Chapter 7

USING DCT FEATURES FOR PRINTING TECHNIQUE AND COPY DETECTION

Christian Schulze, Marco Schreyer, Armin Stahl and Thomas Breuel

Abstract The ability to discriminate between original documents and their photocopies poses a problem when conducting automated forensic examinations of large numbers of confiscated documents. This paper describes a novel frequency domain approach for printing technique and copy detection of scanned document images. Tests using a dataset consisting of 49 laser-printed, 14 inkjet-printed and 46 photocopied documents demonstrate that the approach outperforms existing spatial domain methods for image resolutions exceeding 200 dpi. An increase in classification accuracy of approximately 5% is achieved for low scan resolutions of 300 dpi and 400 dpi. In addition, the approach has the advantage of increased processing speed.

Keywords: Printing technique and copy detection, discrete cosine transformation

1. Introduction

Due to advances in digital imaging techniques, it is relatively simple to create forgeries or alter documents within a short timeframe. According to the American Society of Questioned Document Examiners (ASQDE), modern printing technologies are increasingly used to produce counterfeit banknotes [5] and forged documents [13].

Forensic document examiners are confronted with a variety of questions [9]: Who created the document? Which device created the document? What changes have been made to the document since it was originally produced? Is the document as old as it purports to be? As a result, a variety of sophisticated methods and techniques have been developed since Osborn and Osborn [16] noted that a document may have any one of twenty or more different defects that may not be observed unless one is specifically looking for them.

One challenge when examining machine-printed documents is to distinguish between laser-printed and photocopied documents. Kelly, *et al.* [9] note that "the analysis of a photocopy can run the entire gamut of instruments in a document examination laboratory." The difficulty can be ascribed to the fact that laser printing and photocopying are very similar in operation [6] – both techniques use indirect electrostatic digital imaging to transfer the printing substrate.

The photocopy process has at least two distinct phases: scanning the template document and printing the scanned content. Due to technical limitations of photocopier devices and/or physical conditions, a small amount of the original document information may be lost or altered. These imperfections are observable at the edges of characters in the form of light blurring [3]. The imperfections are more evident after the document is transformed into the frequency domain. This is because, in general, sharp transitions between character and non-character areas result in a larger number of high frequency values compared with the smooth transitions associated with blurred character edges.

This work presents a novel frequency domain approach for document printing technique recognition. The recognition system creates a unique fingerprint for each printing technology from the number and distribution of the frequencies contained in a document. Experimental results demonstrate that using discrete cosine transformation (DCT) coefficients and machine learning techniques can help distinguish between inkjet-printed and laser-printed documents and also detect first-generation photocopies at low scan resolutions.

2. Related Work

Kelly and Lindblom [9] have surveyed the major forensic document techniques applicable to questioned documents. Although the use of digital imaging techniques in forensic examinations of documents is relatively new, recent research efforts have demonstrated that these techniques are very useful for discriminating between non-impact printing techniques.

Print quality is a useful feature for identifying the printing technology used to create a document. Print quality metrics include line width, raggedness and over spray, dot roundness, perimeter and the number of satellite drops [15]. Other useful quality features include the gray level co-occurrence features for the printed letter "e" [14]; and texture features based on gray level co-occurrence and local binary map histograms used in conjunction with edge roughness features that measure line edge roughness, correlation and area difference for printed document charac-

ters [11]. Schulze, *et al.* [17] have evaluated these features in the context of high-throughput document management systems.

The methodology proposed by Tchan [18] is very similar to our approach. Documents are captured at low resolution and printing technologies are distinguished by measuring edge sharpness, surface roughness and image contrast. However, Tchan has only experimented with documents containing squares and circles, not typical office documents.

Color laser printers add another dimension to the document feature space, which has led to the development of alternative detection methods. By evaluating hue component values within the HSV color space, it is possible to distinguish between different printing substrates and, thus, printing techniques [6]. Additionally, the yellow dot protection patterns on documents printed by color laser printers (that are nearly invisible to the unaided human eye) can be used for printer identification [12, 19]. Indeed, the distinctive dot pattern is directly related to the serial number of a particular laser printer or photocopier.

In addition to the printing technology, the physical characteristics of a printing device often leave distinctive fingerprints on printed documents. For example, the manner in which the spur gears hold and pass paper through printing devices can be used to link questioned documents to suspected printers [2].

