12. Tsai2014_Article_DigitalForensicsOfPrintedSourc.pdf

by

Submission date: 10-Oct-2022 09:44PM (UTC+0800)
Submission ID: 1921598887
File name: 12. Tsai2014_Article_DigitalForensicsOfPrintedSourc.pdf (1.88M)
Word count: 9925
Character count: 47526

Multimed Tools Appl (2014) 73:2129–2155
DOI 10.1007/s11042-013-1642-2

Digital forensics of printed source identification for Chinese characters

Min-Jen Tsai · Jin-Shen Yin · Imam Yuadi · Jung Liu

Published online: 11 September 2013

Abstract Recently, digital forensics, which involves the collection and analysis of the origin digital device, has become an important issue. Digital content can play a crucial role in identifying the source device, such as serve as evidence in court. To achieve this goal, we use different texture feature extraction methods such as graylevel co-occurrence matrix (GLCM) and discrete wavelet transform (DWT), to analyze the Chinese printed source in order to find the impact of different output devices. Furthermore, we also explore the optimum feature subset by using feature selection techniques and use support vector machine (SVM) to identify the source model of the documents. The average experimental results attain a 98.64 % identification rate which is significantly superior to the existing known method of GLCM by 1.27 %. The superior testing performance demonstrates that the proposed identification method is very useful for source laser printer identification.

Keywords Digital image forensics · Graylevel co-occurrence Matrix (GLCM) · Discrete Wavelet Transform (DWT) · Feature Selection

1 Introduction

With the advancement and proliferation of digital technologies, a wide variety of multimedia content has been digitalized which makes its duplication or circulation easy through distribution channels. Therefore, people can acquire high-quality images and printed documents by using electronic devices such as computers, cell phones, digital cameras and printers, which have considerable computational power with low cost. Furthermore, with these devices, confidential or private information also can be easily stolen or captured by malicious people. The increasing rate of high-tech crime indicates such a new criminal trend. Digitized images and digital content are now so easily modified and forged that the need for digital forensics has become an important issue both to effectively identify criminal activities and for the protection of intellectual property rights [1–3]. Examples of criminal activities are: threatening letters, counterfeit money and forged documents.

M.-J. Tsai (✉) · J.-S. Yin · I. Yuadi · J. Liu
Institute of Information Management, National Chiao Tung University, 1001 Ta-Hsueh Road, Hsin-Chu 300 Taiwan, Republic of China
e-mail: mjtsai@cc.nctu.edu.tw

While considerable attention has been paid in the past to research issues related on analyzing the English characters like "e", few literatures have yet been much explored for Chinese printed documents. The reason for this is that "e" is the most frequently occurring character in the English language, especially it is a vowel. However, Chinese is not a syllabification based language which does not have the basic letters of the alphabet. On the other hand, culturally, socially, and politically, Chinese is a language of global significance. The users of Chinese text number about 1.2 billion – accounting for over 15 % of the world's population. Countless works of philosophy, literature, science, health, law, art, history, religion, and political science have been written in this language over four thousand years of history. Chinese and English share the distinction of being the world's most widely spoken languages [12–14]. Therefore, it is also the motivation in this study for printed source identification for Chinese characters.

In this research, we concentrate on identifying the source brand or model of the printing devices which create the documents. Among previous works in this area, Zhu et al. [4] used unique features from print content as the print signature to register and authenticate print documents. Mikkilineni et al. [5] believed that by extracting the banding features incurred by printing defects, it is possible to authenticate the monochrome printing devices of the printed documents. This authenticating scheme is based on the fact that printers have different banding frequency sets which are dependent on brand and model. However, it is difficult to obtain the banding frequencies only from text on the printed documents. Therefore, Mikkilineni et al. [6] used graylevel co-occurrence texture features from text in the English documents to identify the source brand or model of monochrome laser printers. On the other hand, Talbot et al. [7] employed a printing discrimination method which is based on invariant moments for the authentication of inkjet printers. Khanna et al. [8] have described some forensic characterizations of a printer that involves finding intrinsic features, such as banding features or texture features, that are caused by electromechanical fluctuations and imperfections. In order to trace the originating printer, Bulan et al. [9] analyzed a printed image and exploited the locally varying geometric distortion in the printing process encountered during specific printing. Ritchey and Rego [31] presented a stego-system which generates stego-objects using context sensitive tiling. Huang and Fang [32] integrate the EXIF metadata of images and error-control codes with watermarking for copyright protection of images. Chan at al. [33] present a user-friendly system based on the use of JPEG- LS median edge predictor to determine the prime number for each block. Choi et al. [10] proposed a method which used the noise features extracted from the statistical analysis of the HH sub-band on discrete wavelet transform (DWT) for 15 RGB channel features and 24 CMYK channel features for identifying the source of color laser printer. Tsai et al. [11] leverage the previous research of [10] by not only using the noise features in the HH sub-band but also LH and HL sub-bands after DWT, for color laser printer identification. Moreover, they use feature selection method with SVM classifier to obtain the important feature set and get a good identification rate.

Another issue for this study is to gain effective identification performance when the number of unknown printed documents increases since the dimensionality of feature space will incur more computational complexity. Hence, the study aims to not only encompass Chinese characters but also proposes a useful identifying system which gives higher performance in printer source identification.

Through literature reviews with substantial experiments for ref. [5, 6], the results are not consistent with the published data due to incorrect formulas in [6]. The authors have achieved compatible outcomes by verifying those formulas which are corrected and listed at the Appendix section in this study. It is the widely accepted that the journal publications not only disseminate user's experiences and case studies in the application and exploitation of established or emerging standards, interfaces and methods, but also offers a forum for discussion on actual projects, standards, interfaces and methods by recognized experts. Since digital media is widely used

currently, the measurement which identifies the characteristics and origin of a printed device, has become a new and necessary field of research. It is to say that the assessment to identify the source model or device plays a crucial role if digital content will serve as evidence in court, similar to its non-digital counterparts. Inspired by the research of [6], this study develops a standardized forensic procedure to identify source device according to device's output which is scanned as an image. Consequently, the main contributions of this paper are the following:

(1) We analyze the images from laser printer source and provide useful forensic characterization of a printer by using GLCM and DWT features.
(2) We propose an efficient identification method for identifying large laser printer sources to get good performance on printer identification for Chinese characters.

The rest of this paper is organized as follows. In Section 2, we will give the detailed description of the proposed theoretical approach on feature extraction and classification for source color laser printer identification. Section 3, numerical results and discussion are illustrated to justify the proposed approach. Finally, the conclusion is drawn and future works are discussed in Section 4.

2 The related works and research methods

2.1 The electrophotographic printer process

The electrophotographic printer process varies among different manufacturers and the printed document is greatly influenced by the printer mechanism.

Figure 1 shows a general electrophotographic printer process [5]. The print process has six steps: charging, exposure, development, transfer, settlement and cleaning.

(1) Charging: The first step is to uniformly charge the optical photoconductor (OPC) drum when users issue a print instruction.

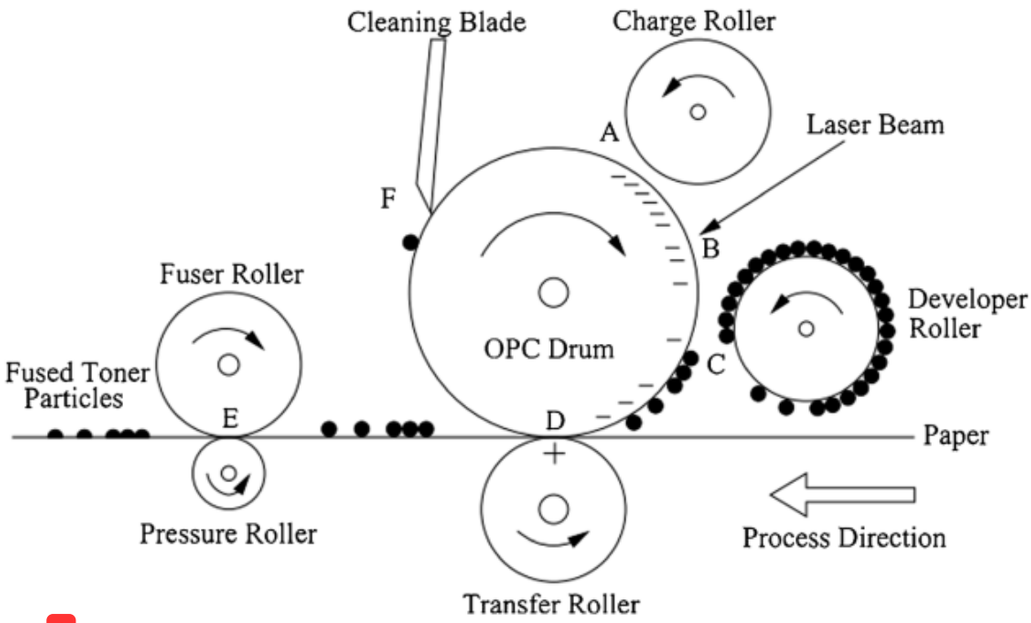


Fig. 1 Diagram of the electrophotographic printer process: **a** charging, **b** exposure, **c** development, **d** transfer, **e** fusing, **F** cleaning

(2) Exposure: A laser scans the drum and releases the electric charge on the surface ("exposing" is also known as "writing" in some documentation).
(3) Development: The discharged locations on the drum attach toner particles which are then attracted to the paper which has an opposite charge.
(4) Transfer: Transfers toner from the photosensitive drum to paper.
(5) Fusing: Melts the toner transferred to the paper to fix it.
(6) Cleaning: Finally a blade or brush to removes any excess toner left on the photoreceptor after transfer, to bring it back to its initial state.

