A PATTERN CLASSIFICATION FRAMEWORK FOR THEORETICAL ANALYSIS OF COMPONENT FORENSICS

Ashwin Swaminathan, Min Wu and K. J. Ray Liu

Electrical and Computer Engineering Department, University of Maryland, College Park

ABSTRACT

Component forensics is an emerging methodology for forensic analysis that aims at estimating the algorithms and parameters in each component of a digital device. This paper proposes a theoretical foundation to examine the performance limits of component forensics. Using ideas from pattern classification theory, we define formal notions of identifiability of components in the information processing chain. We show that the parameters of certain device components can be accurately identified only in controlled settings through semi non-intrusive forensics, while the parameters of some others can be computed directly from the available sample data via complete non-intrusive analysis. We then extend the proposed theoretical framework to quantify and improve the accuracies and confidence in component parameter identification for several forensic applications.

Index Terms – Component forensics, pattern classification, visual sensors, semi non-intrusive forensics.

1. INTRODUCTION

Digital imaging technologies have undergone tremendous growth in recent decades. The resolution and quality of imaging devices have been improving steadily, and at the same time, the cost of the imaging devices have been declining, making them increasingly popular for day-to-day use. Digital images and videos captured by such devices have been used in a number of applications ranging from military and law enforcement to free-lance consumer photography. On the other hand, digital imaging technologies have also been used for illicit applications. For example, an increasing number of movies have been re-shot with camcorders directly from the theater where they are screened, and sold in the market. This kind of piracy incurs a significant loss to the copyright industry. Complementary to watermarking and fingerprinting technologies that help track such illegal reproduction, forensic analysis can help to trace the origin and authenticity of digital data. This bootlegging example raises a number of forensic questions such as what kind of device (*e.g.,* camera or camcorder and what brand/model) was used to capture the data? Was the image/video recorded from a display device and if so, what kind of display device (*e.g.,* flat-screen or projector) was used? Further, what kinds of legitimate processing and undesired alteration have been applied to the image/video since it has left the device? Answers to such forensic questions would facilitate tracing both the person who illegally captured the video by tracing his/her camcorder and the theater from which the video was captured using its display characteristics.

Our recent work [1] has introduced *component forensics* as a methodology for forensic analysis. Component forensics aims at finding the algorithms and parameters employed in each component of the information processing chain to answer who has done what, when, where, and how. We have shown that the *intrinsic fingerprint* traces left behind in the final digital image by the different components of the imaging device could be used as evidence to estimate the component parameters and provide clues to answer forensic questions related to origin and authenticity of digital data. However, as the intrinsic fingerprint traces pass through the different parts of the information processing chain, some of them may be modified or destroyed and some others newly created. In the bootlegging example, some traces of the projector employed in the theater might be lost and new fingerprint traces about the camcorder itself might be inserted. Hence, the data obtained from the final camcorder alone may or may not help compute the parameters of the display device. This leads to further forensic questions as to what components are identifiable and what are not.

A component estimation framework for media forensics was introduced in [2]. Formal theoretical notions were defined to characterize the accuracies in estimating the parameters of several components in the information processing chain. However, in many forensic scenarios, additional side information is available and can be used to improve accuracies. For instance, in the bootlegging example, geographic constraints can be enforced to narrow down on a possible set of theaters (and their display parameters) from where the movie could have been illegally recorded using a camcorder. In the presence of such additional information, the component parameters could be found with a higher accuracy from among the available sample set of algorithms by reformulating the estimation problem as a classification problem. In this work, we develop a theoretical framework for media forensics under the assumption that the component parameters take values from a finite set of possible algorithms, and derive conditions under which a component is *forensically classifiable*. Building upon the proposed framework, we devise methods to improve the confidence of component parameter identification for several forensic applications.

This work complements the theoretical estimation framework proposed in [2] to provide a generalized theoretical framework for media forensics. Related prior work mostly aim at developing techniques for forensic analysis to estimate the parameters of the different components in the information processing chain. In literature, methods have been proposed to estimate in-camera processing such as color interpolation [1,3,4] and white balancing [5], and post-camera manipulations like resampling [6], lighting, luminance, brightness change [6], and JPEG compression [7]. These collection of prior art provides algorithms to estimate the parameters of many types of in-camera and post-camera processing. To our best knowledge, this present work along with [2] are the first ones to provide a theoretical framework to foster systematic analysis that can be applicable to many types of digital devices and their combinations.