Gupta, *et al.* [7] have presented a structured methodology for detecting the scanner-printer combination used to create tampered documents based on document image imperfections. They measure the overall similarity and the similarity in coarse document areas between the original document and the tampered document. The standard deviation and average saturation of the image noise are then computed. Experimental results indicate that this method is very effective.

However, none of the techniques in the literature use frequency domain features to identify the printing technique used to create a questioned document. Also, we were unable to find any research articles focusing on the detection of photocopied documents in real-world scenarios.

3. Printing Process Characteristics

In general, printing is a complex reproduction process in which ink is applied to a printing substrate in order to transmit information in a repeatable form using image-carrying media (e.g., a printing plate) [10].

Inkjet printers use a printhead to emit tiny ink droplets onto the printing paper. As the paper is drawn through the printer, the printhead moves back-and-forth horizontally and usually transfers the ink directly to the paper. Ink deposition is digitally controlled by the printer

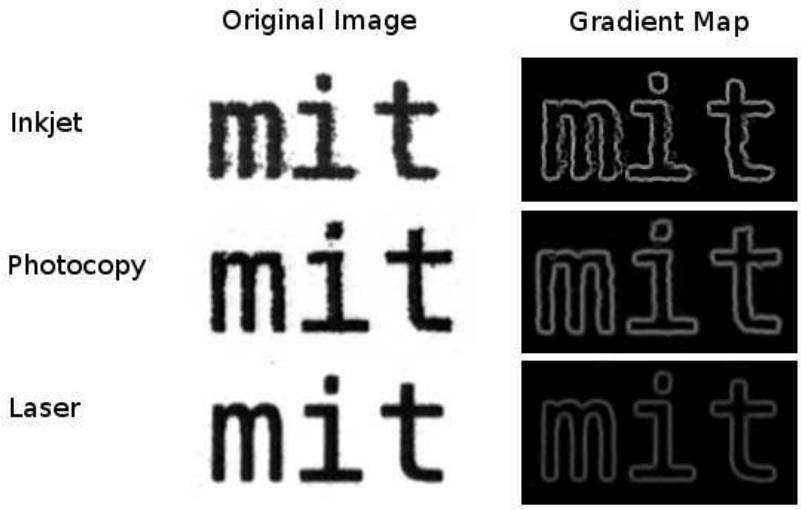

Figure 1. Representative plots of printing technologies.

firmware. Ink is sprayed onto the paper so that multiple gradients of dots accumulate to form an image with variable color tones.

The majority of laser printers and photocopiers engage electrophotographic printing technology. The underlying concept is to generate a visible image using an electrostatic latent image created by surface charge patterns on a photoconductive surface. The main difference between laser printing and photocopying is the image source. In the case of photocopying, the image has to be scanned prior to printing. During the scanning procedure, a scanhead is moved slowly across the illuminated document. The scanhead divides the image into microscopic rows and columns and measures how much light reflects from each individual row-column intersection. The charge collected by each photodiode during the scanning process is proportional to the reflection of a specific area of the paper. The amount of reflectance is then recorded as a dot or picture element (pixel).

Edge sharpness is the most effective feature for visually discriminating between high resolution scans of documents created by different printing techniques. Figure 1 shows representative plots of inkjet, photocopy and laser printing technologies and their corresponding gradient maps. The images are captured at a scanning resolution of 2,400 dpi. The left column shows the original scans while the right column displays the corresponding gradient maps obtained using a Prewitt gradient filter. Note the differences in edge sharpness and edge roughness for the different printing techniques.

3.1 Edge Sharpness and Contrast

The gradient at an image pixel measures the image intensity change in the direction of the highest intensity change at that particular pixel. A sharp edge and, therefore, a high intensity change results in a skin gradient. Examination of the gradient images for the three printing techniques in Figure 1 reveals that the laser-printed document image is characterized by sharp transitions between character and non-character areas. In contrast, the photocopied and the inkjet-printed images show a tendency towards smoother and blurred character edges. This feature is due to printing substrate diffusion in the case of the inkjet-printed image and light diffusion during scanning in the case of the photocopied document.