2.2 The features

Since feature space is quite complicated, it is critical to determine a set of necessary features that can be used to describe each printer uniquely by observing a sample of the output from the printer. We will treat the output scanned document as an "image" and use image analysis tools to determine the features that characterize the printer.

2.2.1 The GLCM-based features

Graylevel Co-occurrence texture features assume that the texture information in an image is contained in the overall spatial relationships among the pixels in the image. This can be done by determining the Graylevel Co-occurrence Matrix (GLCM). GLCM features are estimates of the second order probability density function of the pixels in the image and the features are then statistics obtained from the GLCM [6].

To generate a GLCM there are four directions that could be focused on during the generation of the matrix: they are 0° (or horizontal direction), 45° direction, 90° (or vertical direction), and 135° direction, as shown in Fig. 2. The direction and spatial distance from the reference pixel i will be defined, such as 1 space at 45° direction locates the adjacent pixel j next to the reference pixel i.

First we define the number of pixels in the ROI (region of interest) as shown in Fig. 3, which is the set of all pixels within the printed area of the character, the formula for which is defined in Eq. 1 [6]. It is generated as a binary image map with all the pixels labeled as 1 within ROI, while pixels valued as 0 if they are not within ROI.

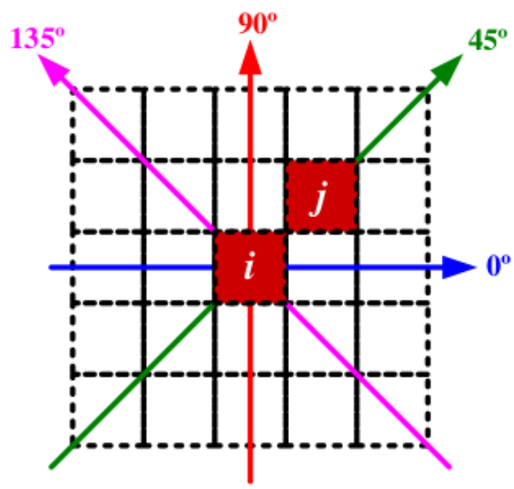


Fig. 2 The four different orientations for generation of GLCM

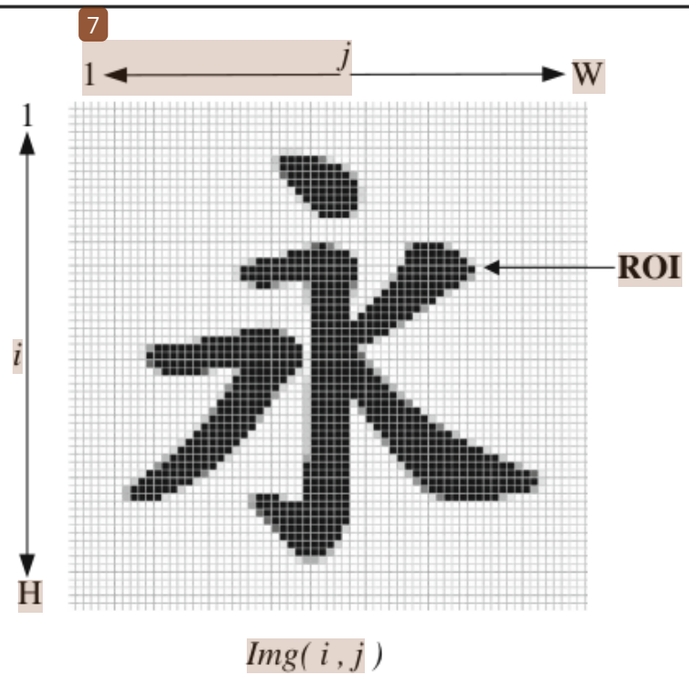


Fig. 3 An example of ROI for Chinese character "永"

$$R = \sum_{(i,j)\in ROI} 1 \tag{1}$$

We can then obtain the estimated values of the normalized GLCM from Eq. 2.

$$GLCM(i,j) = \frac{1}{\sum_{(i,j)} Img(i,j)} Img(i,j) \tag{2}$$

Note: (i, j) indicates the spatial location of image. $Img\ (i, j)$ is the probabilities from location (i, j).

When the GLCM features are generated, we have revised the formulas in [15] since some of the equations are incorrect. In this paper, there are a total number of 22 textural features that could be computed from the GLCM, such as contrast, variance, sum average, etc.. The details of the 22 features are defined by Eqs. 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23and 24.

The mean and marginal probabilities are defined from GLCM:

$$\mu_x = \sum_{i=0}^{N} i \times p_x(i) \tag{3}$$

$$\mu_y = \sum_{j=0}^{N} j \times p_y(j) \tag{4}$$

where

N is the number of distinct gray levels.

$$p_x(i) = \sum_{j=0}^{N} GLCM(i,j)$$

$$p_y(j) = \sum_{i=0}^{N} GLCM(i,j)$$

The standard deviations are defined as following:

$$\sigma_x^2 = \sum_{i=0}^{N} (i-\mu_x)^2 p_x(i) \tag{5}$$

$$\sigma_y^2 = \sum_{j=0}^{N} \left(j-\mu_y\right)^2 p_y(j) \tag{6}$$

The *Energy*, *Energy*, *Entropy*, H_{xyl}, H_{xyl}, *MaxProb*, *Corr* and *DiagCorr* are defined from GLCM:

$$Energy = \sum_{i=0}^{N} \sum_{j=0}^{N} GLCM(i,j)^2 \tag{7}$$

$$Entropy = \sum_{i=0}^{N} \sum_{j=0}^{N} -GLCM(i,j)\log(GLCM(i,j)) \tag{8}$$

$$H_{xy1} = \sum_{i=0}^{N} \sum_{j=0}^{N} -GLCM(i,j)\log\left(p_x(i)p_y(j)\right) \tag{9}$$

$$H_{xy2} = \sum_{i=0}^{N} \sum_{j=0}^{N} -p_x(i)p_y(j)\log\left(p_x(i)p_y(j)\right) \tag{10}$$

$$MaxProb = \max(GLCM(i,j)) \tag{11}$$

$$Corr = \sum_{i=0}^{N} \sum_{j=0}^{N} \frac{GLCM(i,j)(i-\mu_x)\left(j-\mu_y\right)}{\sigma_x\sigma_y} \tag{12}$$

$$DiagCorr = \sum_{i=0}^{N} \sum_{j=0}^{N} GLCM(i,j)|i-j|\left(i+j-\mu_x-\mu_y\right) \tag{13}$$

The graylevel difference histogram (GLDH) is defined as

$$GLDH(k) = \sum_{\substack{i=0 \\ |i-j|=k}}^{N} \sum_{j=0}^{N} GLCM(i,j)$$

The following four features *GLDHenergy*, *GLDHentropy*, *GLDHintertia* and *GLDHhomogeneity* are defined from *GLDH*:

$$GLDHenergy = \sum_{k=0}^{N} GLDH(k)^2 \tag{14}$$

$$GLDHentropy = \sum_{k=0}^{N} -GLDH(k)\log(GLDH(k)) \tag{15}$$

$$GLDHinertia = \sum_{k=0}^{N} k^2 GLDH(k) \tag{16}$$

$$GLDHhomogeneity = \sum_{k=0}^{N} \frac{GLDH(k)}{1+k^2} \tag{17}$$

The graylevel sum histogram (GLSH) is defined as

$$GLSH(k) = \sum_{\substack{i=0 \\ |i+j|=k}}^{N} \sum_{j=0}^{N} GLCM(i,j)$$

From GLSH, we defined 5 features:

$$GLSHenergy = \sum_{k=0}^{2N} GLSH(k)^2 \tag{18}$$

$$GLSHentropy = \sum_{k=0}^{2N} -GLSH(k)\log(GLSH(k)) \tag{19}$$

$$GLSHvar = \sum_{k=0}^{2N} (k-\mu_s)^2 GLSH(k) \tag{20}$$

$$GLSHshade = \sum_{k=0}^{2N} \frac{\left(k-\mu_x-\mu_y\right)^3 GLSH(k)}{\left(\sigma_x^2+\sigma_y^2+2\rho\sigma_x\sigma_y\right)^{3/2}} \tag{21}$$

$$GLSHprom = \sum_{k=0}^{2N} \frac{\left(k-\mu_x-\mu_y\right)^4 GLSH(k)}{\left(\sigma_x^2+\sigma_y^2+2\rho\sigma_x\sigma_y\right)^{4/2}} \tag{22}$$

where

$$\mu_s = \sum_{k=0}^{2N} k \times GLSH(k);$$

ρ is the correlation as defined in Eq. 12;
μ_x, μ_y, σ_x^2,and σ_y^2 are defined in Eqs. 3, 4, 5, 6, respectively.

