http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 10 (11): 2112-2118, 2011
ISSN 1812-5638 / DOI: 10.3923/itj.2011.2112.2118
© 2011 Asian Network for Scientific Information

Copy Paper Brand Source Identification using Commodity Scanners

[1]Shize Shang, [1]Xiangwei Kong and [2]Xin'gang You
[1]School of Information and Communication Engineering,
Dalian University of Technology, Dalian, People's Republic of China
[2]Beijing Institute of Electronic Technology and Application, Beijing, China

Abstract: In print forensics, copy paper brand source identification can be used to expose the copy paper source of the forged contracts or official documents. In this study, a novel method for identifying the copy paper brand is proposed for forensics application using only a commodity scanner and without modifying the document. The scanned document image margin is cut into image blocks. After the image preprocessing, 114-D texture features are extracted from these image blocks including Gray Level Co-occurrence Matrix (GCM) features and Fourier spectrum features. The decision for each piece of paper is given by voting the decision results of the image blocks. The experimental results are provided finally to demonstrate the feasibility of the proposed method.

Key words: Print forensics, copy paper brand identification, texture features, voting decision

INTRODUCTION

Rapid technology development lead to electronic devices appear in everywhere. Printers and duplicators have been used in a growing number of applications in companies and government departments. Most of the documents produced by these output equipments are contracts, vouchers, official documents and so on. As the importance of these documents, some people tamper the documents to obtain more benefits and social order is destructed. So it is necessary to prove their authenticity. Consequently, a series of print forensics issues arise and become an important part of information security.

In the print forensics literature, there have been equipment source forensics, such as printer identification (Wu *et al*., 2009; Mikkilineni *et al*., 2005; Schulze *et al*., 2008) and scanner identification (Khanna *et al*., 2009; Khanna and Delp, 2009). For document forgery detection (Kee and Farid, 2008) models the degradation in a document caused by printing and authenticates text documents by detecting inconsistencies across the document. For detecting forged currency, tickets and authenticating passports, the uniqueness detection for blank paper is discussed in Clarkson *et al*. (2009). During the printer forensics processing, we not only need forgery detection and equipment source forensics but also paper source identification and forensics. What we can do to solve the problem currently is based on either physical methods or chemical methods (Jian *et al*., 1995) uses Fluorescence Spectrophotometry and Double Total t-test to distinguish different types of paper. Near Infrared Spectroscopy (NIRS) is used by Weiyan (2009) to identify the composition of paper. Sarkar *et al*. (2010) uses spectroscopy to identify different types of paper, the correlation coefficients are computed between unknown spectrum and library spectra to identify the paper types. Although the chemical methods reach high accuracy, they damage the documents and needs chemical reagents. The physical methods don't damage the documents but they need expensive special devices. Also the two types of methods require some professionals and long testing time.

As the texture structure on paper is relative stable, different brands of paper and even different sides of the same paper have different texture types. It is relate to the production of raw materials and production process. So the texture on paper can be considered as "fingerprint" and it's hard to forgery. So it is appropriate to discusses the paper brand identification based on texture features. In this study, we propose a method to identify the paper brand only using a commodity scanner and without modifying the document. After the image preprocessing, 114-D features are extracted from image blocks cut from document margin. SVM is used to classify the texture classes including the texture on different sides for each piece of paper. The band source identification for copy paper is given by voting the decision results of the image blocks. The experimental results indicate that we can efficiently identify the copy paper brand with high accuracy in small regions whose size is greater than or equal to 1×1 cm^2 on paper.