The paper is organized as follows. The proposed theoretical analysis framework is described in Section 2. In Section 3, we illustrate this framework with a particular example from digital cameras and present methods to improve the confidence scores in parameter classification. Final conclusions are drawn in Section 4.

2. PROPOSED THEORETICAL FRAMEWORK

In this section, we introduce a theoretical framework for component forensics and examine the conditions under which the parameters of a component can be identified accurately. We define a

Email contact: {ashwins, minwu, kjrliu} @eng.umd.edu.

component as a basic unit in the information processing chain. For instance, the color filter array, color interpolation algorithms, and white balancing operations can be considered as different components in a digital camera. Each of these components may employ different kinds of algorithms and/or parameters in each instantiation of the device. Such differences can be employed for forensic analysis, for instance, to build robust camera identifiers to determine the brand/make of the camera used to capture the digital image [1].

Consider a system with N_c components $\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_{N_c}\}$. Let $\Re_x$ denote a set of all possible inputs that can be given to the k^{th} component $\mathcal{C}_k$, and let $\Re_y$ contain the corresponding outputs. Let $\Theta^k = \{\theta_1^k, \theta_2^k, \ldots, \theta_{N_a}^k\}$ denote the set of all possible algorithms/ techniques that could be employed in k^{th} component $\mathcal{C}_k$ of the system. In this work, we formulate the problem of component forensics as identifying the exact algorithm $\theta^k \in \Theta^k$ used by each of the processing blocks $\mathcal{C}_k$.

Estimating the value of the component parameter, θ^k, can be done in three different ways depending on the nature of the available inputs [1], namely, by intrusive, semi non-intrusive, and completely non-intrusive forensic analysis. In *intrusive* component forensics, the analysts have the device or system in hand and can break it apart. They can then study each component of the system separately, design appropriate inputs, examine the corresponding outputs, and make a decision as to which set of parameters was used in the component via classification approaches. That is,

Definition 1: A component $\mathcal{C}_k$ is said to be *intrusively classifiable* or *i-classifiable* if for each possible algorithm θ_i^k used by the component, and $\forall x \in \Re_x$,

$$p(\theta_i^k|y, x) \geq p(\theta_j^k|y, x) \quad \forall j \in \{1, 2, \ldots, N_a^k\}, \text{and } j \neq i,$$

and there exists at least one input $x^* \in \Re_x$ and its corresponding output y^* for which

$$p(\theta_i^k|y^*, x^*) > p(\theta_j^k|y^*, x^*) \ \forall j \in \{1, 2, \ldots, N_a^k\}, \text{and } j \neq i.$$

Here, N_a^k is the total number of possible algorithms for the component $\mathcal{C}_k$; x and y denote the corresponding input and output of the component, respectively, and are vectors of appropriate dimensions; and p denotes the probability distribution function. The forensic analyst can then employ maximum a posteriori estimation techniques [8] to identify the component parameters $\hat{\theta}^k$.

In addition to computing the parameters of the internal building blocks of the components, it is also important to know the confidence level on the parameter estimation result. A higher confidence value would increase the trustworthiness of the decision made by the forensic analyst in applications involving infringement/licensing to determine potential technology breach [1]; and also in cases involving tampering detection.

Definition 2: For an *i-classifiable* component, $\mathcal{C}_k$, with component parameters, θ_i^k, the confidence score $\gamma_i^k(x)$ for correct classification under the input x is defined by the difference between the likelihood of the correct decision and the maximum of the corresponding likelihoods of the making a wrong decision. Expressed mathematically,

$$\gamma_i^k(x) = p(\theta_i^k|y, x) - \max_{j=1,2,\ldots,N_a^k, j\neq i} p(\theta_j^k|y, x). \qquad (1)$$