In addition to the 20 gray level features above, two extra features are also included and defined in Eqs. 23 and 24. These are the variance and entropy of the pixel values in the ROI.

$$ROI_{\sigma^2} = \frac{1}{R}\sum_{(i,j)\in ROI}\left(Img(i,j)-\mu_{Img}\right)^2 \tag{23}$$

$$ROI_{entropy} = \sum_{\alpha=0}^{N} -P_{Img}(\alpha)\log\left(P_{Img}(\alpha)\right) \tag{24}$$

where

$$\mu_{Img} = \frac{1}{R}\sum_{(i,j)\in ROI} Img(i,j);$$

$$P_{Img}(\alpha) = \frac{1}{R}\sum_{(i,j)\in ROI} 1_{Img(i,j)=\alpha}$$

2.2.2 The wavelet-based features

In the spatial domain, which is the most frequent representation in the computer world, an image is comprised of many pixels and can easily be stored by a 2D matrix. In addition to representation in the spatial domain, an image can also be represented in the frequency domain through the well-known spread spectrum approach [24, 25], i.e., Discrete Fourier Transform (DFT), Discrete Cosine Transform (DCT), or Discrete Wavelet Transform (DWT). Spectrum analysis is another form of textural analysis in which direction and wavelength. Based on the research of [10, 11], the feature set in this paper will be extended to include the four HH, LH, HL, LL sub-bands that have high/low frequency coefficients after the DWT. The method of extracting DWT-based statistics analysis features is briefly explained as follows.

Having a good feature set is critical in identifying the printed image from an unknown laser printer. Gou et al. [10] use statistical method to extract noise based features as the fingerprints of a scanner. The patterns for each sub-band in the printed images from different laser printers are observed in Fig. 4. It is noted that spatial noise patterns from different laser printers are also generated when the printed image is decomposed into different wavelet sub-bands. Therefore, in this paper, applying statistical techniques as described in the following obtain image features from the DWT sub-bands of the printed images and the DWT-based features will be utilized for characterizing laser printers.

In total there are 12 statistical features used in the classification of printers:

- Standard deviation: This feature is used to measure the variability of the gray level image of the DWT sub-bands.
- Skewness: This feature is used to measure the asymmetry of the probability distribution for the gray level image in the DWT sub-bands.
- Kurtosis: This feature is used to assess the peakness of the probability distribution for the gray level image of the DWT sub-bands.

Fig. 4 Different sub-bands after DWT. (**a**) original printed image. (**c**), (**e**), (**g**) are the scanned images from different color laser printers. (**b**), (**d**), (**f**), (**h**) are the wavelet sub-bands of (**a**), (**c**), (**e**), (**g**) respectively after 1 level DWT decomposition

The three statistical features of sdv (Standard Deviation), ske (skewness) and kur (kurtosis) are defined by Eqs. 25, 26 and 27. The 12 wavelet-based features are listed in Table 1.

$$sdv = \sqrt{\frac{1}{N}\sum_{i=0}^{N}\sum_{j=0}^{N}\left(Img(i,j)-\overline{Img(i,j)}\right)^2} \tag{25}$$

$$ske = \frac{1}{N\cdot sdv^3}\sum_{i=0}^{N}\sum_{j=0}^{N}\left(Img(i,j)-\overline{Img(i,j)}\right)^3 \tag{26}$$

$$kur = \frac{1}{N\cdot sdv^4}\sum_{i=0}^{N}\sum_{j=0}^{N}\left(Img(i,j)-\overline{Img(i,j)}\right)^4 \tag{27}$$

where

$$\overline{Img(i,j)} = \frac{1}{N}\sum_{i=0}^{N}\sum_{j=0}^{N}\left(Img(i,j)\right)$$

2.3 Support vector machines (SVM)

A support vector machine (SVM) is a supervised learning technique from the field of machine learning and is used for classification and regression analysis. In this study, building an SVM [16] classifier is the most important step in the process of identifying a printer model from its images. We use SVM for both feature ranking and classification. The reason is that the most significant advantage of an SVM approach is the creation of a way to build a non-linear classifier by replacing the dot product in a linear transformation with a non-linear kernel function. Among the 3 kernel functions discussed in [17], we chose the RBF-based kernel function to build the non-linear classifier for our study.

Table 1 The 12 wavelet-based features

Symbol	Feature description
W1	Sdv of HH sub-band
W2	Ske of HH sub-band
W3	Kur of HH sub-band
W4	Sdv of LL sub-band
W5	Ske of LL sub-band
W6	Kur of LL sub-band
W7	Sdv of LH sub-band
W8	Ske of LH sub-band
W9	Kur of LH sub-band
W10	Sdv of HL sub-band
W11	Ske of HL sub-band
W12	Kur of HL sub-band

2.4 Decision fusion

As more features are used, the processing time to classify the printer source model also increases. In order to reduce the computing time and complexity, we applied the feature-selection technique to the problem of printer model identification in [18] by using algorithms such as Plus-m-minus-r [19] and SFFS [20]. In [18], we considered each individual feature-selection algorithm as an expert and used decision fusion techniques to form an important feature subset through the consensus of experts.

There are three kinds of aggregators to which fusing mechanisms can be applied.

1. Count-based aggregation
   After reviewing all alternatives available in the identification results, the best feature set for n alternatives would be select. The best n alternatives by recommending a label where n will be predefined before aggregation. Then, the alternative having the most label counts will be selected as the final solution.
2. Rank-based aggregation
   The experts will assess the performance order of all alternatives as assessment results during aggregation. Aggregating functions, such as the Borda count or the Resolution Process of GDM (RPGDM) with fuzzy preference relation [21], can be used to fuse the preferences of each expert into the final preference order. The alternative ranked first in preference order will be chosen as the final solution.
3. Confidence-based aggregation
   Similar to the count-based aggregation, the confidence-based aggregation is measured for each alternative that is not label-based by each expert. The measurement value represents the confidence level with which an expert conceives the corresponding alternative as the best alternative. The aggregating functions are used to get the final confidence level of each alternative such as sum, multiply, minimum, maximum, median, or an Ordered Weighted Average (OWA) [22].

In this paper, we consider feature selection algorithms as experts and feature sets as alternatives. It is difficult to select an optimal feature subset in a series of inclusion and exclusion steps. Giving each feature a ranking order or confidence level in the final subset is also a hard work. Therefore, we decide use the count-based aggregation as the feature selection algorithm of decision fusion. Whenever a feature is chosen into the optimal subset by a selecting algorithm, that feature gets a recommending label. Thus, based on majority vote, the features with the most labels are selected into the final optimal subset.

2.5 The scanner issues

There are several scanning factors which affect the image quality as following [15]:

- Resolution/threshold: increasing resolution enables the capture of finer detail. At some point, however, added resolution will not result in an appreciable gain in image quality, only larger file size. The key is to determine the resolution necessary to capture all significant detail present in the source document.
  The threshold setting in bitonal scanning defines the point on a scale, ranging from 0 (black) to 255 (white), at which the gray values captured will be converted to black or white pixels.
- Bit Depth: increasing the bit depth, or number of bits used to represent each pixel, enables the capture of more gray shades or color tones. Dynamic range is the term used

to express the full range of tonal variations from lightest light to darkest dark. A scanner's capability to capture dynamic range is governed by the bit depth used and output as well as system performance. Increasing the bit depth will affect resolution requirements, file size, and the compression method used.

- Enhancement: enhancement processes improve scanning quality but their use raises concerns about fidelity and authenticity. Typical enhancement features in scanner software or image editing tools include descreening, despeckling, deskewing, sharpening, use of custom filters, and bit-depth adjustment.
- Color: capturing and conveying color appearance is arguably the most difficult aspect of digital imaging. Good color reproduction depends on a number of variables, such as the level of illumination at the time of capture, the bit depth captured and output, the capabilities of the scanning system, and mathematical representation of color information as the image moves across the digitization chain and from one color space to another.
- System Performance: the equipment used and its performance over time will affect image quality. Different systems with the same stated capabilities (e.g., dpi, bit depth, and dynamic range) may produce dramatically different results. System performance is measured via targets that check for resolution, tone reproduction, color rendering, noise, and artifacts.
- Operator Judgment and Care: the skill and care of a scanning operator may affect image quality as much as the inherent capabilities of the system. We have noted the effect of threshold in bitonal scanning; operator judgment can minimize line drop out or fill-in. When digital cameras are used, the lighting becomes a concern, and the skills of the camera operator will come into play. A quality control program must be instituted to verify consistency of output.

