SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

March 2018, Vol. 61 032103:1–032103:14
doi: 10.1007/s11432-016-0426-1

Countering JPEG anti-forensics based on noise level estimation

Hui ZENG[1], Jingjing YU[1], Xiangui KANG[1*] & Siwei LYU[2]

[1]*Guangdong Key Laboratory of Information Security, School of Data and Computer Science,*
Sun Yat-sen University, Guangzhou 510006*, China;*
[2]*Department of Computer Science, University at Albany, SUNY, NY* 12222*, USA*

Received June 24, 2016/Revised September 9, 2016/Accepted October 8, 2016/Published online August 14, 2017

Abstract Quantization artifact and blocking artifact are the two types of well-known fingerprints of JPEG compression. Most JPEG forensic techniques are focused on these fingerprints. However, recent research shows that these fingerprints can be intentionally concealed via anti-forensics, which in turn makes current JPEG forensic methods vulnerable. A typical JPEG anti-forensic method is adding anti-forensic dither to DCT transform coefficients and erasing blocking artifact to remove the trace of compression history. To deal with this challenge in JPEG forensics, in this paper, we propose a novel countering method based on the noise level estimation to identify the uncompressed images from those forged ones. The experimental results show that the proposed method achieves superior performance on several image databases with only one-dimensional feature. It is also worth emphasizing that the proposed threshold-based method has explicit physical meaning and is simple to be implemented in practice. Moreover, we analyze the strategies available to the investigator and the forger in the case of that they are aware of the existence of each other. Game theory is used to evaluate the ultimate performance when both sides adopt their Nash equilibrium strategies.

Keywords game theory, quantization artifact, blocking artifact, JPEG forensics, anti-forensics, noise level estimation

Citation Zeng H, Yu J J, Kang X G, et al. Countering JPEG anti-forensics based on noise level estimation. Sci China Inf Sci, 2018, 61(3): 032103, doi: 10.1007/s11432-016-0426-1

1 Introduction

The integrity of digital images has been severely challenged with the development of sophisticated image editing tools (e.g., Adobe Photoshop), which can modify contents of digital images with minimal visible traces. Accordingly, the research field of digital image forensics, which aims to authenticate the integrity of digital images, has experienced rapid developments in the past decade to meet this challenge. This, in return, leads to the development of anti-forensic techniques, the aim of which is to counteract forensic detections [1–4]. Kwok et al. [1] increased the internal bit depth of a contrast enhanced image to remove the peak-gap artifacts in the image histogram, which makes the forensic methods based on such artifacts become invalid. Milani et al. [2] modified the first digit statistics of a double compressed image to fool the Benford' law-based double compression detectors. Qian et al. [3] proposed a denoising algorithm to minimize the distortion caused by JPEG anti-forensics. Fan et al. [4] proposed a JPEG anti-forensic method based on estimating the quantization noise in the DCT domain. The interplay between forensic and anti-forensic techniques has created an interesting dynamics in research efforts [5, 6]. In [5], the

[*] Corresponding author (email: isskxg@mail.sysu.edu.cn)

interplay between the forensics and anti-forensics in source identification was modeled as a zero sum game. Stamm et al. [6] analyzed the interplay between a forensic investigator and an anti-forensic forger in the video frame deletion scenario with game theory based on a sequential move assumption. The forensic investigator was assumed to choose his strategy first and allow the forger to respond. Under the sequential move assumption, the max-min strategy was adopted to find the solution of the game.

JPEG is arguably the most widely used digital image compression standard. Determining if an image has been previously JPEG compressed is an important cue in the forensic analysis of the integrity of an image. To date, the most reliable characteristics that can be used to this end are the quantization effect and blocking artifacts in JPEG compression7. Features reflecting JPEG quantization effect or blocking artifacts are also used to detect double JPEG compression [7–9], image tampering [10–12], or perform other forensic analyses [13].

Although these forensic techniques above are effective in their own sight, they may be defeated by anti-forensic techniques that are designed to avoid their inspection. For instance, Stamm et al. [14] proposed a JPEG anti-forensic method, where anti-forensic dither is added to the DCT coefficients of a JPEG compressed image to imitate the histogram of the DCT coefficients of the original image. The same authors extended their work to conceal blocking artifacts by deblocking [15]. These anti-forensic techniques can conceal the traces of JPEG compression without introducing noticeable perceptible artifact.