3.2 Edge Roughness and Degradation

Edge roughness and degradation denote the divergence of the printed character shape from the original template character shape. The gradient maps in Figure 1 reveal different degrees of character shape degradation. A high degree of edge roughness is observed for the inkjet-printed document; in contrast, little edge roughness is seen for the laser-printed document. As with edge sharpness, edge roughness is determined by several factors such as printer resolution, dot placement accuracy, rendering algorithms and interactions between colorant and paper.

4. Frequency Domain Comparisons

Figure 2 shows the results of a pairwise printing technique comparison in the frequency domain. The DCT [1] is first calculated for images produced by each document creation technique; the DCT frequency spectrum results are then averaged for each document creation technique and compared. The left image shows the result obtained by comparing the average laser-printed spectrum (white) versus the average spectrum of the photocopied documents (black). The middle image shows the average laser-printed spectrum (white) versus the average inkjet-printed spectrum (black). The right image shows the average inkjet-printed spectrum (white) versus the average spectrum of the photocopied documents (black). The frequency comparison spectra images in Figure 2 reveal clear differences between the techniques. Comparing the average DCT coefficient spectrum of laser-printed documents with those produced by the other two printing technologies shows a radial symmetric pattern. This pattern is not as distinctive, but is still recognizable when comparing the spectra of the inkjet-printed and photocopied documents.

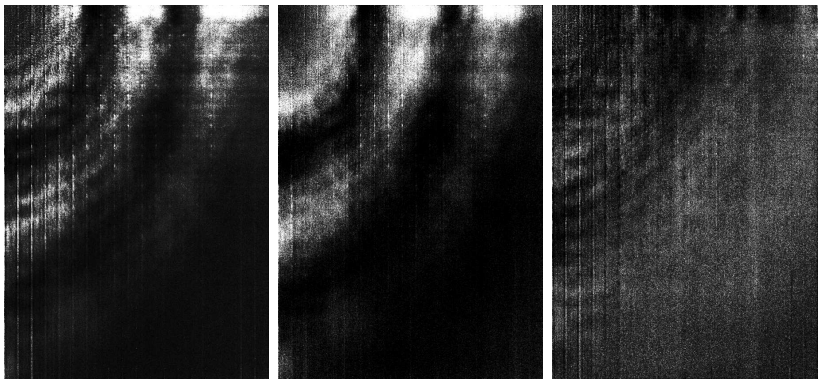

Figure 2. Frequency domain comparison of printing technique classes.

The difference can be traced to the relationship between the spatial and frequency domains. Sharp edges in the spatial domain yield increased DCT coefficient values for high frequencies. Compared with characters in photocopied and inkjet-printed documents, laser-printed characters have sharper transitions between character and non-character regions. This property is evident in the DCT coefficient values.

Comparison of the images in Figure 2 also indicates that the DCT coefficients corresponding to horizontal and vertical frequencies exhibit a highly discriminant behavior. This is due to the properties of Latin fonts for which large numbers of sharp transitions exist between character and non-character regions in the horizontal and vertical directions. These sharp transitions are very fragile to edge blurring induced by inkjet printing and photocopying. As a result, both printing techniques show a tendency to smaller coefficient values in the high frequencies of the horizontal and vertical components of the DCT spectrum.

5. DCT Coefficient Distribution Analysis

The frequency difference images in Figure 2 and the printing technique characteristics discussed above underscore the idea that printing techniques are distinguishable according to their frequency spectra. This especially holds for frequency sub-band coefficients corresponding to vertical and horizontal image intensity variations in the spatial domain.

A two-step procedure is used to determine the DCT coefficient distribution and the distribution strength within the frequency spectrum sub-bands for an arbitrary document image. First, the DCT coefficients of a particular sub-band are extracted from the frequency spectrum. Next, statistical features are calculated from the sub-band coefficients.

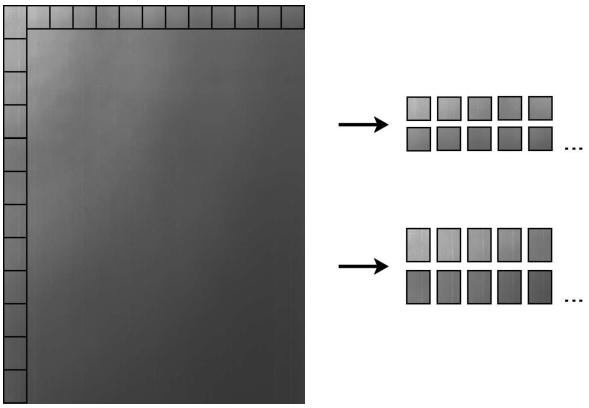

Figure 3. Frequency sub-band coefficient detection and extraction.