As can be seen from the equation, the confidence score $\gamma_i^k(x)$ is a function of the input x and can be improved by selecting proper inputs. To facilitate discussions, let us define $\mathbf{q}^k(x) = [p(\theta_1^k|y, x), p(\theta_2^k|y, x), \ldots, p(\theta_{N_a^k}|y, x)]$. If for an input, x', $\mathbf{q}^k(x') = [0, \ldots, 1, \ 0, \ \ldots, 0]$ with 1 at the i^{th} location, the decision of choosing the i^{th} class is made with a very high confidence and $\gamma_i^k(x')$ equal to 1. On the other hand, if $\mathbf{q}^k(x'') = [\frac{1}{N_a^k} - \varepsilon, \ldots, \frac{1}{N_a^k} + \frac{N_a^k - 1}{N_a^k}\varepsilon, \ \frac{1}{N_a^k} - \varepsilon, \ \ldots, \frac{1}{N_a^k} - \varepsilon]$ where ε is a small positive real number, there is an almost equal probability that the given data sample comes from any of the N_a^k classes. In this case, the decision is made with a very low confidence with $\gamma_i^k(x'') \approx 0$. In this example, x' and x'' represent the best and the worst possible inputs for identifying the component parameters. For other inputs, x, the value of $\gamma_i^k(x)$ would lie in the interval $[0, \ 1]$, with a higher value indicating more confidence in the decision made. Motivated by this discussion, we define a notion of an optimal input:

Definition 3: An *optimal input* $\hat{x}_i^k$ to the k^{th} component of the system that employs the algorithm θ_i^k is defined as the one that maximizes the confidence score, i.e., $\hat{x}_i^k = \arg\max_{x \in \Re_x} \gamma_i^k(x)$. The corresponding confidence score, $\eta_i^{k(i-int)} = \gamma_i^k(\hat{x}_i^k)$, then represents the overall maximum confidence in *intrusively* classifying the parameters of $\mathcal{C}_k$.

In semi non-intrusive and completely non-intrusive forensics, analysts are not allowed to break open the device or system. In the scenario of *semi non-intrusive* forensics, the analysts have access to the system as a black box, and can design appropriate inputs to the system and collect the corresponding output data in order to analyze the processing techniques and compute the parameters of the individual components. To examine this scenario, we define $\phi_j = [\theta_{j_1}^1, \theta_{j_2}^2, \ldots, \theta_{j_{N_c}}^{N_c}]$ to represent the set of algorithms (and parameters) employed by the entire system. Assuming that the component parameters in the k^{th} component can take N_a^k possibilities, we have a total of $N_a = \prod_{k=1}^{N_c} N_a^k$ possible algorithm choices for the system. The task for the forensic analyst is now reduced to finding which of these N_a algorithms is used by the system in question.

Definition 4: A system is said to be *semi non-intrusively classifiable* or *s-classifiable* if for each possible algorithm ϕ_i used by the component,

$$p(\phi_i|y, x) \geq p(\phi_j|y, x) \quad \forall j \in \{1, 2, \ldots, N_a\}, \text{and } j \neq i,$$

and there exists at least one input $x^* \in \Re_x$ and its corresponding output y^* such that

$$p(\phi_i|y^*, x^*) > p(\phi_j|y^*, x^*) \ \forall j \in \{1, 2, \ldots, N_a\}, \text{and } j \neq i.$$

Here, x and y denote the inputs and its corresponding outputs, respectively, of the overall system. The confidence in correct identification for this scenario can be defined similar to (1).

In the *completely non-intrusive* forensics scenario, the forensic analyst is provided only with some sample data produced by the device or system and does not have access to nor other knowledge about its inputs. In this case, we can define:

Definition 5: A system is said to be *non-intrusively classifiable* or *n-classifiable* if for each possible algorithm ϕ_i used by the component, and all possible outputs $y \in \Re_y$,

$$p(\phi_i|y) \geq p(\phi_j|y) \quad \forall j \in \{1, 2, \ldots, N_a\}, \text{and } j \neq i,$$

and there exists at least one input $x^* \in \Re_x$, such that the corresponding output, y^*, satisfies

$$p(\phi_i|y^*) > p(\phi_j|y^*) \quad \forall j \in \{1, 2, \ldots, N_a\}, \text{and } j \neq i.$$

The corresponding confidence score for a system to be non-intrusively classifiable given the output y^* when the actual algorithm employed is ϕ_i is given by

$$\eta_i^{(n-int)} = p(\phi_i|y^*) - \max_{j=1,2,\ldots,N_a^k, j\neq i} p(\phi_j|y^*).$$

We now establish the following results. Due to space constraints, the proofs have been omitted from this paper.

Lemma 1: If a system is *n-classifiable*, then it is *s-classifiable*.