Corresponding Author: Shize Shang, School of Information and Communication Engineering, Dalian University of Technology, Dalian, People's Republic of China

THE PROCESS OF COPY PAPER BRAND IDENTIFICATION

The process of copy paper brand identification is illustrated in Fig. 1. After the paper samples are scanned on two sides, the image preprocessing has been done so as to get image blocks and better image quality. The document image on margin will be cut into blocks for fixed size and then the histogram of the image blocks will be equalized to enhance the texture contrast. From the image blocks, 114-D features will be extracted from the image blocks including GCM features and Fourier spectrum features. As the paper has two sides and which side on document is printed is unknown for the tester. We suppose that each piece of paper has different types of texture structures on two sides. So we use SVM to classify all types of texture including the texture of both sides on paper. As some types of copy paper have similar texture structures on two sides and they are classified wrong to each other. We only consider whether the paper brand decision is right or not and ignore the wrong decision on two sides of the paper, so we can achieve high accuracy in each class. On the document margin, all the blank blocks are selected on each side of the paper, if more than 50% of the blocks from different sides are decided to the same class X, this paper is classified into class X. Finally, the paper brand classification results have been made by voting decision.

IMAGE PREPROCESSING

The study is scanned to gray image by commodity scanner. The scanning resolution is 1200 dpi and the bit depth is 8. The proposed method is applied to detect the margin on printed document. So the size of image blocks we choose to identify should be small enough. Generally speaking, The margin width of the document ranges from 1 to 3 cm, so the size we choose in the image ranges from 1×1 to 3×3 cm^2 on paper.

As the blank image blocks whose pixel value is close to 255, the image histogram is equalized to enhance the texture contrast before the feature extraction. In order to get rid of the interference of impurities in histogram equalization, we choose the mean of image block but not for the minimum. The image block I is equalized as follows:

$$I' = \frac{I - \text{mean}(I)}{\max(I) - \text{mean}(I)} \times 255 + 128 \quad (1)$$

Figure 2 shows paper images of 8 brands after histogram equalization, with the 2×2 cm^2 paper block. As shown in Fig. 2a-h, the paper texture is similar to out-of-order texture which lacks repetitiveness. Also we can see the clear structure of texture after histogram equalization.

TEXTURE FEATURE EXTRACTION

Here, we extract 114-D features from image blocks including 64-D GCM features and 50-D Fourier spectrum features which describe the texture structures from spatial and frequency, respectively. We extract GCM features in four directions (0°, 45°, 90°, 135°) and extract the Fourier features for angle type and radiative type.

GCM features: Supposing that S is the pixel-couples' collection with specific space contact in target area and $A = \{[(x_1, y_1), (x_2, y_2)] \in S \mid f(x_1, y_1) = i \text{ and } f(x_2, y_2) = j\}$, it represents the number of pixel-couples with specific spatial relations whose values are i and j, respectively. Then the GCM p is defined as:

$$p(i,j) = \frac{|A|}{|S|} \quad (2)$$

The numerator and the denominator represent the number of pixel-couples in A and S, respectively

