 90.0% classification rate. The drawback of PRNU requires synchronization [8] or registration [7] of the scanned sample images, which could cause inaccuracy due to non-linearity of the distortion by printers or cameras.

To improve the classification rate, we investigate alternative statistical tools which can provide robust features without complicated image pre-processing for printer model/brand identification. In this paper, we propose to apply Benford's law based on first digit statistics to multi-size block DCT coefficients for printer identification. The use of generalized Benford's law has already been applied to block-DCT coefficients for image processing [11] and for JPEG compression rate detection in [12].

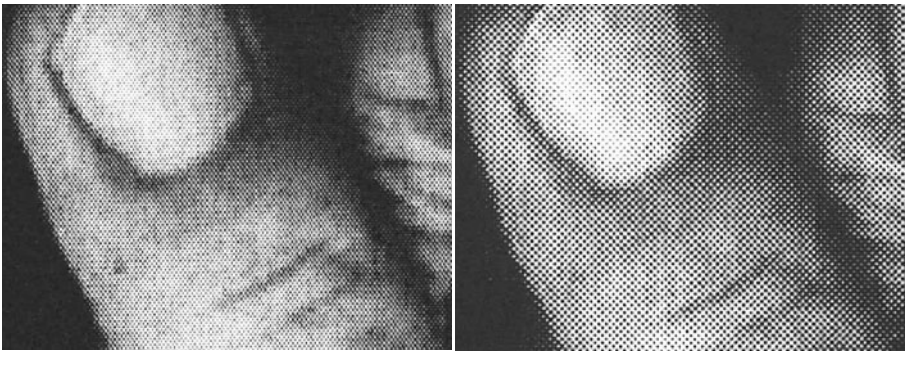
The contributions of this paper are: (1) Develop a multi-size block based DCT coefficients Benford's Law for forensic features extracted from printed documents. These features are used for printer identification. 2) *Composite signature* considers both the impacts of halftoning and printing distortion. 3) Use of support vector machine (SVM) classifier to identify the brand and model of printers. Our scheme achieves an average of 96.0% classification rate of identification for five distinct printer brands and an average of 94.0% classification rate for six diverse printer models out of those five brands.

The rest of the paper is organized as below. In Section 2, forensic characterizations are addressed and designed for printer identification; Section 3 proposes our approach on probability distribution of the first digit of multi-size block based DCT coefficients as feature vectors. Section 4 illustrates experiment settings and an application of SVM for printer model and brand identifications. Section 5 provides a conclusion and discussion of future work.

2 Design of Forensic Characterization for Printer Identification

2.1 Printer Principle and Halftoning

Most printers undergo a halftoning process before an image is physically printed. This process digitalizes a grayscale or color image into perceptually good quality of binary patterns. Common halftoning algorithms are error diffusion [13], clustered dot halftone etc. [14]. For example, cluster dot halftoning is often used in EP printers to generate periodically different levels of gray spot patterns illustrated in Figure 1a and Figure 1b. These figures show halftone patterns of a finger image printed on HP laserjet 4250 and Lexmark X646e printers respectively. The differences of halftone dot size and gray levels produced by these two brand printers can be clearly seen in the two figures.



(a) HP Laserjet 4250 (b) Lexmark X646e

Fig. 1. Output of Two Different Brand Printers

However, it is not only the different halftone patterns introduced by different printers but other numerous geometric distortions can also impact the visual dots introduced by EP or Inkjet printers. Bulan et al. determined that the patterned visual dots printed on the paper was mainly due to the variations in laser scanning speed over a scanline, and to the velocity of the Optical Photo-Conductor (OPC) drum causing non-uniform spacing between raster lines [7]. These mechanical distortions can also be manifested as banding in the printers regarded as a feature being used for forensics by Mikkilineni et al. [1].