Since this study only considers the text information, the scanned data will be converted into grayscale only. Therefore, several issues like bit depth, enhancement, dynamic range and color would not influence the final judgment. We fully agree the system performance and the experience of the scanner operator will affect the image quality which in the long run will make substantial difference for source identification. In this study, the scanned documents are printed texts which are generally surrounded by border space. The extracted characters are carefully examined. In addition, the scanner technology improves significantly recently and we have checked several brand scanners like Hp, Epson, Brother with low dpi resolution under proper scanned procedures. The results make almost no difference for the identification ratio even the operators are different, and the training and testing sets are also different. Therefore, under normal operational procedures with good maintenance, the system performance and operator judgment will not be the factors in this study accordingly to streamline the whole procedures.

2.6 The proposed approach

The identifying printer procedures are briefly described as following and the flow chart is shown in Fig. 5.

(1) Collecting the printed documents from different printers.
(2) Scanning the documents with 8 bit/pixel (grayscale) by a BenQ 3300U scanner. In the next, all the Chinese characters "永" in the document are extracted. The reason for this is that "永" (means "eternal" or "eternity") contains all the basic types of strokes for the Chinese calligraphy. Fig. 6 shows the eight basic strokes from the character "永". All modern Chinese characters are drawn with a palette consisting of eight basic strokes [23]. Generally all strokes are painted from top to bottom and left to right – with exceptions for characters number six, which is draw upwards. In the following overview each type of stroke is shown (drawn in black colour) within actual characters (drawn in grey colour) [23] in Table 2.

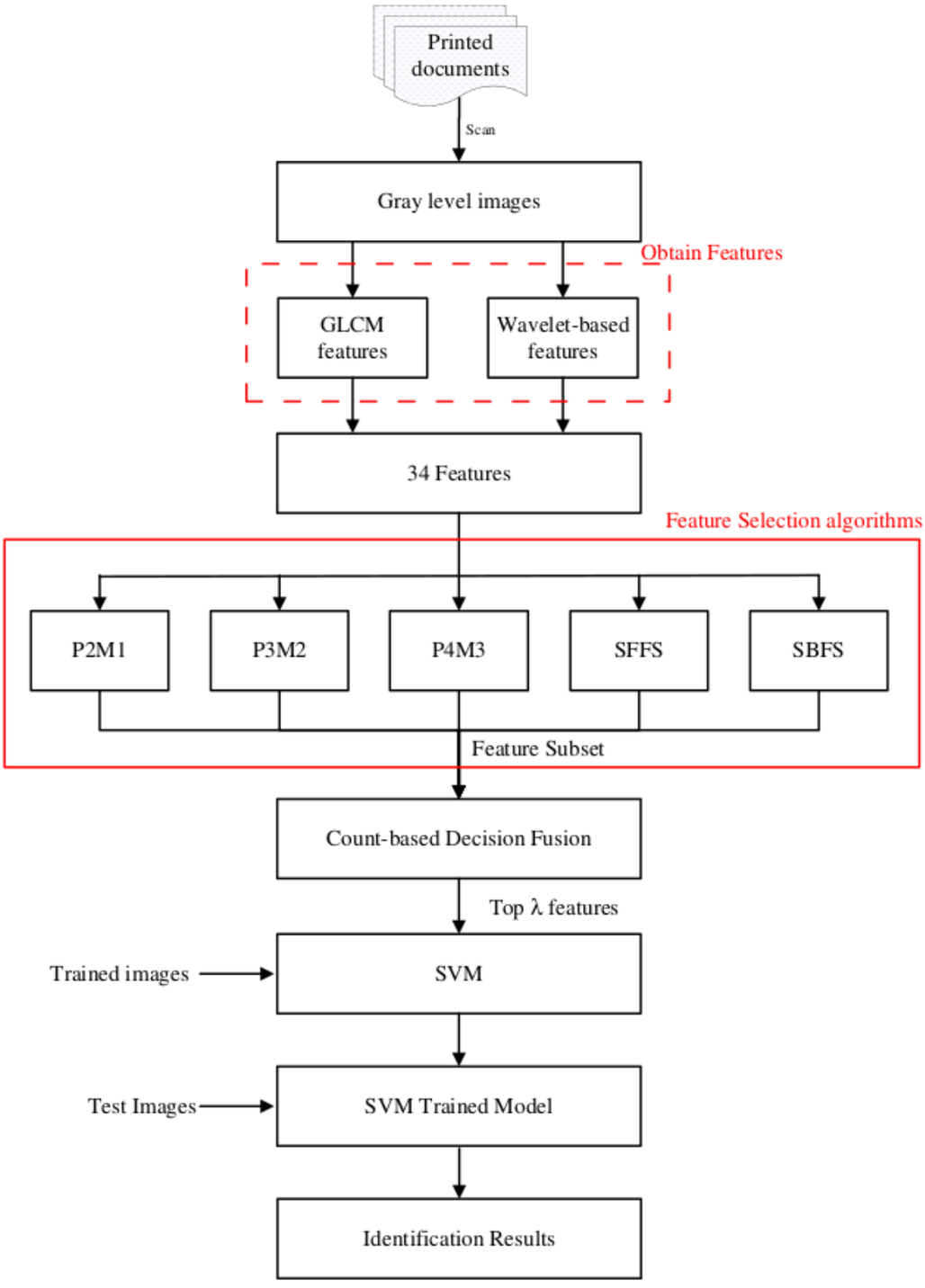


Fig. 5 Procedure of identifying source laser printers

(3) Based on our previous research [11], we extend the feature set of the proposed model by using 22 GLCM-related features and 12 wavelet-related features. Because the method of GLCM features in [15] was proven to effectively identify a printer source device.

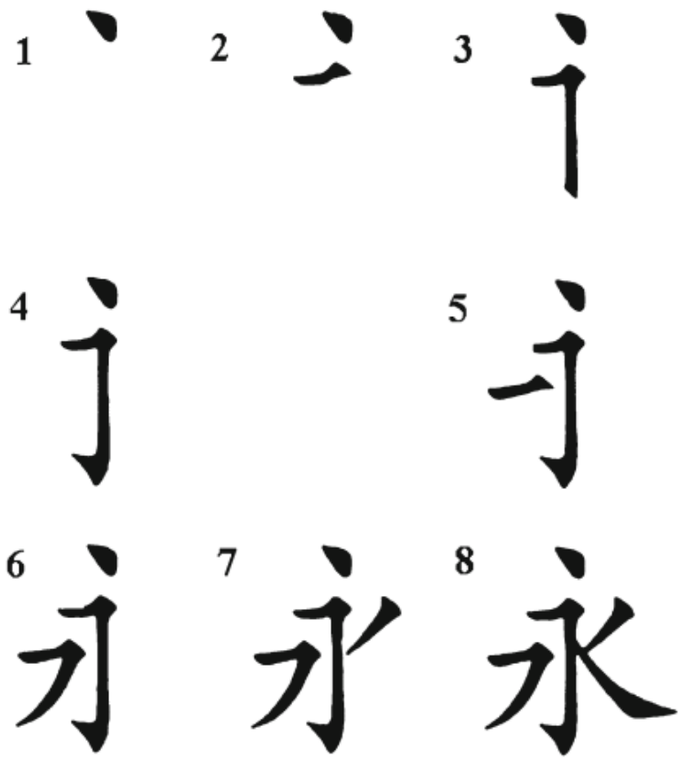


Fig. 6 The eight basic strokes from the character "永"

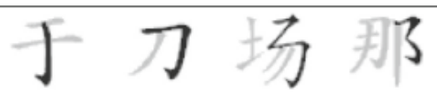
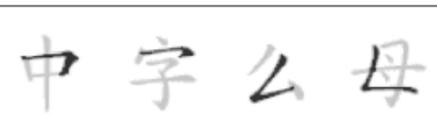
(4) After obtaining 34 features, we implement the feature selection algorithms to aggregate the feature subsets. In the final feature subset, the algorithm will select the top λ features form 34 features in order to get the highest identification rate.
(5) Suppose that the highest rate is achieved when λ features are selected, subsequently, the SVM trained model will be built by using the top λ features from the training images.
(6) Finally, the test images will be fed into the trained model to predict the printer source model and get the identification results.

3 Experiments and discussion

Before we discuss the experimental settings, we must first determine the sample size needed for the experimental analysis to give good (or statistically significant) identification results. Due to the modest sample size and low false positive rate, we have adopted the setting of Mikkilineni [15] who applied 500 images for training and 300 images for testing in order to make a fair comparison. Second, to conduct verification for experimental results of [15], we have performed the same experiment for English character "e" using GLCM features and the accuracy rate is as high as 97 % which is shown in Table 3. Therefore, we are confident that GLCM features can be used for printer source identification.