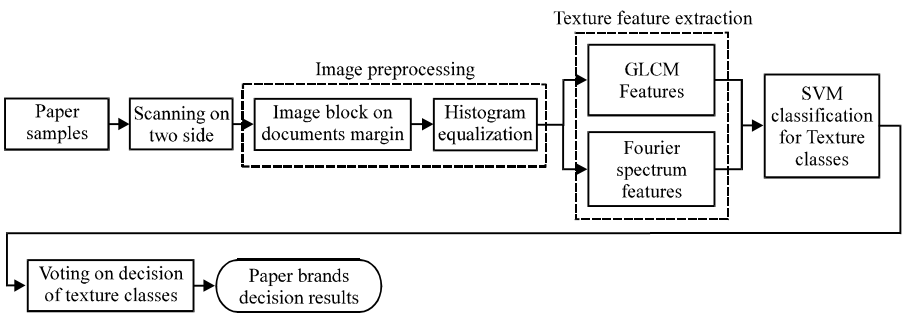

Fig. 1: The process of copy paper brand identification

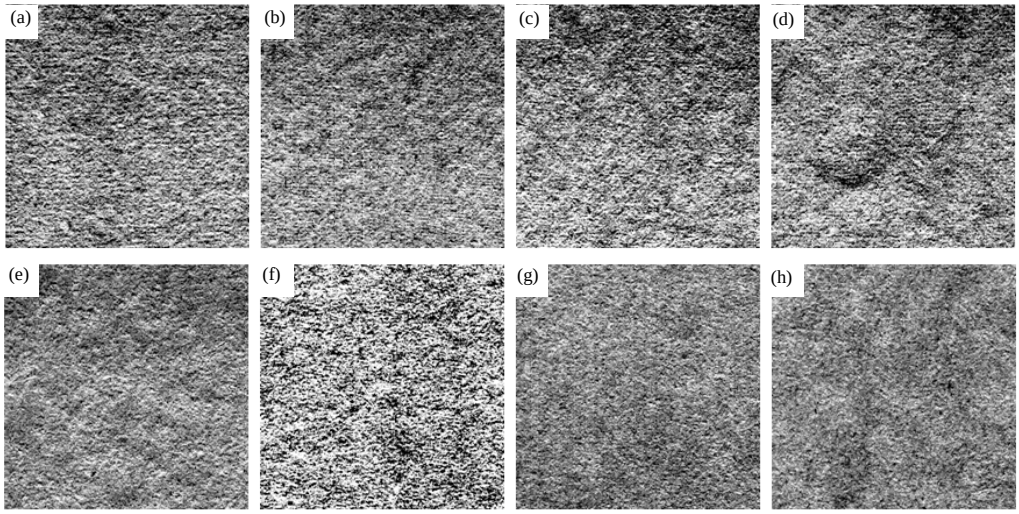

Fig. 2(a-h): Example images of 8 different brands paper after histogram equalization; (a) Blue Flagship, (b) UPM-XinLe, (c) UPM-Jia Yin, (d) Golden color, (e) Tango Blueness sword, (f) HP, (g) Double A and (h) Forerunner

Table 1: GCM features for paper brand identification

No.	Expression	Remark
1	$W_1 = \sum_{i=1}^{N}\sum_{j=1}^{N} p^2(i,j)$	
2	$W_2 = -\sum_{i=1}^{N}\sum_{j=1}^{N} p(i,j)\log p(i,j)$	
3	$W_3 = -\sum_{i=1}^{N}\sum_{j=1}^{N} \lvert i-j\rvert\, p(i,j)$	
4	$W_4 = -\sum_{i=1}^{N}\sum_{j=1}^{N} \frac{p(i,j)}{k+\lvert i-j\rvert}$	$k = 0.5$
5	$W_5 = \sum_{i=0}^{N-1} t^2[\sum_{i=1}^{N}\sum_{j=1}^{N} P(i,j)] \quad \lvert i-j\rvert = t$	
6	$W_6 = \frac{1}{\sigma_x\sigma_y}\sum_{i=1}^{N}\sum_{j=1}^{N} ijP(i,j) - \mu_x\mu_y$	μ_x and σ_x represent the mean and variance of $P_x(I, j)$, μ_y and σ_y represent the mean and variance of $P_y(I, j)$
7	$W_7 = \sum_{i=1}^{N}\sum_{j=1}^{N} (i-\mu)^2 P(i,j) = \sum_{i=1}^{N} (i-1)^2 P_x(i)$	
8	$W_8 = \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{1}{1+(i-j)^2} P(i,j)$	
9	$W_9 = \sum_{i=2}^{2N} iP_{x+y}(i)$	
10	$W_{10} = \sum_{i=2}^{2N} (i-W_{10})^2 P_{x+y}(i)$	
11	$W_{11} = \sum_{i=2}^{2N} P_{x+y}(i)\log[P_{x+y}(i)]$	
12	$W_{12} = -\sum_{i=1}^{N}\sum_{j=1}^{N} P(i,j)\log[P(i,j)]$	
13	$W_{13} = \sum_{i=0}^{N-1} (i-d)^2 P_{x-y}(i)$	$d = \sum_{i=0}^{N-1} iP_{x-y}(i)$
14	$W_{14} = -\sum_{i=2}^{N-1} P_{x-y}(i)\log[P_{x-y}(i)]$	
15	$W_{15} = \frac{W_{12} - E_1}{\max(E_x, E_y)}$	$E_1 = -\sum_{i=1}^{N}\sum_{j=1}^{N} P(i,j)\log[P_x(i)P_y(j)]$, $E_x = -\sum_{i=1}^{N} P_x(i)\log[P_x(i)]$, $E_y = -\sum_{j=1}^{N} P_y(i)\log[P_y(i)]$
16	$W_{16} = \sqrt{1-\exp[-2(E_2 - W_{12})]}$	$E_2 = -\sum_{i=1}^{N}\sum_{j=1}^{N} P_x(i)P_y(j)\log[P_x(i)P_y(j)]$