2.2 Designing Features for Forensics in Printer Identifications

Pattern noise features have demonstrated to be a good forensic characterization for printer identifications [1]. However, the challenge is to separate the fixed component from the random component of the noise. In Khanna et al.'s approach [1], by averaging the sampled images, they reduced the random component while

enhancing fixed noise parts. Their approach also required the sample images with pixel-wise alignment so that the averaging of noise additions would not be misplaced. The image alignment process has shown to be not trivial. Moreover, the noise feature vectors were composed of ten dimensional features associated with each scanned image giving rise to increased computational requirements to SVM calculation [1].

Buluan's forensic features [7] consisted of a collection of distortion displacement vectors over the halftone dots. Their features required the scanned image rotation compensation caused by printers, which is also difficult if the distortion is a non-linear transformation.

In this paper, we identified that a good forensic feature should have the following properties:

- Independent to the image content.
- Robustness of the feature vector means the random noise does not impact the forensic characterization significantly.
- Composite signature means feature vector should reflect forensic characterization of intrinsic signature and extrinsic signature.
- Efficiency with moderate numbers of feature vectors.

Gonzelez et al. [11] found that the first digit probability distribution of DCT coefficients of an image followed closely to the Benford's law. Their generalized Benford's law was represented by a Fourier approximation to an empirical digit distribution. Li et al. also developed a generalized Benford's law to detect double JPEG compression [15]. Their research has shown that the Benford's law is independent to the image content, and robust as forensic features.

3 Multi-size Block DCT Coefficients Statistics

The Benford's law, the first digit distribution of d (d $\in$ 1,...,9), follows a logarithmic scale, as shown in Equation 1

$$p(d) = log_{10}(1 + \frac{1}{d}), d = 1, 2, ..., 9 \tag{1}$$

Gonzelez et al. [11] mentioned that Benford's law has the following properties:

> *[Property 2]. Suppose that a random variable X follows Benford's law; then the random variable $Z = \alpha X$ will follow Benford's law for an arbitrary α if and only if X is strong Benford.*
>
> *[Property 3]. Let X follows Benford's law, and Let Y be another random variable independent of X. Then, the random variable $Z = XY$ is strong Benford.*

3.1 Multi-size Block Benford's Law

To test the validity of the Benford's law for printer identification, we assume the impact of the printer's halftoning and geometric distortion on the test images will be either additive or multiplicative noise. This noise will change the local spatial pixel positions or gray levels that would also change the DCT coefficients. However, the change of the first digit probability distribution of a single block DCT coefficients cannot reflect the decorrelation of noise features in test images made by printers, thus the first digit statistics of DCT coefficients with various block sizes is possible to detect the weak noise energy associated with the different parts of the images and in the change of the block DCT coefficients. If X is 8x8 block-DCT coefficients, multi-size block DCT can choose 2-power times size of X, let $\alpha = 0, 1, 2, 3, 4$ as a multiplicative factor.

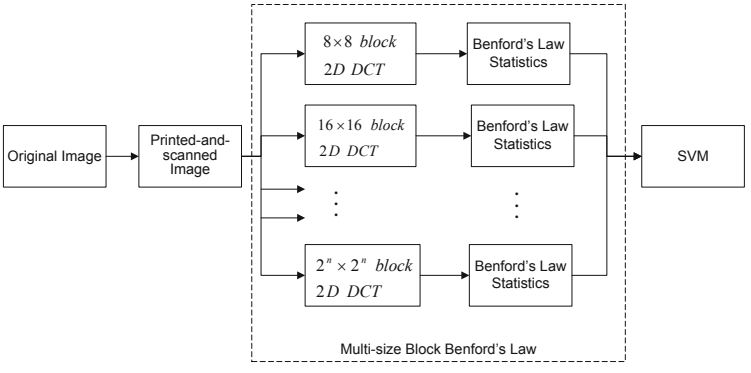


Fig. 2. Multi-size Block DCT Coefficients Statistics Flowchart

The process of statistics extraction of multi-size block DCT coefficients is presented in Figure 2. A test image is printed by one of experimental printers at resolution of 600*600 dpi. and the printout is scanned into digital signal at resolution of 600*600 dpi by Infotec ISC 3535. Multi-size blocks are applied to the scanned printout, where SVM features are extracted from the first digit probability distribution of multi-size block DCT coefficients. For an Image g(i,j) $\forall i, j \in 0, 1, ..., n-1$, its n*n block 2D DCT transform G(u,v) $\forall u, v \in 0, 1, ..., n-1$, is defined as shown in Equation 2 to Equation 4