Four experiments are conducted with Chinese characters in this study to verify the proposed method and the experimental design chart is shown in Fig. 7. First, we used 12 different printers to print the same file with ".doc" format. The brands and models of 12

Table 2 Each type of stroke

no	means	stroke
1	Dot (dian3)	
2	Horizontal (heng2)	
3	Vertical (shu4)	
4	Hook (gou1)	
		
5	Rising (ti2)	
6	Slanting to the left (pie3)	
7	Turning (zhe2)	
8	Slanting to the right (na4)	

printers used in the study are shown in Table 4. Next step, we scanned all printed documents with 8 bits/pixel (grayscale) by scanner BenQ 3300U with default scanner setting and extracted the "永" letter images from the file. In all experiments, 500 "永" letter images from a printer are randomly selected to train the SVM classifier, whereas at least another 300 images, randomly taken from the same document data set, are tested during the identification procedure of the printer source model.

Table 3 The confusion matrix of identification results using GLCM features by method [15] (%)

Avg=97.00	1	2	3	4	5	6	7	8	9	10	11	12
1	99.50	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.20	98.40	0.00	0.90	0.00	0.00	0.10	0.10	0.00	0.00	0.00	0.30
3	0.00	0.10	99.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.60	0.00
4	0.00	0.80	0.00	99.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	0.00	0.00	0.10	0.00	0.00	98.80	0.00	0.00	0.00	0.00	1.10	0.00
7	0.10	0.00	0.00	0.00	0.00	0.00	89.10	0.00	3.90	6.90	0.00	0.00
8	0.00	0.10	0.00	0.00	0.00	0.00	0.00	99.80	0.00	0.10	0.00	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00	3.50	0.10	91.20	4.90	0.00	0.40
10	0.00	0.00	0.00	0.00	0.00	0.00	5.10	0.00	3.20	91.40	0.00	0.30
11	0.00	0.00	0.30	0.00	0.00	0.70	0.00	0.00	0.00	0.00	99.00	0.00
12	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.80	0.00	98.70
False positive rate	0.02	0.16	0.02	0.01	0.00	0.12	0.89	0.01	0.78	0.75	0.03	0.14

The brands and models of 12 printers in Table 3 are shown in Table 4

3.1 Experiment I: different scanner resolution comparison

A pixel is the smallest element on the display screen. A screen contains thousands of pixels each of which can be made up of one or more dots or a cluster of dots. The more pixels or dots that make up the display screen, the clearer the resolution or image will be. Scanner resolution has a similar concept to a display screen. Working with scanned documents means converting them into digital images. A high resolution image appears crisper, and its texture will often be more clear and vibrant. Unfortunately, a high-resolution image will also need more processing time to identify the source printer (for example, scanner time, training time, testing time, and so on.).

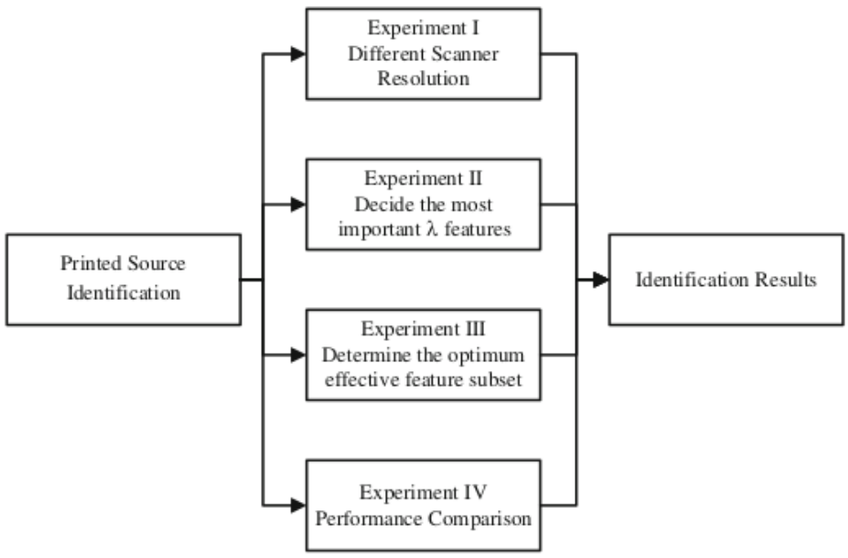


Fig. 7 Experimental design Chart

Table 4 Printers used for classification

No	Brand	Model
1	HP	ColorLaserJet CP2025
2	HP	DesignJet 111
3	HP	LaserJet 1536dnf MPF
4	HP	LaserJet 2200D
5	HP	ColorLaserJet 3800dn
6	HP	LaserJet 4050
7	HP	LaserJet 4100
8	HP	LaserJet 4250n
9	HP	LaserJet 4300
10	HP	LaserJet 4350n
11	HP	LaserJet M1120 MFP
12	Sharp	AR m205

In [15], the research set the scanner resolution of 2400 dpi for their experiment. However, such a setting is little use in practice since scanning a single A4 document takes more than 10 min and the file size will be more than 100 Mb. In order to explore the resolution's contribution for the identification accuracy rate, we adopt different resolutions, as shown in Table 5. We set the character "永" with the size 10 pt DFKai-sb(標楷體) for testing.

The steps of experiment I are listed as follows:

(1) 10 sets of images from our scanned image database of 12 printer sources are randomly generated. In each set, there are 500 images which are selected from each printer as training data and another 300 images for test data. The 22 GLCM features, 12 wavelet-based features, and all features (22 GLCM features+12 wavelet-based features) are then calculated.
(2) Apply the SVM engine to build the prediction models using GLCM features, wavelet-based features, and all features.
(3) Feed the test image subsets to the corresponding model trained in step 2 for the printer source prediction.
(4) Repeat step 1 through 3 ten times to obtain the predicted results.

As shown in Table 6, higher resolution gets better identification result. The accuracy rates when using all features at 300 dpi and 600 dpi are higher than 98 % and the results show that the resolutions with 300dpi or 600 dpi have sufficient information for printer identification. In addition, the results for 300dpi and 600dpi differ by less than 1 % which are in the acceptable range. Therefore, the 300dpi resolution will be used in this paper.

Table 5 The pixel size of "永" for different scanner resolutions by BenQ 3300U

Resolution	The pixel size
150 dpi	28×28
300 dpi	57×51
600 dpi	94×92

Table 6 The accuracy rate for different scanner resolutions by BenQ 3300U (%)

Feature set Resolution	GLCM features 22 features	Wavelet-based features (DWT) 12 features	All (GLCM+DWT) 34 features
150 dpi	92.79	64.72	94.48
300 dpi	97.18	85.74	98.23
600 dpi	97.65	91.93	99.14

3.2 Experiment II: Decide the most important λ features

To search the most important features and reduce the evaluation time without the loss of accuracy, the adaptive feature selection algorithm is implemented. According to [19, 20], we implemented five feature selection algorithms in Java: SFFS, SBFS, plus-2-minus-1 (P2M1), plus-3-minus-2 (P3M2), and plus-4-minus-3(P4M3). The number of chosen features is decided based on the accuracy rate for all 34 features (GLCM+wavelet-based features). The steps of experiment II are listed as follows:

(1) 10 sets of images from 12 printer sources are randomly generated. In each set, there are 500 images which are selected from each printer as training data and another 300 images for test data.
(2) The feature selection algorithm is executed by adding or removing one feature at a time to find the optimum identification rate. The selection order during execution is recorded to choose the most important features.
(3) Repeat step 2 for 10 different image sets.
(4) The diagram of accuracy rate versus number of features is plotted to decide the value of λ for the most important features. As Fig. 8 shows, the accuracy rate for different feature sets reaches a maximum near 24 features. Hence, we choose 24 as the value of λ.

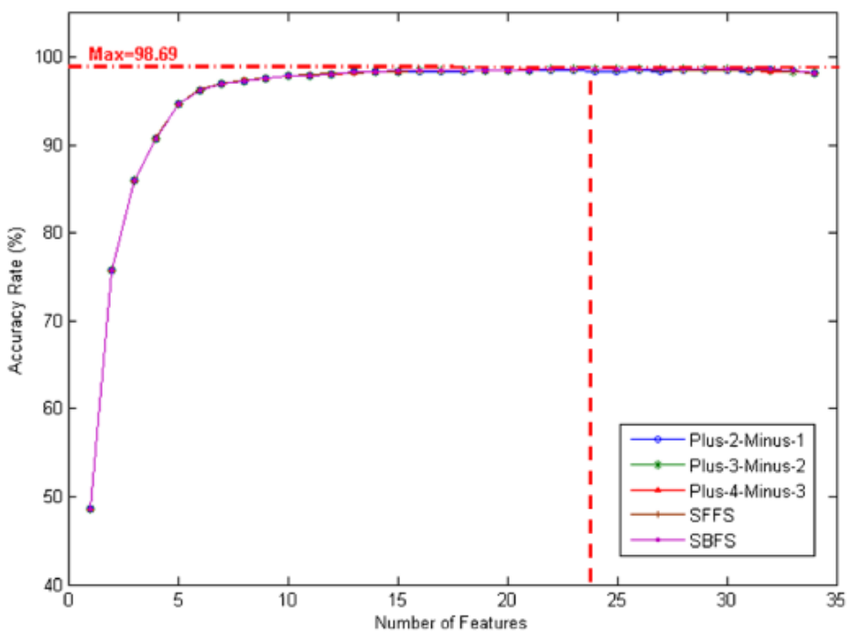


Fig. 8 Predicting accuracy rate versus number of features used

3.3 Experiment III: Determine the optimum effective feature subset (24 features)

We conducted this experiment by using images from 12 printers to verify the effectiveness of the 24 most important features. The steps to execute the experiment III are similar to experiment II and the 24 most effective features are tabulated in Table 7.