(|.| means the number). The spatial relations contain four directions (0°, 45°, 90°, 135°). Suppose that:

$$P_x(i) = \sum_{j=1}^{N} p(i,j) \quad i = 1,2,\cdots,N \tag{3}$$

$$P_y(i) = \sum_{i=1}^{N} p(i,j) \quad i = 1,2,\cdots,N \tag{4}$$

$$P_{x+y}(k) = \sum_{i=1}^{N}\sum_{j=1}^{N} p(i,j) \quad k = i+j = 2,3,\cdots,2N \tag{5}$$

$$P_{x-y}(k) = \sum_{i=1}^{N}\sum_{j=1}^{N} p(i,j) \quad k = |i-j| = 0,1,\cdots,N-1 \tag{6}$$

where, N is the maximum of pixel value. In each direction, we obtain 16-D features (Haralick and Shapiro, 1992; Russ, 2002) summarized in Table 1. So the total numbers of GCM features is 64.

Fourier spectrum features: In this study, we discuss the features extraction of Fourier spectrum in polar coordinates. Suppose that S (r, θ) is the spectrum, where θ is direction and r is frequency. S_θ (r) represents one-dimensional spectrum function in a fixed direction and S_r (θ) represents one-dimensional spectrum function at a fixed frequency. If the texture possesses spatial periodicity or affirmatory directivity, the energy spectrum will have peaks in the corresponding frequency. The spectrum features are extracted in spectrum blocks. Figure 3a and b show two types of approaches including angle type and radiative type, respectively (because the spectrum in four quadrants is symmetrical, only the first quadrant is used).

The features of angle type (Yujin, 2005) are defined as:

$$A(\theta,\theta+\alpha) = \sum\sum |F|^2(u,v) \tag{7}$$

where, $|F|^2$ is the energy spectrum, $\theta \le \tan^{-1}(v/u) \le \theta+\alpha$, $0<u, v<N/2$, $\theta = 0°, 3°, 6°, ..., 87°$, $\alpha = 3°$. N×N is the size of image block. We divide the spectrum into 30 blocks, so we get 30-D angle type features and they reflect the sensitivity for energy spectrum in the texture direction.

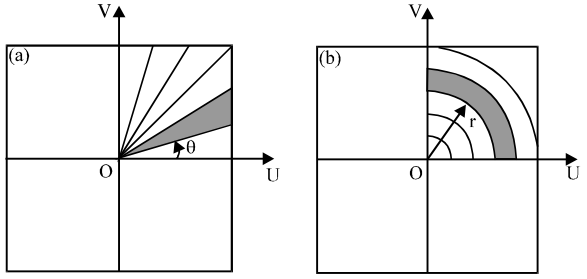

Fig. 3(a-b): Two types of block for fourier spectrum

The features of radiative type (Yujin, 2005) are defined as:

$$R(r,r+r') = \sum\sum |F|^2(u,v) \tag{8}$$

where, $|F|^2$ is the energy spectrum, $r^2 \le u^2+v^2 \le (r+r')^2$, $0<u, v<\frac{N}{2}$,

$$r = 0, \frac{1}{20}\times\frac{N}{2}, \frac{2}{20}\times\frac{N}{2}, \cdots, \frac{19}{20}\times\frac{N}{2}, r' = \frac{1}{20}\times\frac{N}{2}$$

N×N is the size of image block. We divide the spectrum into 20 blocks, so we get 20-D radiative type features and they have some connection with texture roughness.