In our previous work [16], a noise level based method was proposed to uncover such JPEG anti-forensics. However, this method does not work for images that contain abundant textures. In this paper, by further analyzing the characteristic of JPEG anti-forensics, we propose a new feature to expose traces of JPEG anti-forensics, which is the L_2 norm of the noise level of an image and the noise level of the compression residual (CR) of the image. Specially, CR is defined as the difference between a test image $\boldsymbol{I}$ and its compressed version $\boldsymbol{I}'$, i.e. $\mathrm{CR}(\boldsymbol{I}) = \boldsymbol{I} - \boldsymbol{I}'$. As will be illustrated in Section 3, introducing the noise level of the CR can make up the limitation of the previous method on the images that contain abundant textures. The experimental results show that the proposed method can provide improved performance when tested on the different image databases with only this single feature. Furthermore, we analyze the strategies available to the forger and the forensic investigator, and model the interplay between them as a zero-sum game. The ultimate performances for both sides under mixed strategy Nash equilibrium are evaluated with the game theory.

The rest of the paper is organized as follows. Section 2 gives a brief review of the JPEG anti-forensics and some prior work related to this paper. Section 3 describes the proposed counter anti-forensic method. The interplay between the forensic investigator and the forger in the JPEG forensics scenario is modeled as a zero-sum game in the Section 4. Section 5 shows the experimental results on the uncompressed color image database (UCID) [17] and BOSSbase image database [18]. The conclusion is given in Section 6.

2 Background and prior work

2.1 Anti-forensics of JPEG compression

JPEG compression starts by segmenting an input image into non-overlapping 8×8 pixel blocks, then it uses 2-D DCT to transform each block data into 64 DCT coefficients. Let X, Y be the DCT coefficient of an uncompressed image and its JPEG compression, respectively. For the DCT coefficients at the (i, j)-th position, $i, j \in 1, \ldots, 8$, X will be quantized as $Y = q_{i,j}\mathrm{round}(X/q_{i,j})$, where $q_{i,j}$ is the corresponding quantization step. The blocking operation introduces discontinuity in spatial domain (Figure 1(b)), and the quantization introduces a comb-like histogram of DCT coefficients as shown in Figure 2(b).

To hide such compression evidence, Stamm et al. [14] first uses dithering to remove the quantization artifact:

$$Z = Y + D, \tag{1}$$

where Y is a DCT coefficient, D is additive dither, and Z is the anti-forensically modified DCT coefficient.

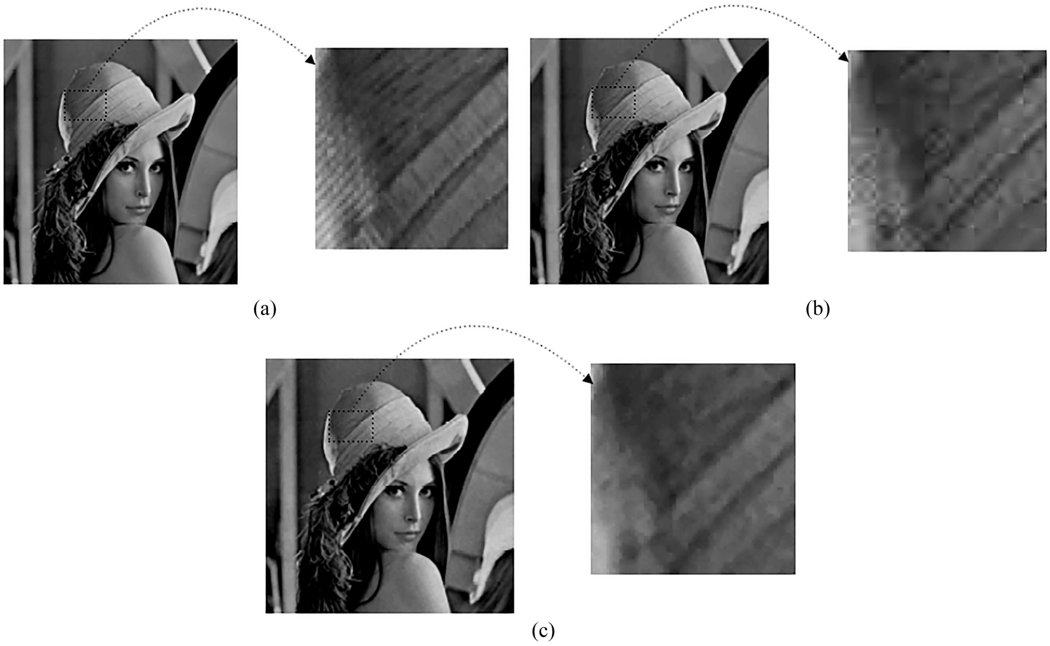

Figure 1 (a) Blocking artifact of the uncompressed Lena image; (b) the same image after JPEG compression with quality factor $Q = 75$; (c) the forged image.