This lemma suggests that if a component is non-intrusively classifiable, then its parameters can also be identified semi non-intrusively. Further, we can show that the average confidence values obtained using semi non-intrusive analysis is greater than or equal to the ones obtained via completely non-intrusive analysis, *i.e.,* for all possible algorithms ϕ_i, $\eta_i^{(s-int)} = \max_{x \in \Re_x} \gamma_i(x) \geq \eta_i^{(n-int)}$. This result follows from the fact that semi non-intrusive forensics provides more control to the forensic analyst who can design better inputs to improve the overall performance. Similarly,

Lemma 2: If a system is *s-classifiable*, then each of its components are *i-classifiable*.

This result follows intuitively from the fact that in intrusive forensics, the analyst can isolate each component separately and compute its parameters with a higher control over the experimental setup. The converse of Lemma 2 is not generally true. To examine the conditions under which an *i-classifiable* component is *s-classifiable*, we introduce the notion of an *ϵ-consistent component*. A component is said to be *ϵ-consistent* if the following two conditions are satisfied:

- for all outputs y_1 and y_2 with $d_Y(y_1, y_2) \leq \epsilon$, the estimates of the corresponding inputs x_1 and x_2 satisfy $d_X(x_1, x_2) \leq \epsilon$, where d_X and d_Y are appropriately chosen distance metrics in the input and the output space, respectively; and
- for all inputs x_1 and x_2 with $d_X(x_1, x_2) \leq \epsilon$, the estimates of the corresponding outputs y_1 and y_2 satisfy $d_Y(y_1, y_2) \leq \epsilon$.

We now have the following theorem:

Theorem 1: If all the components in a system are ϵ-consistent and the k^{th} component with parameter θ_i^k is *i-classifiable* with a confidence score $\eta_i^{k(i-int)}$, then the k^{th} component is *s-classifiable* with confidence score $\eta_i^{k(s-int)}$ approximately given by

$$\eta_i^{k(s-int)} \approx \eta_i^{k(i-int)} - 2(N_c - 1)\epsilon \left| \frac{\partial \gamma_i^k(x)}{\partial x} \right|_{x = \hat{x}_i^k}. \qquad (2)$$

The above expression is obtained by considering the first-order approximation of the confidence score and dropping the higher-order terms. Theorem 1 gives the conditions under which the knowledge about the intrusive forensics can be extended to semi non-intrusive forensics. The theorem also suggests that $\eta_i^{k(i-int)} \geq \eta_i^{k(s-int)}$, and therefore the confidence score for parameter identification from semi non-intrusive forensics is lower than (or at most equal to) the ones that can be attained from intrusive forensics. It can be further shown that the equality is attained only when all the components are 0-consistent. This result is expected because intrusive forensic methodology gives more control than semi non-intrusive forensics, as the forensic analyst can break the device or system open to examine each of its individual components in greater detail. On the other hand, in the case of semi non-intrusive forensic analysis, the analyst would need to come up with good inputs to be given to the overall system and study the interactions between various system components based on the overall input/output response.

3. CASE STUDIES WITH DIGITAL CAMERAS

In this section, we illustrate the proposed theoretical framework using illustrative examples from digital cameras and the proposed techniques can be extended to other kinds of digital devices and combinations of digital devices such as in the scenario of the bootlegging example.

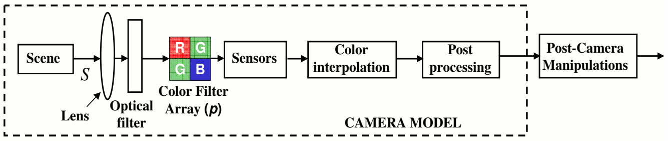

Fig. 1. Information processing chain in digital cameras showing its individual components

Mathematical Model of Camera Components: Fig. 1 shows the information processing chain in digital cameras. Rays of light from the scene pass through the lens and the optical filter and are finally recorded by the sensors. Most cameras employ a color filter array (CFA) to sample the data from the real world scene [1,3]. The CFA is a thin film on the sensors and selectively allows a certain component of light to pass through them to the sensors. Due to this selective sampling, some pixels values in the image are directly obtained from the sensor and the remaining pixels are interpolated using the values captured from the surrounding neighborhood [3]. In our recent work [1], we show that the camera's color interpolation module can be approximated to be linear in three different regions of the image, divided based on the local gradient values. Let x be the input to the camera's color interpolation module. The corresponding output r after color interpolation satisfies $r(m, n, c) = x(m, n, c)$, for the pixels that are directly obtained from the sensor, and $r(m, n, c) = \sum_{k,l} \alpha(k, l, c) x(m - k, n - l, c)$ for the interpolated pixels in each texture region with $\alpha(., ., .)$ denoting the corresponding filter coefficients.