$$G(u,v) = \frac{2}{n} \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} a(i)a(j)cos(\frac{\pi u(2i+1)}{2n})cos(\frac{\pi v(2j+1)}{2n})g(i,j) \qquad (2)$$

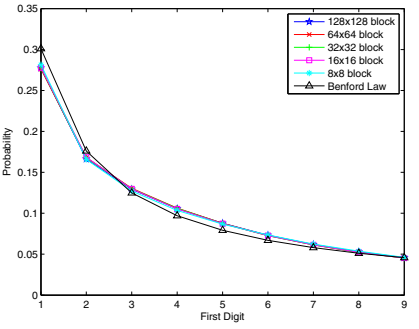
with

$$a(i) = \begin{cases} \frac{1}{\sqrt{2}} & \text{for } i = 0, \\ 1 & \text{for } i > 0 \end{cases} \qquad (3)$$

and

$$a(j) = \begin{cases} \frac{1}{\sqrt{2}} & \text{for } j = 0, \\ 1 & \text{for } j > 0 \end{cases} \qquad (4)$$

(a) A test image printed on Xerox Phaser 4800, and scanned in grayscale image



(b) Multi-size block DCT coefficients distribution vs original Benford's law

Fig. 3. A test image and its multi-size block DCT coefficients statistics

To verify this idea, we printed five copies of a test image on Xerox laser printer as an example, and scanned in grayscale image as shown in Figure 3a. To reduce the random noise impact, white margin is removed at each scanned image because the scanned image size is larger than that of the test image. Each processed image is individually calculated based on the first digit distribution on DCT coefficients with different block sizes $8\times8, 16\times16, 32\times32, 64\times64$, and 128×128. By averaging Benford's law statistics over five copies of the test image to reduce randomness, we found that the results followed closely with the Benford's law distribution as shown in Figure 3b. Therefore, empirical distribution of the multi-size block DCT coefficients can be generalized to Benford's law as indicated in [11][1].

4 Experiments and Result Analysis

4.1 Multi-size Block DCT Coefficients Statistics

For our experiments, we use five high resolution images downloaded from photo.net as illustrated in Figure 4. These images are printed on six model printers in Table 1. Each test image is printed five copies for each printer. The printed images are then scanned into A4 size images with 600 dpi grayscale JPEG format. We use a bounding box to remove the white margin of the scanned images

[1] Remark: Multi-size block DCT coefficients can be regarded as a composite signature discussed in 2.2. Since the influence of halftoning and printing distortion are reflected in the spatial pixels of the test image. This influence will change block-DCT coefficients distribution. However, from a channel model point of view, this influence is a composite noise of halftoning and the printing. The magnified impact of this noise signal is obtained by sampling the test image based on multi-size block DCT coefficients.

(a) (b) (c) (d) (e)

Fig. 4. Test Images

Table 1. Utilized Printers in Experiment

Brand	Model Parameters	DPI
HP deskjet	5940	600
Cannon DeskJet	MPS60	300
HP LaserJet	4250	600
Lexmark Laser	X646e	600
Xerox Phaser	5800	600
Xerox Phaser	4800	600

so that each scanned image can be resized to the same size image of 1024x1024 pixels by cubic interpolation. The resized images are transformed into DCT domains with multi-size blocks and first digit probability statistics are applied to each block DCT coefficients. We assign a label to each printer and calculate first digit distribution as 9 forensic feature vectors for SVM input.