The steps to execute the experiment are as follows:

(1) 10 sets of images from 12 printer sources are randomly generated. In each set, there are 500 images which are selected from each printer as training data and another 300 images for test data.
(2) The feature selection algorithm is executed by adding or removing one feature at a time to find the optimum identification rate. The selection order during execution is recorded to choose the most important features.
(3) Repeat step 2 for 10 different image sets.
(4) Using the recorded feature-selection order, the counter-based decision fusion algorithm is used to decide the final top 24 selected features from the results of 10 tests.

As shown in Table 8 and Table 9, the accuracy rate when using the all feature set and the 24 important features are 98.29 % and 98.31 %. Moreover, the average accuracy rate of the 24 important features set, optimized by the proposed method, increases by 0.02 % – more accurate than the results using all 34 features. The high average accuracy rate justifies the effectiveness of our proposed method in identifying the printer source model when the optimum feature subset is used.

In addition, here we also use another Chinese character "的" (the most common used preposition in Chinese, means "of") with the 24 most important features in Table 7 to verify the results by 7 printers, as shown in Table 10. The accuracy rate is also high under the proposed approach.

3.4 Experiment IV: Performance comparison of different font size and font type

20

In a typical forensic printer identification scenario, it may not be possible to obtain all the same font size and font type from the unknown printed documents. Thus, in this experiment,

Table 7 The 24 most important features

Symbol	Features description	Symbol	Features description
GLCM1	Mean of r values	GLCM21	ROI Variance
GLCM2	Mean of c values	GLCM22	ROI Entropy
GLCM3	Variance of r values	W1	Sdv of HH subband
GLCM4	Variance of c values	W3	Kurtosis of HH subband
GLCM9	Max Prob of GLCM	W4	Sdv of LLL subband
GLCM10	Corr of GLCM	W5	Skewness of LL subband
GLCM11	DiagCorr of GLCM	W6	Kurtosis of LL subband
GLCM12	GLDHenergy	W7	Sdv of LH subband
GLCM13	GLDHentropy	W8	Skewness of LH subband
GLCM18	GLSHvar	W9	Kurtosis of LH subband
GLCM19	GLSHshade	W10	Sdv of HL subband
GLCM20	GLSHprom	W12	Kurtosis of HL subband

Table 8 The confusion matrix of identification results using all features (%)

Avg=98.29	1	2	3	4	5	6	7	8	9	10	11	12
1	98.80	0.87	0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00
2	1.27	96.07	0.00	0.00	0.00	0.00	2.27	0.10	0.23	0.00	0.00	0.07
3	0.10	0.00	99.90	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.07	0.13	0.00	99.27	0.00	0.00	0.23	0.30	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.00	0.00	99.80	0.00	0.00	0.00	0.00	0.20	0.00
7	0.80	1.77	0.10	0.00	0.00	0.00	95.40	0.73	0.77	0.43	0.00	0.00
8	0.00	0.07	0.00	0.27	0.00	0.00	0.40	99.27	0.00	0.00	0.00	0.00
9	0.07	0.03	0.00	0.00	0.00	0.00	0.83	0.00	97.37	1.67	0.00	0.03
10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.63	96.90	0.00	2.47
11	0.00	0.00	0.13	0.00	0.00	0.23	0.00	0.00	0.00	0.00	99.63	0.00
12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	2.80	0.00	97.10
False positive rate	0.12	0.25	0.01	0.01	0.00	0.02	0.32	0.02	0.08	0.72	0.01	0.25

we will explore the different text format environment and verify the proposed method with different font size and font type in document content, as shown in Table 11.

The steps to execute the experiment are as follows:

(1) T sets of images from 12 printer sources are randomly generated. And the printer sources acquirement is based on the eight variables in Table 11. In each set, there are 500 images which are selected from each printer as training data and another 300 images for test data.
(2) Repeat step 2 for 10 different image sets.
(3) Use the SVM engine to build the four prediction models by all 34 features, the optimum effective 24 features subset from experiment III, 22 GLCM features, and 12 DWT features.

Table 9 The confusion matrix of identification results using 24 features (%)

Avg=98.31	1	2	3	4	5	6	7	8	9	10	11	12
1	98.60	0.80	0.03	0.00	0.00	0.00	0.53	0.00	0.03	0.00	0.00	0.00
2	1.27	96.03	0.03	0.00	0.00	0.00	2.60	0.00	0.07	0.00	0.00	0.00
3	0.13	0.00	99.80	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00
4	0.17	0.10	0.00	99.27	0.00	0.00	0.20	0.27	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.00	0.00	99.80	0.00	0.00	0.00	0.00	0.20	0.00
7	0.90	1.40	0.03	0.00	0.00	0.00	95.60	0.70	0.90	0.47	0.00	0.00
8	0.00	0.07	0.00	0.23	0.00	0.00	0.30	99.40	0.00	0.00	0.00	0.00
9	0.07	0.13	0.00	0.00	0.00	0.00	0.73	0.00	97.50	1.53	0.00	0.03
10	0.00	0.00	0.00	0.00	0.00	0.00	0.13	0.00	0.70	96.73	0.00	2.43
11	0.00	0.00	0.03	0.00	0.00	0.20	0.00	0.00	0.00	0.00	99.77	0.00
12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.20	2.63	0.00	97.17
False positive rate	0.15	0.28	0.01	0.01	0.00	0.02	0.30	0.02	0.07	0.75	0.01	0.22

Table 10 The confusion matrix of identification results for Chinese character "的" by using 24 features (%)

Avg=98.21		Printer No.						
		1	3	4	5	6	7	12
Printer No.	1	100.00	0.00	0.00	0.00	0.00	0.00	0.00
	3	0.00	100.00	0.00	0.00	0.00	0.00	0.00
	4	0.00	0.00	100.00	0.00	0.00	0.00	0.00
	5	0.02	0.00	0.02	99.96	0.00	0.00	0.00
	6	0.00	0.00	0.00	0.00	94.24	0.24	5.52
	7	0.00	0.00	0.00	0.00	0.54	99.24	0.22
	12	0.00	0.00	0.00	0.00	5.88	0.06	94.06
False positive rate		0.00	0.00	0.00	0.01	0.48	0.01	0.33

(4) Feed all 34 features, the optimum effective 24 features subset from experiment III, 22 GLCM features and 12 DWT features test image subsets to the corresponding model trained in step 3 for the printer source prediction.
(5) Compare the predicted source in step 4 with its original source to get the identification accuracy rate.

To further investigate the effect of feature set from the eight variables, we tabulate the performance in Table 12.

As shown in Table 12, the average accuracy rate when using the optimum effective 24 features subset is the highest. From Table 12, there is almost 2 % difference for optimum effective 24 features and 22 GLCM features for 14 pt font. Adding DWT features with feature selection can help to improve the accuracy rate.

Regarding the computation time, it depends on the equipment applied for the experiments. In addition, high speed processer and solid state drive (SSD) can reduce the total computation cost if the research funding can support the expense. From algorithm analysis, there are immediately about one thirds features saved during the calculation by using the feature selection and the computation saving is meaningful for large scale study.

The average experimental results attain a 98.64 % identification rate by using the optimum features which is significantly superior to the existing known method of GLCM

Table 11 Eight variables considered for forensic identification experiment

No	Font type	Font size
1	DFKai-sb(標楷體)	08 pt
2	DFKai-sb(標楷體)	10 pt
3	DFKai-sb(標楷體)	12 pt
4	DFKai-sb(標楷體)	14 pt
5	PMingliu(新細明體)	08 pt
6	PMingliu(新細明體)	10 pt
7	PMingliu(新細明體)	12 pt
8	PMingliu(新細明體)	14 pt

Table 12 The performance of variable font size and font type

		All 34 features	The optimum effective 24 features	22 GLCM features	12 DWT features
DFKai-sb (標楷體)	8 pt	97.71 %	97.75 %	95.43 %	84.44 %
	10 pt	98.29 %	98.31 %	97.08 %	86.06 %
	12 pt	98.20 %	98.21 %	96.57 %	85.78 %
	14 pt	98.51 %	98.58 %	96.50 %	87.06 %
PMingliu (新細明體)	8 pt	98.84 %	98.90 %	97.76 %	88.47 %
	10 pt	99.05 %	99.08 %	98.54 %	90.57 %
	12 pt	98.80 %	98.82 %	97.89 %	85.62 %
	14 pt	99.46 %	99.51 %	99.21 %	87.56 %

by 1.27 % from Table 12. These results demonstrate the effectiveness of our proposed method for improving the accuracy rate of identifying the printer source.