CLASSIFICATION AND VOTING DECISION

As a piece of paper has two sides and some brands of paper have different types of texture structures on different sides of the paper. We consider one brand of paper having two types of texture. SVM is used to classify the texture classes. For some types of paper which have similar texture on two sides, the texture is classified wrong to each other. The decision processing on two sides is done to reduce the error. All the image blocks are selected on the document margin. The final decision for paper types is given by voting decision results of image blocks.

SVM classification: As some brands of paper have different types of texture structures on two sides. We will make mistake if considering one piece of paper having one type of texture. In order to solve this problem, one brand of paper is considered having two types of texture and they are classified by SVM. C-support vector classification with the non-linear RBF kernel (Chang and Lin, 2001) is used in this paper. RBF kernel is defined as:

$$K(x_i,x) = \exp\left(-\gamma\|x_i - x\|^2\right) \tag{9}$$

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

Voting decision: In this study, one piece of paper is considered having two types of texture and all the types of texture are classified by SVM. The texture on different

Table 2: A list of paper types

No.	Brand	Weight (g)	No.	Brand	Weight (g)
1	BLUE FLAGSHIP	70	12	ShuangHui	70
2	UPM-XinLe	70	13	QiangBing	70
3	UPM-JiaYin	70	14	ACEPRINT	70
4	GOLDEN COLOR	70	15	SenWang	70
5	TANGO Blueness Sword	70	16	TESCO	70
6	HP	70	17	Outdo	80
7	Double A	80	18	ZhiWang	80
8	FORERUNNER	70	19	QiCai	80
9	Outdo	70	20	LanNiao	70
10	ZhiWang	70	21	LuMei	70
11	JinMing	70			

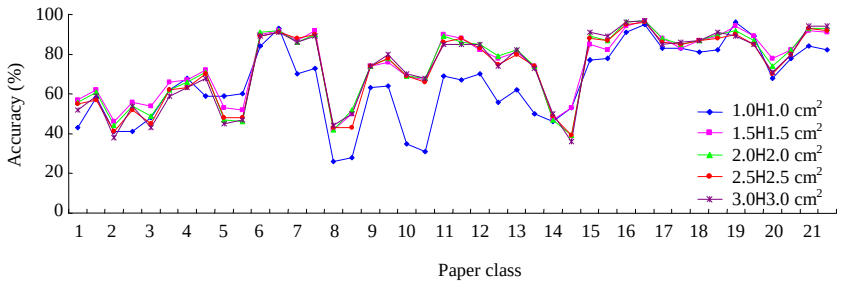

Fig. 4: The classification accuracy of blocks for 42 classes texture

sides of the paper will be classified wrong to each other if they have similar texture structures and the accuracy of these types is low. Before the voting decision, decision processing on two sides has been done to reduce the error. We only consider whether the paper brand decision is right or not, so we can achieve high accuracy in each brand if we don't consider the wrong decision on different sides of the paper.

Select all the blank blocks on the document margin to detect and the final result is given by voting all the blocks decision in each piece of paper. if more than 50% of the blocks from different sides are decided to the same class X, this paper is classified into class X. If the result does not meet the above condition, the class of the document is considered as unknown.