After color interpolation, the image r goes through a post-processing stage where white balancing and color correction are performed to give $z(m, n, c) = \sum_{j=1}^{3} \beta(c, j) r(m, n, j)$, for $c = 1, 2, 3$. Finally, the image z may be JPEG compressed to reduce storage space. Compression can be modelled as quantization in the DCT domain, and can be represented as additive noise ζ in the pixel domain. The final camera output image is $y = z + \zeta$.

Camera Component Analysis using the Proposed Framework: Here, we examine the conditions under which the different camera components are identifiable under the three forensic analysis scenarios. Such camera components as color interpolation, white balancing, and JPEG compression are separately *i-classifiable*. More specifically, given the input and the output to each of these components, the interpolation parameters α in each texture region, and the white balancing parameters β can be obtained by solving a set of linear equations. Such approaches have been employed to estimate the interpolation coefficients [1] and the white balancing parameters [5]. To obtain the parameters of JPEG compression, statistical analysis based on binning techniques on the DCT coefficients have been employed [7].

Next, we consider the camera components together and examine the conditions under which they are *s-classifiable* and *n-classifiable*, respectively. Combining the equations for r and z, the input-output relationship for the interpolated pixels satisfying the camera model can be written as

$$y(m, n, c) = \sum_{j=1}^{3} \sum_{k,l} \alpha(k, l, j) \beta(c, j) x(m-k, n-l, j) + \zeta(m, n, c), \qquad (3)$$

Concatenating all the elements of $y(m, n, c)$ to form a vector $\mathbf{y}$ and representing (3) in matrix form, we obtain $\mathbf{y} = A_{\alpha\beta}\mathbf{x} + \mathbf{n}$. Here,

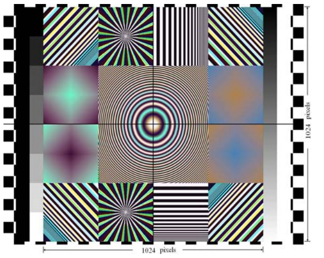

Fig. 2. Heuristic input pattern for semi non-intrusive forensics

$A_{\alpha\beta}$ denotes a matrix of appropriate dimension formed using the component parameters α and β, and $\mathbf{n}$ represents the compression noise satisfying $\mathcal{N}(0, \Sigma_n)$. The probability distribution of $\mathbf{y}$ given the component parameters $A_{\alpha\beta}$ and $\mathbf{x}$ is then Gaussian distributed with mean $\mu_y = A_{\alpha\beta}\mathbf{x}$ and covariance matrix $\Sigma_y = \Sigma_n$.

Estimating the camera parameters $\phi = A_{\alpha\beta}$ and/or its individual component parameters (α and β) provide valuable evidence for forensic analysis [1]. In the absence of compression noise, $\Sigma_n = 0$ and the set of equations in (3) can be solved to obtain $A_{\alpha\beta}$, and the results classified to belong to one of the possible algorithms in the algorithm space. Therefore, color interpolation and white balancing are *s-classifiable* in the absence of noise. Further, in the absence of noise, both components are 0-consistent and from Theorem 1, $\eta^{k(s-int)} = \eta^{k(i-int)}$. This suggests that breaking the device open to estimate the color interpolation or white balancing parameters does not give better results. In the presence of additive noise or compression, $\Sigma_n \neq 0$, and hence semi non-intrusive analysis does not provide the same confidence as intrusive forensics. Further, the confidence in parameter identification via semi non-intrusive forensics depends on the choice of the input.

Applications to Input Pattern Design for Semi Non-Intrusive Forensics: In this part, we show that with appropriate design, the confidence score for semi non-intrusively estimating the color interpolation and the white balancing coefficients can be increased. As shown earlier, the camera output $\mathbf{y}$ follows $\mathcal{N}(A_{\alpha\beta}\mathbf{x}, \Sigma_n)$. Substituting for the probability distribution of $\mathbf{y}$ and computing the confidence score, we can show that the *optimal* input for camera component forensics is the one that maximizes the distance, $||(A_{\alpha\beta}(i) - A_{\alpha\beta}(j))\mathbf{x}||$. Here, $A_{\alpha\beta}(i)$ and $A_{\alpha\beta}(j)$ correspond to two different possible values for the $A_{\alpha\beta}$ from the algorithm space. It can be shown that the solution for this maximization problem, $\hat{\mathbf{x}}$, is along the direction of the eigenvector corresponding to the largest eigenvalue of the matrix $(A_{\alpha\beta}(i) - A_{\alpha\beta}(j))$.