3.5 Discussion

There are several issues that the authors would like to address in this section.

(1) Determining the appropriate samples

As described previously in the introduction, there is no research conducted on printed source identification with Chinese characters as content. In [15], Mikkilineni et al. used the English letter "e" as the experimental samples for printer identification because "e" is the most commonly used letter in the English language. However, Chinese contains many more different characters than English. Therefore, finding a set of suitable Chinese characters to be the experimental samples is still under investigation.

(2) The typical generic features

Hence, we only utilize GLCM-based and DWT-based features in the printer forensic system. However, a well functioning forensic system should explore the generic features of different device source identification and still maintain the effectiveness of its forensic rate. Since there is no such discussion available in our literature survey, this could be the topic of further investigative research.

(3) Printer Manufacturers and the number of printers

In this study, we collect 12 different printers and most printers are made by the same manufactures-HP. The possible reason for this situation is that the HP printer is the most widely used printer brand with over 40 % market share in Taiwan [26]. The future research could collect more different printer brands or increase the number of printers for examining the interaction between the printed documents and printer manufactures.

(4) Color images for color laser printers

Due to the most collected sample were from the monochrome laser printers, we only consider gray-scale images from laser printers in this study. However, people may use the color laser printers to print their documents. The implementation of the collecting color printed files will be needed to be studied.

(5) Complexity analysis of the proposed method

The computation complexity of proposed method is low from the view of mathematical analysis. The whole complexity should be discussed for wavelet transform, GLCM and mathematical static calculation respectively.

Suppose the synthesis filters are h (low-pass) and g (high-pass) for wavelet transform. Take $|h|=2N$, $|g|=2M$, and assume $M \geq N$. The cost of the standard algorithm for CDF 9/7 filters [27] is $4(N+M)+2$ and could be sped up by the lifting algorithm in [28] to $2(N+M+2)$. The computation of wavelet transform is linear time mathematics.

On the other hand, GLCM statistics are employed in the spatial domain. Eqs. 3, 4, 5 and 6) are marginal means and variances. The next seven features (Eqs. 7, 8, 9, 10, 11, 12 and 13))are the energy of the normalized GLCM, three entropy measurements, the maximum entry in the GLCM, and two correlation metrics. Eqs 14, 15, 16 and 17) are the energy, entropy, inertia and local homogeneity of difference histogram GLDH. Another 5 features, Eqs 18, 19, 20, 21 and 22, are obtained from the sum histogram GLSH and they are the energy, entropy, variance, cluster shade and cluster prominence of GLSH. Additional two features, Eqs. 23 and 24 are the variance and entropy of the pixel values in the ROI. Since GLCM features (3–22) are manipulated within the range of distinct gray level (0–255), the complexity is linear. Assuming the image width is l where l is the pixel size as defined in Table 5, the complexity of GLCM features (Eqs. 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23 and 24) is generally less than or equivalent to the complexity of the local variance $O(l^2)$ for each character and the total complexity is no more than $O(l^2)$.

The complexity of statistical features of standard deviation, skewness and kurtosis operated at the DWT coefficients is dominated by the kurtosis complexity which is$\approx O(l^4)$. In summary, the total amount of complexity approximately equals to $O(l^4)$. Since l is small (as shown in Table 5), the total amount of calculation is also low in practical applications.

(6) ROI analysis

In this study, the output scanned document as an "image" and use image analysis tools to determine the features that characterize the printer. The printed areas of the document have fluctuations which can be viewed as texture. This research assumes that the texture in a document is predominantly affected in the processing direction. Consequently, the generation of GLCM is considering the variation in the direction of 90° during the initial ROI selection. To further improve the ROI, some image enhancement algorithms need to be applied to select ROI. The main aim of enhancement is making objects homogeneous while increasing contrast between objects and background. Certain edge detection algorithms, such as Canny operator [29] and edge detection with local Scale Control [30] could help to find the proper boundary of ROI and further improve the accuracy rate.

4 Conclusion and future research

This study focused on analyzing the relationship between digital printers and the printed documents with Chinese characters through the help of support vector machines and decision fusion. The proposed approach utilizes feature selection algorithms to choose the top λ (λ=24 based on experimental results) important features from the GLCM-based and DWT-based features. From the experiments, it is determined that the identification accuracy rate can achieve 98.64 % when 12 printers are examined. By integrating these parameters, the data shows a high printer source identification rate by our approach, proving the efficiency of its forensic application. In summary, the main contributions of this paper are listed as follows:

(1) We analyze the images (Chinese characters) from laser printer source and provide useful forensic characterization of a printer by using features based on GLCM-based and DWT-based features.

(2) We propose an efficient identification method for identifying large laser printer source which gives both good performance on printer identification and significantly reduces the total computation time.
(3) The selection of feature dimension should be appropriate based on the capability of the computer facilities to avoid huge computation cost. However, large scale investigation is inevitable, especially, large number of Chinese characters or printer sources. Therefore, feature selection plays significant role to reduce the feature dimension and finally relieve the burden of the total computation tasks.

For future research, we will not only explore more features to enrich our feature set and improve the identification accuracy rate but also collect as many printers as possible per brand to examine the identification performance.

Acknowledgements This work was supported by the National Science Council in Taiwan, Republic of China, under Grant NSC99-2410-H-009-053-MY2 and NSC101-2410-H-009-006-MY2.

Appendix

The revised/adjusted formulas in this study and in Mikkilineni [15]

	The formulas in this study	The Formulas in Mikkilineni [15]
Revised formula	$GLDHenergy = \sum_{k=0}^{N} GLDM(k)^2$	$GLDHenergy = \sum_{k=0}^{N} GLDM(k)$
	$GLSHenergy = \sum_{k=0}^{2N} GLSH(k)^2$	$GLSHenergy = \sum_{k=0}^{2N} GLSH(k)$
	$GLSHshade = \sum_{k=0}^{2N} \frac{(k-\mu_x-\mu_y)^3 GLSH(k)}{(\sigma_x^2+\sigma_y^2+2\rho\sigma_x\sigma_y)^{3/2}}$	$GLSHshade = \sum_{k=0}^{2N} \frac{(k-\mu_x-\mu_y)^3 GLSH(k)}{(\sigma_x^2+\sigma_y^2+2\rho\sigma_x\sigma_y)^{3/2}}$
	$GLSHprom = \sum_{k=0}^{2N} \frac{(k-\mu_x-\mu_y)^4 GLSH(k)}{(\sigma_x^2+\sigma_y^2+2\rho\sigma_x\sigma_y)^{4/2}}$	$GLSHprom = \sum_{k=0}^{2N} \frac{(k-\mu_x-\mu_y)^4 GLSH(k)}{(\sigma_x^2+\sigma_y^2+2\rho\sigma_x\sigma_y)^{4/2}}$
Adjusted formula	$\mu_x = \sum_{i=0}^{N} i \times p_x(i)$	$\mu_x = \sum_{i=0}^{N} p_x(i)$
	$\mu_y = \sum_{j=0}^{N} j \times p_y(j)$	$\mu_y = \sum_{j=0}^{N} p_y(j)$
	$\sigma_x^2 = \sum_{i=0}^{N} (i-\mu_x)^2 p_x(i)$	$\sigma_x^2 = \sum_{i=0}^{N} i^2 \times p_x(i) - \mu_x^2$
	$\sigma_y^2 = \sum_{j=0}^{N} \left(j-\mu_y\right)^2 p_y(j)$	$\sigma_y^2 = \sum_{j=0}^{N} j^2 \times p_y(j) - \mu_y^2$