4.2 SVM Classifier

In this paper, LIBSVM [16] is used for implementation of multi-SVM classification, which provides a parameter selection tool using RBF kernel with cross validation via parallel grid search. RBF kernel is selected because of its nonlinearity, less computation complexity with only two hyperparameters C and γ that need be tuned to find the best classification performance. v-fold cross-validation is to divide training set into v subsets of equal size. One subset is predicted by using the classifier trained on remaining $v-1$ subset. In our experiments a five-fold cross validation was used to estimate the accuracy of each parameter combination. In order to find the best C and γ, a search of $C \in [log_{10}(-5)..log_{10}(15)]$ and $\gamma \in [log_{10}(3)..log_{10}(-15)]$ is selected to train the first digit probability as features. The maximum accuracy of SVM is recorded. For $C = 32$ and $\gamma = 0.5$, the maximum classification accuracy for five brand printers can be achieved approximately up to 96% in Figure 5.

For brand identification test, the average classification accuracy is approximately 96.0% for printers which include HP Laserjet 4250, Xerox Phaser 5800,

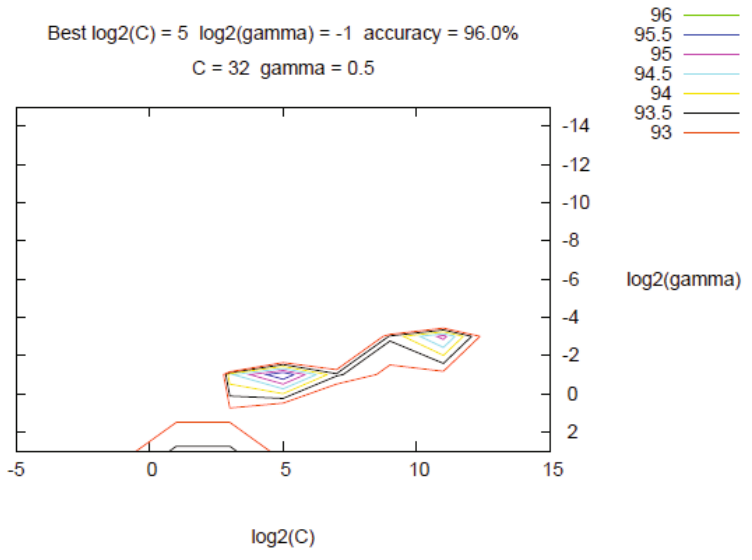


Fig. 5. Five Brands of Printers in SVM model

Table 2. Six Distinct Printer Model Classification Confusion Matrix

		Predict					
	Brand	HP Laserjet 4250	Xerox Phaser 5800	Cannon MSP60	Lexmark X646e	HP deskjet 5940	Xerox Phaser 4800
Train	HP Laserjet 4250	92.0%				4.0%	4.0%
	Xerox Phaser 5800		98.0%	2.0%			
	Cannon MSP60			100%			
	Lexmark X646e	8.0%			92.0%		
	HP Deskjet 5940					100%	
	Xerox Phaser 4800			16.0%			84.0%

Cannon MSP60, Lexmark X646e and HP deskjet 5940. It is assumed that the laser and inkjet printers are considered different. We then apply the present scheme to classify distinct six printer models from thess five printer brands. Table 2 shows the model classification confusion matrix for six different laser and inkjet printer models. The average classification accuracy rate achieved is approximately 94.0% which indicates that our scheme can be used to successfully identify different printer models.

5 Conclusion and Future Work

In this paper, we presented a novel multi-size block Benford's law scheme for identifying laser and inkjet printers. Different from the existing schemes, our forensic feature vectors were composed of 9 dimensional features based on the first digit distribution of multi-size block DCT coefficients statistics. Our multi-size block Benford's Law model achieved good classification accuracy to both printer brands and printer models. The average accuracies were approximately 96.0% and 94.0% respectively.

In our future work, we plan to develop mixed forensic features for printer model identifications in conjunction with noise features in printed documents. It might be a new direction not only providing the high classification rate but also to provide the application such as forgery detection in the printed documents.