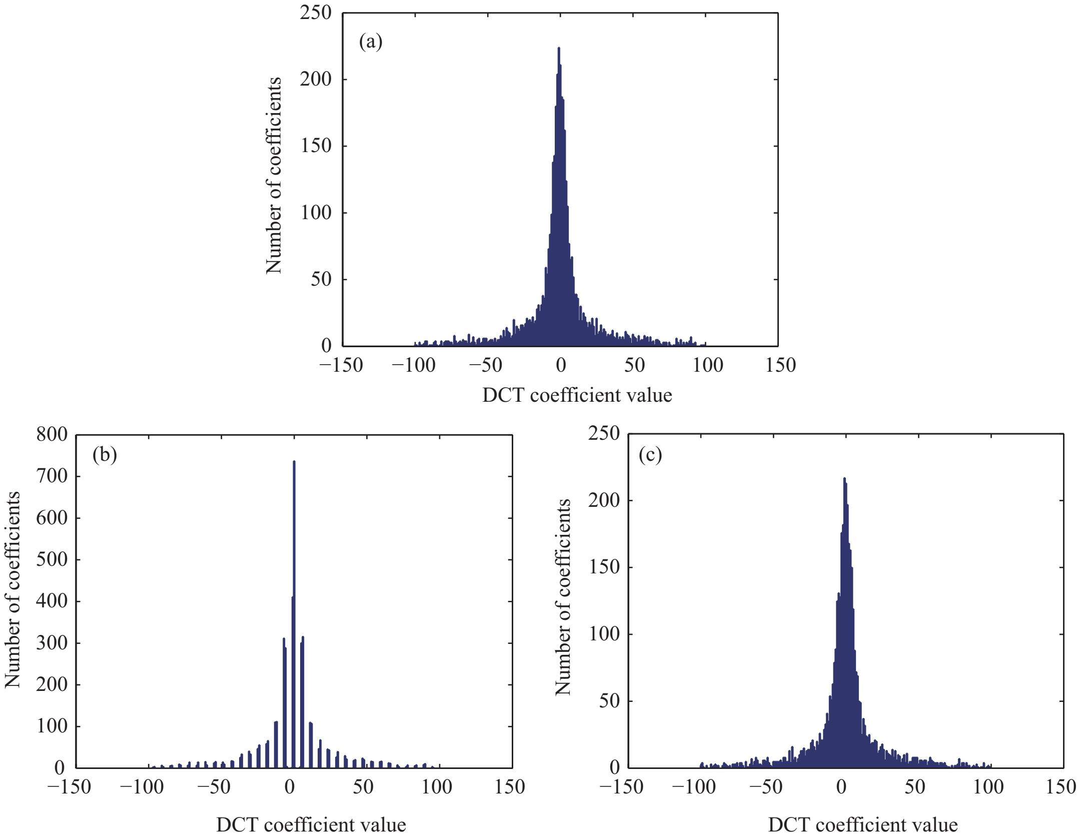

Figure 2 (Color online) Histogram of DCT coefficient of the (2, 2) subband from (a) uncompressed Lena image, (b) Lena after JPEG 75 compression, and (c) the forged image obtained after adding dither in DCT domain.

For the AC coefficients, their histograms are adjusted to match Laplacian distribution [19]. For the DC coefficient, uniformly distributed anti-forensic dither is added. Figure 2(c) illustrates that the anti-forensic