- **Frequency Sub-Band Extraction:** As shown in Figure 2, a high discriminative character is evident in the horizontal and vertical frequency sub-bands of the spectra. Therefore, these frequency sub-band coefficients are obtained from the spectrum of each document by extracting a set of horizontal and vertical sub-band boxes as illustrated in Figure 3. The left-hand side figure shows the detected horizontal and vertical frequency sub-band coefficients; the right-hand side shows the extracted frequency sub-band coefficients. The aspect ratio of the sub-band boxes is set to $\frac{1}{\sqrt{2}}$ in order to capture the frequency distortion caused by the aspect ratio of the document images.

- **Statistical Feature Extraction:** Let k be the number of frequency sub-band boxes box_i obtained from the normalized frequency spectrum $F'(u,v)$ of a document. To obtain the unique horizontal and vertical frequency sub-band pattern of the document, the mean and standard deviation of the coefficients are computed according to the following equations for each sub-band box_i:

$$\begin{aligned} \mu(box_i) &= \frac{1}{MN}\sum_{m=0}^{M}\sum_{n=0}^{N}F'(m,n) \\ \sigma(box_i) &= (\frac{1}{MN}\sum_{m=0}^{M}\sum_{n=0}^{N}(F'(m,n)-\mu)^2)^{\frac{1}{2}} \end{aligned}$$

where $m, n \in box_i$ and M, N indicate the size of box_i. This produces a $2k$-dimensional feature vector for each document image.

6. Evaluation

This section describes the experimental setup used to evaluate our approach and the results obtained.

6.1 Experimental Setup

Existing document image databases (e.g., UW English Document Image Database I-III, Medical Article Records System (MARS), MediaTeam Oulu Document Database, Google 1,000 Books Project) do not include annotations on the printing techniques used to create the documents. Therefore, a document image database annotated with the needed ground truth information was created. As in previous works [11, 17], the "Grünert" letter in the 12 pt "Courier New" normal font with a 12 pt line height was used as a template to create the ground truth document image database. This template document implements the DIN-ISO 1051 standard. The database used consisted of 49 laser and 14 inkjet printouts and 46 photocopied documents. Every document in the database was created by a different printer or photocopier, covering all the major manufacturers.

To create a realistic evaluation scenario, half of the photocopied documents were generated using laser-printed templates while the other half were based on inkjet-printed templates. Only first-generation photocopies were added to the database.

All the documents were scanned at resolutions of 100 dpi, 200 dpi, 300 dpi, 400 dpi and 800 dpi. A Fujitsu 4860 high-speed scanning device was used for scan resolutions lower than 400 dpi. This scanner is designed for high-throughput scanning and, therefore, the maximal scan resolution is limited to 400 dpi. In order to test the DCT feature performance at higher resolutions, an EPSON 4180 device was used to produce 800 dpi scans. All the document images obtained were stored in the TIFF file format to avoid further information loss.

Classification Methodology We classified the printing techniques of the scanned document images using support vector classification utilities provided by the LibSVM library [4]. The support vector machine (SVM) classification was performed using a radial-basis function kernel with optimized parameters. Optimal kernel parameters for C and γ were obtained by a coarse grid search in the SVM parameter space within the intervals $C = [2^{-5},2^{15}]$ and $\gamma = [2^{-15},2^{3}]$ as suggested in [8].