EXPERIMENTAL RESULTS

This section shows experimental results to identify the copy paper brand, where 21 types of paper from 19 brands shown in Table 2 are used in the experiments and 50 pieces of A4 paper (210×297 sq mm) for each type. Each piece of paper is divided into 2×2 = 4 blocks, so the total number of pieces of paper is 200 for each type. All types of paper have been scanned by Epson perfection 1200 model with 1200 dpi scanning resolution. The image format is BMP and the bit depth is 8. We consider the texture on different sides as different texture classes. So the total number of texture classes is 42.

Choose 8 pieces of paper randomly for training set and the remaining pieces as testing set. After scanning, the images are divided into image blocks, the sizes we choose are 472×472, 709×709, 945×945, 1181×1181 and 1421×1421 pixels, corresponding to 1×1, 1.5×1.5, 2×2, 2.5×2.5 and 3×3 sq cm of paper block while the scanning resolution is 1200 dpi. The experiments are repeated 10 times and the classification results for 42 classes are shown in Fig. 4.

From the results in Fig. 4, we can see the classification accuracy for both sides of paper in some classes is very low which is because the texture structures on different sides of the same paper is very similar in these classes and the image blocks are identified as the other. If we only consider whether the class decision is right or not and ignore the decision on different sides of paper, the experimental results are shown in Fig. 5, from which we can see the classification accuracy increases in some classes.

As the paper block size we choose is smaller than 3×3 cm on paper. The final decision is given by voting the decision results of image blocks in each piece of paper. In this study, all the blocks are selected from each side of paper, if more than 50% of the blocks from different sides are decided to the same class X, this paper is classified into class X. The experiments are repeated 10 times and the results are shown in Table 3.

In practical situations, most of the blank regions are on the margin. So the image block size subjects to the

Table 3: The experimental results for paper types (%)

No.	Block size (cm^2) 1×1	1.5×1.5	2×2	2.5×2.5	3×3	No.	Block size (cm^2) 1×1	1.5×1.5	2×2	2.5×2.5	3×3
1	88.16	100.00	100.00	100.00	99.90	12	71.83	100.00	100.00	99.88	99.20
2	75.51	98.91	99.92	99.48	98.37	13	71.42	99.37	99.76	99.88	98.41
3	82.04	99.37	100.00	99.25	97.67	14	94.28	100.00	100.00	100.00	99.90
4	100.00	100.00	100.00	99.94	99.90	15	100.00	100.00	100.00	100.00	100.00
5	97.55	99.68	100.00	99.77	98.83	16	100.00	100.00	100.00	100.00	100.00
6	100.00	100.00	100.00	100.00	100.00	17	88.97	100.00	100.00	100.00	99.90
7	100.00	100.00	100.00	100.00	100.00	18	90.61	100.00	100.00	100.00	99.95
8	100.00	100.00	100.00	100.00	99.86	19	100.00	100.00	100.00	100.00	100.00
9	77.14	99.53	99.92	99.48	98.00	20	77.95	100.00	100.00	100.00	99.39
10	67.75	99.68	100.00	99.54	98.27	21	100.00	100.00	100.00	100.00	100.00
11	74.28	100.00	100.00	100.00	99.90						

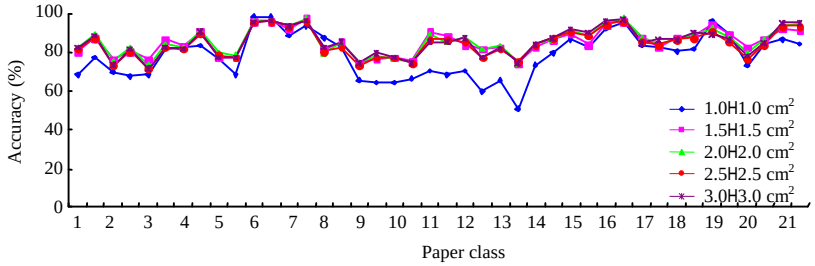

Fig. 5: The classification accuracy of blocks for 42 classes texture after decision processing

width of the margin. We select all the blank blocks in the document to detect and the final result is given by voting all the blocks decision. If more than 50% of the blocks from different sides are decided to the same class X, this paper is classified into class X. If the result does not meet the above condition, the class of the document is considered as unknown.