References

1. Khanna, N., Mikkilineni, A.K., Chiu, G.T., Allebach, J.P., Delp, E.J.: Survey of scanner and printer forensics at purdue university. In: Srihari, S.N., Franke, K. (eds.) IWCF 2008. LNCS, vol. 5158, pp. 22–34. Springer, Heidelberg (2008)
2. Zhao, X., Ho, A.T.S., Shi, Y.Q.: Image forensics using generalized benfords law for accurate detection of unknown jpeg compression in watermarked images. In: 16th International Conference on Digital Signal Processing (DSP), Greece (July 2009)
3. Chiang, P.-J., Khanna, N., Mikkilineni, A., Segovia, M., Suh, S., Allebach, J., Chiu, G., Delp, E.: Printer and scanner forensics. IEEE Signal Processing Magazine 26, 72–83 (2009)
4. Mikkilineni, A.K., Arslan, O., Chiang, P.-J., Kumontoy, R.M., Allebach, J.P., Chiu, G.T.-C., Delp, E.J.: Printer forensics using svm techniques. In: Proceedings of the IS&T's NIP21: International Conference on Digital Printing Technologies, Baltimore, MD, vol. 21, pp. 223–226 (October 2005)
5. Mikkilineni, A.K., Chiang, P.-J., Ali, G.N., Chiu, G.T.-C., Allebach, J.P., Delp, E.J.: Printer identification based on graylevel co-occurrence features for security and forensic applications. In: Security, Steganography, and Watermarking of Multimedia Contents, pp. 430–440 (2005)
6. Nitin, K., Mikkilineni, A.K., Chiang, P.-J., Ortiz, M.V., Shah, V., Suh, S., Chiu, G.T.-C., Allebach, J.P., Delp, E.J.: Printer and sensor forensics. In: IEEE Workshop on Signal Processing Applications for Public Security and Forensics, Washington, D.C, USA, April 11-13 (2007)
7. Bulan, O., Mao, J., Sharma, G.: Geometric distortion signatures for printer identification. In: Proc. IEEE Intl. Conf. Acoustics Speech and Sig. Proc., Taipei, Taiwan, pp. 1401–1404 (2009)
8. Lukas, J., Fridrich, J., Goljan, M.: Digital camera identification from sensor pattern noise. IEEE Transactions on Information Forensics and Security 1, 205–214 (2006)
9. Chen, M., Fridrich, J., Goljan, M., Lukas, J.: Determining image origin and integrity using sensor noise. IEEE Transactions on Information Forensics and Security 3, 74–90 (2008)
10. Filler, T., Fridrich, J., Goljan, M.: Using sensor pattern noise for camera model identification. In: 15th IEEE International Conference on Image Processing, ICIP 2008, pp. 1296–1299 (12-15, 2008)

11. Perez-Gonzalez, F., Heileman, G., Abdallah, C.: Benford's law in image processing. In: Proc. IEEE International Conference on Image Processing, vol. 1, pp. 405–408 (2007)
12. Fu, D., Shi, Y.Q., Su, W.: A generalized Benford's law for JPEG coefficients and its applications in image forensics. In: Proceedings of SPIE, vol. 6505, p. 65051L (2007)
13. Floyd, R., Steinberg, L.: An adaptive algorithm for spatial greyscale. Proceedings of the. Society for Information Display 17(2), 75–77 (1976)
14. Ulichney, R.: Digital Halftoning. MIT Press, Cambridge (1987)
15. Li, B., Shi, Y.Q., Huang, J.: Detecting double compressed jpeg image by using mode based first digit features. In: IEEE International Workshop on Multimedia Signal Processing (MMSP 2008), Queensland, Australia, pp. 730–735 (October 2008)
16. Chen, P.-H., Lin, C.-J.: LIBSVM: a library for support vector machines (2001) Software available at, `http://www.csie.ntu.edu.tw/~cjlin/libsvm`