Table 1. Classification accuracy results.

dpi	DCT [%]	$\mathbf{v_{DCT}}$ $[\frac{\mathbf{p}}{\mathbf{min}}]$	Gradient [%]	$\mathbf{v_{grad}}$ $[\frac{\mathbf{p}}{\mathbf{min}}]$
100	72.64	44.78	75.47	50.85
200	80.95	15.50	80.00	14.60
300	85.85	6.42	81.13	6.26
400	92.92	2.22	88.23	2.01
800	99.08	0.53	97.17	0.78

Performance Evaluation To evaluate the prediction capability of the extracted features without losing the generalization ability of the learned model, we applied a stratified 10-fold cross validation. To evaluate the classification performance of the applied feature data, we calculated (for each classification trial) the accuracy based on the percentage of correctly classified documents in the testing data set. The average accuracy across all trials was then computed to give the overall result of the stratified cross-validation. The results were compared with a gradient feature based on the work of Tchan [18]. We selected this feature because it was the best performing feature from the set of implemented spatial domain features.

6.2 Experimental Results

Table 1 presents the classification accuracy results for printing techniques using the DCT feature and the best performing spatial domain feature. The processing speed is given in pages per minute. The values in Table 1 show that the DCT feature produces better classification accuracy rates than the gradient feature (the best performing spatial domain feature) for scan resolutions $\geq$ 200 dpi. Significantly better performance of $\approx 5\%$ is seen with the DCT feature for resolutions of 300 dpi and 400 dpi. This observation is important for high-throughput systems because the maximal scan resolution of these systems is usually limited to 400 dpi. Note that the classification accuracy of the DCT feature for the three classes (inkjet, laser and copy) at 400 dpi exceeds 90%.

Figure 4 presents the classification accuracy (top) and processing speed (bottom) for DCT and gradient-based features. The throughput achieved for resolutions of 200 to 400 dpi is larger for the DCT feature, while at 100 dpi and 800 dpi the gradient-based feature has larger values. The speed measurements were made on a system equipped with an Intel Core 2 Duo T7300 (2 GHz) processor and 1 GB memory using a single core.

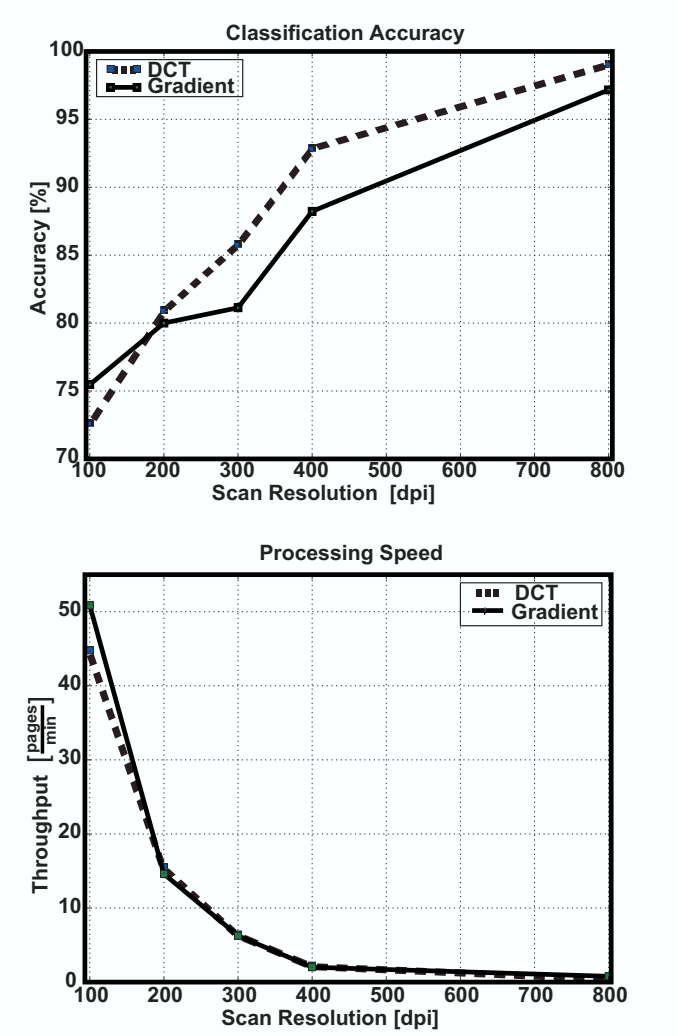

Figure 4. Classification accuracy and processing speed.

7. Conclusions

This paper has presented a novel approach for document printing technique recognition using features calculated in the frequency domain. It demonstrates that the frequency distribution and number of frequencies within a transformed document image are directly related to the edge and noise characteristics visible in the spatial domain. Therefore, the DCT coefficients obtained by transforming a document image can serve as a fingerprint for the printing technology used to create the document.