References

1. Bulan O, Junwen M, Sharma G (Apr. 2009) "Geometric distortion signatures for printer identification," *in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.1401–1404
2. Canny J (1986) A computational approach to edge detection. IEEE Trans Pattern Anal Machine Intell PAMI-8:679–698
3. Chan CS, Chang CC, Vo HP (2012) A User-Friendly Image Sharing Scheme Using JPEG-LS Median Edge Predictor. Journal of Information Hiding and Multimedia Signal Processing 3(4):340–351
4. Chiclana F, Herrera F, Herrera-Viedma E (1998) Integrating three representation models in fuzzy multipurpose decision making based on fuzzy preference Relations. Fuzzy Set Syst 97:33–48
5. Chinese Calligraphy, http://www.zein.se/patrick/chinen9p.html
6. Choi JH, Im DH, Lee HY, Oh JT, Ryu JH, Lee HK (Nov. 2009) "Color laser printer identification by analyzing statistical features on discrete wavelet transform," *in Proceedings of IEEE International Conference on Image Processing (ICIP)*, pp. 1505–1508
7. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20:273–297
8. Cox IJ, Kilian J, Leighton FT, Shamoon T (1997) Secure spread spectrum watermarking for multimedia. IEEE Transactions on Image Processing 6:1673–1687
9. Cox IJ, Miller ML, Bloom JF, Fridrich J, Kaler T (2008) Digital Watermarking and Steganography. Morgan Kaufmann Publishers, Elsevier
10. Daubechiesl I, Sweldens W (1998) Factoring wavelet transforms into lifting steps. J Fourier Anal Appl 4(3):247–269
11. Elder JH, Zucker SW (1998) Local scale control for edge detection and blur estimation. IEEE Trans Pattern Anal Machine Intell 20:699–716
12. Haralick RM, Shanmugam K, Dinstein I (1973) "Textural features for image classification,". IEEE Transactions on Systems, Man, and Cybernetics SMC–3(6):610–621. doi:610
13. http://apps.carleton.edu/curricular/asln/Chinese/
14. http://www.library.cornell.edu/preservation/tutorial/contents.html
15. http://www.moneydj.com/kmdj/news/newsviewer.aspx
16. http://www.zh.wikpedia.org/wiki/漢字
17. http://zh.wikipedia.org/wiki/以人口排列的语言列表
18. Huang HC, Fang WC (2010) Metadata-Based Image Watermarking for Copyright Protection. Simulation Modelling Practice and Theory 18(4):436–445
19. Khanna N, Mikkilineni AK, Martone AF, Ali GN, Chiu GTC, Allebach JP, Delp EJ (2006) A survey of forensic characterization methods for physical devices. Proceedings of Digital Forensic Research Workshop 3:S17–S28
20. Kundur D, Lin CY, Macq B, Yu H (June 2004) "Special issue on enabling security technologies for digital rights management," *in Proceedings of the IEEE*, pp. 879–882
21. Mikkilineni AK, Ali GN, Chiang PJ, Chiu GTC, Allebach JP, Delp EJ (2004) Signature-embedding in printed documents for security and forensic applications. Proceedings of the SPIE International Conference on Security, Steganography, and Watermarking of Multimedia Contents 5306:455–466
22. Mikkilineni AK, Chiang PJ, Ali GN, Chiu GTC, Allebach JP, Delp EJ (2005) Printer identification based on graylevel co-occurrence features for security and forensic applications. Proceedings of the SPIE International Conference on Security 5681:430–440
23. Pudil P, Novovicova J, Kittler J (1994) "Floating search methods in feature selection," *Pattern Recognition Letters*, pp.1119–1125
24. Ritchey PC, Rego VJ (2012) A Context Sensitive Tiling System for Information Hiding. Journal of Information Hiding and Multimedia Signal Processing 3(3):212–226
25. Stearns S (1976) "On selecting features for pattern classifiers," *In: 3rd International Conf. Pattern Recognition*. Coronado, California, pp.71–75
26. Talbot V, Perrot P, Murie C (Sep. 2006) "Inkjet printing discrimination based on invariant moments," *in Proceedings of the IS&T's NIP22: International Conference on Digital Printing Technologies*, pp. 427–431
27. Tsai MJ, Liu J, Wang CS, Chuang CH (May 2011) "Source Color Laser Printer Identification Using Discrete Wavelet Transform and Feature Selection Algorithms," *IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 2633–2636
28. Tsai MJ, Wang CS, Liu J, Yin JS (2012) Using Decision Fusion of Feature Selection in Digital Forensics for Camera Source Model Identification. Computer Standards & Interfaces 34:292–304
29. Vapnik V (2000) The Nature of Statistical Learning Theory, 2nd edn. Springer-Verlag, Inc., New York

30. Villasenor JD, Belzer B, Liao J (1995) Wavelet filter evaluation for image compression. IEEE Transactions on Image Processing 4(8):1053–1060
31. World Intellectual Property Organization (WIPO), http://www.wipo.int/
32. Yager RR (1998) "On ordered weighted averaging aggregation operators in multicriteria decision making,". IEEE Transactions on Systems, Man and Cybernetics 18:183–190
33. Zhu B, Wu J, Kankanhalli M (2003) "Print signatures for document authentication". CCS'03, Washington

Min-Jen Tsai received the B.S. degree in electrical engineering from National Taiwan University in 1987, the M.S. degree in industrial engineering and operations research from University of California at Berkeley in 1991, the engineer and Ph.D. degrees in Electrical Engineering from University of California at Los Angeles in 1993 and 1996, respectively. He served as a second lieutenant in Taiwan army from 1987 to 1989. From 1996 to 1997, he was a senior researcher at America Online Inc. In 1997, he joined the institute of information management at the National Chiao Tung University in Taiwan and is currently a full professor. His research interests include multimedia system and applications, digital right management, digital watermarking and authentication, digital forensic, enterprise computing for electronic commerce applications. Dr. Tsai is a member of IEEE, ACM and Eta Kappa Nu.

Jin-Shen Yin is currently a Ph.D. student at the institute of information management, National Chiao Tung University.

Imam Yuadi is currently a Ph.D. student at the institute of information management, National Chiao Tung University.

Jung Liu is currently a Ph.D. student at the institute of information management, National Chiao Tung University.

12. Tsai2014_Article_DigitalForensicsOfPrintedSourc.pdf

ORIGINALITY REPORT

14%	13%	8%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.cerias.purdue.edu Internet Source	3%
2	www.mysciencework.com Internet Source	2%
3	tudr.thapar.edu:8080 Internet Source	1%
4	knol.google.com Internet Source	1%
5	sdiwc.net Internet Source	1%
6	www.carleton.edu Internet Source	1%
7	engineering.purdue.edu Internet Source	1%
8	Submitted to Colorado Technical University Online Student Paper	<1%
9	Submitted to Calumet High School Student Paper	<1%
10	Jahanirad, Mehdi, Ainuddin Wahid Abdul Wahab, and Nor Badrul Anuar. "An evolution of image source camera attribution approaches", Forensic Science International, 2016. Publication	<1%
11	scottminer.rbind.io Internet Source	<1%

12	www.byui.edu Internet Source	<1 %
13	baadalsg.inflibnet.ac.in Internet Source	<1 %
14	www.jihmsp.org Internet Source	<1 %
15	www.researchgate.net Internet Source	<1 %
16	mdpi-res.com Internet Source	<1 %
17	it.canyonsdistrict.org Internet Source	<1 %
18	web.archive.org Internet Source	<1 %
19	archive.org Internet Source	<1 %
20	Pei-Ju Chiang. "Printer and Scanner Forensics: Models and Methods", Studies in Computational Intelligence, 2010 Publication	<1 %
21	Shen Yue. "Lifting wavelet transform based adaptive filter for active power filters", 2008 27th Chinese Control Conference, 07/2008 Publication	<1 %
22	Mikkilineni, Aravind K., Pei-Ju Chiang, Gazi N. Ali, George T. C. Chiu, Jan P. Allebach, Edward J. Delp III, and Ping W. Wong. "", Security Steganography and Watermarking of Multimedia Contents VII, 2005. Publication	<1 %
23	citeseerx.ist.psu.edu Internet Source	<1 %
	www.science.gov	

24	Internet Source	<1%
25	docs.lib.purdue.edu Internet Source	<1%
26	patentimages.storage.googleapis.com Internet Source	<1%
27	Megha Borole, Satish R. Kolhe. "A feature-based approach for digital camera identification using photo-response non-uniformity noise", International Journal of Computational Vision and Robotics, 2021 Publication	<1%
28	pdffox.com Internet Source	<1%
29	Khanna, N.. "A survey of forensic characterization methods for physical devices", Digital Investigation, 200609 Publication	<1%
30	Zhisheng Gao, Peng Shi, Hamid Reza Karimi, Zheng Pei. "A mutual GrabCut method to solve co-segmentation", EURASIP Journal on Image and Video Processing, 2013 Publication	<1%
31	docplayer.net Internet Source	<1%
32	tel.archives-ouvertes.fr Internet Source	<1%
33	Ali Reza Akoushideh, Asadollah Shahbahrami, Babak Mazloom-Nezhad Maybodi. "High performance implementation of texture features extraction algorithms using FPGA architecture", Journal of Real-Time Image Processing, 2012 Publication	<1%

34	Musrrat Ali, Chang Wook Ahn, Millie Pant. "An efficient lossless robust watermarking scheme by integrating redistributed invariant wavelet and fractional Fourier transforms", Multimedia Tools and Applications, 2017 Publication	<1 %
35	academic.oup.com Internet Source	<1 %
36	epdf.tips Internet Source	<1 %
37	www.igi-global.com Internet Source	<1 %

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		