Based on the above theory, we design a possible candidate input image heuristically [5] to compute the parameters of camera components as shown in Fig. 2. To simulate the camera capture process, the input image is interpolated using two different interpolation techniques: bicubic that does not adapt to image content, and the adaptive color plane interpolation method [3] that adapts to image gradient values. The interpolation coefficients are estimated and used as an input to a two-class support vector machine (SVM) classifier for identification. This SVM has been trained with the coefficients obtained from natural images correspondingly interpolated with each of the same two different techniques. We study the robustness in parameter estimation under JPEG compression. In Fig. 3, we plot the confidence values obtained on classification under different quality levels of JPEG compression both for the designed pattern and for natural images. We notice from the figure that as the JPEG quality factor reduces and compression noise becomes stronger, the confidence of correctly identifying the interpolation coefficients reduces. Additionally, we observe that the confidence score obtained with the designed pattern is higher than the average scores obtained with natural images; demonstrating the superiority of designed pattern for semi non-intrusive analysis.

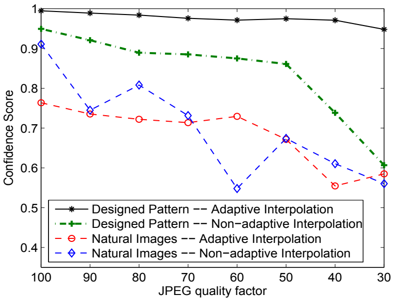

Fig. 3. Confidence score as a function of JPEG quality factor for (a) natural images (b) designed pattern

4. CONCLUSIONS

In this paper, we have developed a theoretical framework for media forensics for components with a finite number of possibilities in the parameter space. The proposed framework employs ideas from pattern classification theory to answer forensic questions about what components and processing operations are classifiable and what are not. We have define formal notions of identifiability of components under different scenarios, and have quantified the confidence in which the component parameters can be computed in each case. We have shown that the confidence in identifying the component parameters depends on the nature of available inputs and testing conditions, and that intrusive forensics gives higher confidence than semi non-intrusive forensics and semi non-intrusive analysis is better than completely non-intrusive scenario. We then apply the theoretical framework to design good inputs for semi non-intrusive forensics; and show that the confidence in parameter identification can be improved via such an approach. The proposed theoretical model can also be extended to study post-device processing operations such as tampering, and to provide a theoretical foundation for media forensics.

5. REFERENCES

[1] A. Swaminathan, M. Wu, and K. J. Ray Liu, "Non-Intrusive Component Forensics of Visual Sensors Using Output Images," in *IEEE Trans. on Info. Forensics and Security*, vol. 2, no. 1, pp. 91-106, March 2007.

[2] A. Swaminathan, M. Wu, and K. J. R. Liu, "A Component Estimation Framework for Information Forensics," *IEEE Workshop on Multimedia Signal Processing*, pp. 397-400, Crete, Greece, October 2007.

[3] A. C. Popescu and H. Farid, "Exposing Digital Forgeries in Color Filter Array Interpolated Images," *IEEE Trans. on Signal Processing*, vol. 53, no. 10, part 2, pp. 3948–3959, October 2005.

[4] S. Bayram, H. T. Sencar, and N. Memon, "Improvements on Source Camera-Model Identification Based on CFA Interpolation," *Proc. of the WG 11.9 Intl. Conf. on Digital Forensics*, Orlando, FL, January 2006.

[5] A. Swaminathan, M. Wu, and K. J. R. Liu, "Optimization of Input Pattern for Semi Non-Intrusive Component Forensics of Digital Cameras," *Proc. of the ICASSP*, Honolulu, HI, April 2007.

[6] A. C. Popescu and H. Farid, "Statistical Tools for Digital Forensics," *Intl. Workshop on Info. Hiding*, Toronto, Canada, May 2004.

[7] J. Lukas and J. Fridrich, "Estimation of Primary Quantization Matrix in Double Compressed JPEG Images," *Proc. of the DFRWS*, Aug. 2003.

[8] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, John Wiley & Sons, Inc., Second Edition, 2000.