Experimental results verify that the frequency domain approach outperforms spatial domain techniques. The results are particularly significant in the case of document images scanned at resolutions from 200 dpi to 800 dpi. Moreover, no major increase in the processing time is observed despite the fact that a transformation of the scanned document into the frequency domain is necessary to extract DCT based features. Consequently, the approach presented in this paper is also well-suited to high-throughput document processing scenarios.

References

[1] N. Ahmed, T. Natarajan and K. Rao, Discrete cosine transform, *IEEE Transactions on Computers*, vol. 23(1), pp. 90–93, 1974.

[2] Y. Akao, K. Kobayashi and Y. Seki, Examination of spur marks found on inkjet-printed documents, *Journal of Forensic Science*, vol. 50(4), pp. 915–923, 2005.

[3] H. Baird, The state of the art of document image degradation modeling, *Proceedings of the Fourth International Association for Pattern Recognition Workshop on Document Analysis Systems*, pp. 1–16, 2000.

[4] C. Chang and C. Lin, LIBSVM: A Library for Support Vector Machines, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan (www.csie.ntu.edu.tw/∼cjlin/libsvm).

[5] J. Chim, C. Li, N. Poon and S. Leung, Examination of counterfeit banknotes printed by all-in-one color inkjet printers, *Journal of the American Society of Questioned Document Examiners*, vol. 7(2), pp. 69–75, 2004.

[6] H. Dasari and C. Bhagvati, Identification of printing process using HSV colour space, *Proceedings of the Seventh Asian Conference on Computer Vision*, pp. 692–701, 2006.

[7] G. Gupta, R. Sultania, S. Mondal, S. Saha and B. Chanda, A structured approach to detect the scanner-printer used in generating fake documents, *Proceedings of the Third International Conference on Information Systems Security*, pp. 250–253, 2007.

[8] C. Hsu, C. Chang and C. Lin, A Practical Guide to Support Vector Classification, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan (www.csie.ntu.edu.tw/∼cjlin/papers/guide/guide.pdf), 2003.

[9] J. Kelly and B. Lindblom (Eds.), *Scientific Examination of Questioned Documents*, CRC Press, Boca Raton, Florida, 2006.

[10] H. Kipphan, *Handbook of Print Media*, Springer, Heidelberg, Germany, 2001.

[11] C. Lampert, L. Mei and T. Breuel, Printing technique classification for document counterfeit detection, *Proceedings of the International Conference on Computational Intelligence and Security*, pp. 639–644, 2006.

[12] C. Li, W. Chan, Y. Cheng and S. Leung, The differentiation of color laser printers, *Journal of the American Society of Questioned Document Examiners*, vol. 7(2), pp. 105–109, 2004.

[13] J. Makris, S. Krezias and V. Athanasopoulou, Examination of newspapers, *Journal of the American Society of Questioned Document Examiners*, vol. 9(2), pp. 71–75, 2006.

[14] A. Mikkilineni, P. Chiang, G. Ali, G. Chiu, J. Allebach and E. Delp, Printer identification based on gray level co-occurrence features for security and forensic applications, *Proceedings of the SPIE*, vol. 5681, pages 430–440, 2005.

[15] J. Oliver and J. Chen, Use of signature analysis to discriminate digital printing technologies, *Proceedings of the International Conference on Digital Printing Technologies*, pp. 218–222, 2002.

[16] A. Osborn and A. Osborn, Questioned documents, *Journal of the American Society of Questioned Document Examiners*, vol. 5(1), pp. 39–44, 2002.

[17] C. Schulze, M. Schreyer, A. Stahl and T. Breuel, Evaluation of gray level features for printing technique classification in high-throughput document management systems, *Proceedings of the Second International Workshop on Computational Forensics*, pp. 35–46, 2008.

[18] J. Tchan, The development of an image analysis system that can detect fraudulent alterations made to printed images, *Proceedings of the SPIE*, vol. 5310, pp. 151–159, 2004.

[19] J. Tweedy, Class characteristics of counterfeit protection system codes of color laser copiers, *Journal of the American Society of Questioned Document Examiners*, vol. 4(2), pp. 53–66, 2001.