CONCLUSION AND FUTURE WORK

This study describes a new method to identify paper brand source only using a commodity scanner. The margin is divided into blocks and the histogram is equalized. GCM features and Fourier spectrum features are extracted from the image blocks on document margin. The total number of features is 114. SVM is used to classify the texture classes. Final decision for paper brands is given by voting the decision results of image blocks. The experimental results indicate that our method is robust to discuss the paper brand identification based on texture features.

From the experimental results shown in Fig. 4, we can find some paper brands have the same texture structures on both sides of paper, such as TANGO Blueness Sword, FORERUNNER, ACEPRINT, Double A and so on. The classification results in Table 3 shows the classification accuracy is close to 100% for 21 types of paper when the block size is greater than or equal to 1.5×1.5 cm^2.

In future works, we plan to investigate how to deal with unknown samples which are not in the training set to increase the feasibility of this method and discuss the texture changes resulting from paper aging. It will allow us to study deeper into the texture structure of paper.

ACKNOWLEDGMENT

This study is supported by the National Natural Science Foundation of China under Grant No. 60971095, and also the Fundamental Research Funds for the Central Universities.

REFERENCES

Chang, C.C. and C.J. Lin, 2001. LIBSVM: A library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Clarkson, W., T. Weyrich, A. Finkelstein, N. Heninger, J.A. Halderman and E.W. Felten, 2009. Fingerprinting blank paper using commodity scanners. Proceedings of the 30th IEEE Symposium on Security and Privacy, May 17-20, Berkeley, CA., pp: 301-314.

Haralick, R.M. and L.G. Shapiro, 1992. Computer and Robot Vision. Addison-Wesley, USA.

Jian, W., S. Qingfang, X. Limei and W. Yanji, 1995. Identification of the papers by fluorescence spectrophotometry and double total t-test. Chinese J. Anal. Chem., 23: 1181-1184.

Kee, E. and H. Farid, 2008. Printer profiling for forensics and ballistics. Proceedings of the 10th ACM Workshop on Multimedia and Security, Sept. 22-23, Oxford, pp: 3-9.

Khanna, N. and E.J. Delp, 2009. Source scanner identification for scanner documents. Proceedings of the IEEE International Workshop on Information Forensics and Security, Dec. 6-9, London, pp: 166-170.

Khanna, N., A.K. Mikkilineni and E.J. Delp, 2009. Scanner identification using feature-based processing and analysis. IEEE Trans. Inform. Forensics Security, 4: 123-139.

Mikkilineni, A.K., P.J. Chiang, G.N. Ali, G.T.C. Chiu, J.P. Allebach and E.J. Delp III, 2005. Printer identification based on graylevel co-occurrence features for security and forensic applications. Proc. SPIE, 5681: 430-440.

Russ, J.C., 2002. The Image Processing Handbook. CRC. Press, USA.

Sarkar, A., S.K. Aggarwal and D. Alamelu, 2010. Laser induced breakdown spectroscopy for rapid identification of different types of paper for forensic application. Anal. Methods, 2: 32-36.

Schulze, C., M. Schreyer, A. Stahl and T.M. Breuel, 2008. Evaluation of graylevel-features for printing technique classification in high-throughput document management systems. Lecture Notes Comput. Sci., 5158: 35-46.

Weiyan, L., 2009. Study on Application of Near Infrared Spectroscopy Technique for the Classification of Paper Raw Materials. Nanjing Forestry University, China.

Wu, Y., X. Kong, X. You and Y. Guo, 2009. Printer forensics based on page document's geometric distortion. Proceedings of the 16th IEEE International Conference on Image Processing, Nov. 7-10, Cairo, pp: 2909-2912.

Yujin, Z., 2005. Image Engineering (II): Image Analysis. Tsinghua University Press, China.