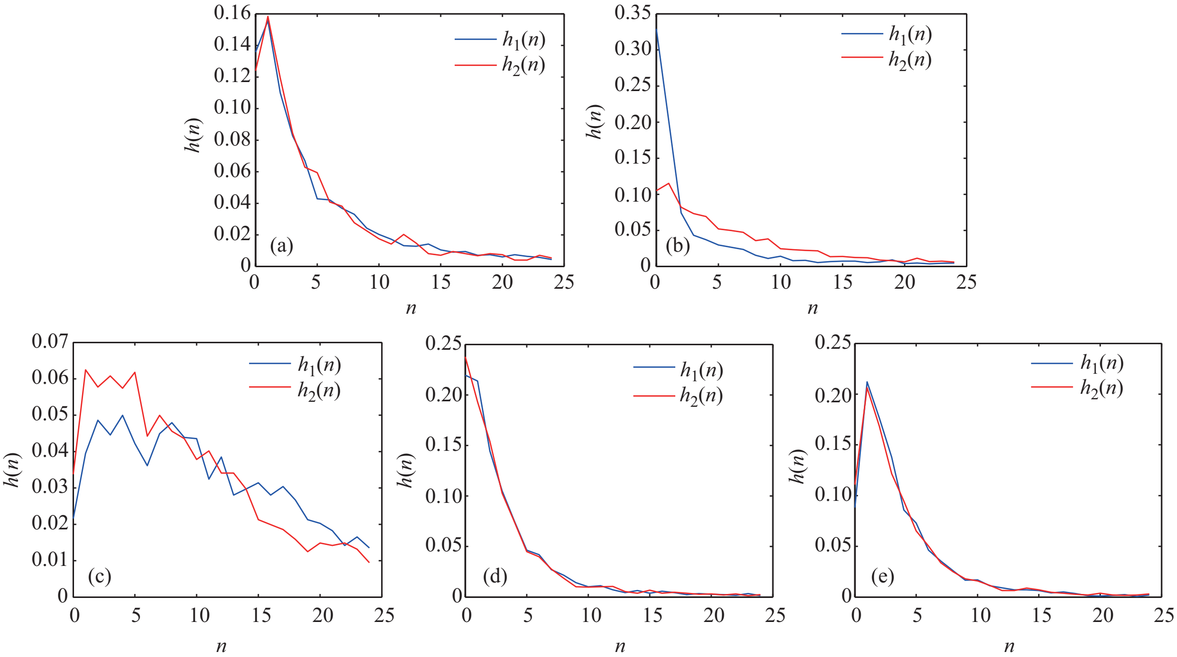

Figure 3 (Color online) Histogram ($h_1(n)$ and $h_2(n)$) of neighbor pixel differences from (a) an uncompressed image, (b) the same image after JPEG 75 compression, (c) after adding anti-forensic dither, (d) after median filtering, and (e) after adding noise to obtain the final forged image.

method can make the distribution of DCT coefficients be similar to that of the uncompressed one which is shown in Figure 2(a).

Although adding anti-forensic dither can remove the evidence of quantization, it cannot remove blocking artifact. The JPEG blocking detection method in [20] operates as follows. It collects two kinds of pixel difference measurements of an image. One kind is the average diagonal difference of four neighbor pixels at the center of each block, which is denoted as R_1. The second kind is the average diagonal difference of the four neighbor pixels spanning across block boundaries, which is denoted as R_2. The histograms of the R_1 and R_2 values of an image are denoted as $h_1(n)$ and $h_2(n)$, respectively. Figures 3 (a) and (b) show $h_1(n)$ and $h_2(n)$ of an uncompressed Lena image and its JPEG compressed version, respectively. It is observed that the difference between $h_1(n)$ and $h_2(n)$ can be used to measure the blocking artifact, i.e., the larger the difference, the greater the probability that the image has been JPEG compressed. Figure 3(c) shows $h_1(n)$ and $h_2(n)$ of the image after anti-forensic dither is added. It is observed that there also exists distinct difference between $h_1(n)$ and $h_2(n)$. Thus it is necessary for the forger to further remove blocking artifact [15]:

$$\boldsymbol{v} = \mathrm{med}_s(\boldsymbol{u}) + \boldsymbol{n}, \tag{2}$$

where $\boldsymbol{u}$ represents an image after anti-forensic dither is added, $\mathrm{med}_s(\ )$ denotes median filtering operation with a square window of size $s \times s$ pixels, and $\boldsymbol{n}$ is a zero mean Gaussian noise. The standard deviation of $\boldsymbol{n}$ is denoted as σ_1 and called as the strength of the added noise in the sequel.

Figure 3(d) shows $h_1(n)$ and $h_2(n)$ of the same image after median filtering is applied. It is observed that the difference between $h_1(n)$ and $h_2(n)$ has been significantly reduced. However, another obvious artifact appears. Note that $h_1(1) < h_1(0)$ in the Figure 3(d), whereas $h_1(1)$ is typically greater than $h_1(0)$ for an uncompressed image as shown in Figure 3(a). This difference is due to the streaking artifact of median filtering [21]. To avoid the streaking artifact, it is necessary for the forger to add Gaussian noise after applying median filtering. Figure 3(e) shows the histograms $h_1(n)$ and $h_2(n)$ of the final forged image, from which it is observed that $h_1(n)$ and $h_2(n)$ are similar to that shown in Figure 3(a). Such deblocking effectiveness can also be observed from Figure 1(c), in which the blocking artifact is significantly reduced compared to the compressed image as shown in Figure 1(b). Before close the subsection, we give the diagram of the whole anti-forensic algorithm in Figure 4, and the motivation of

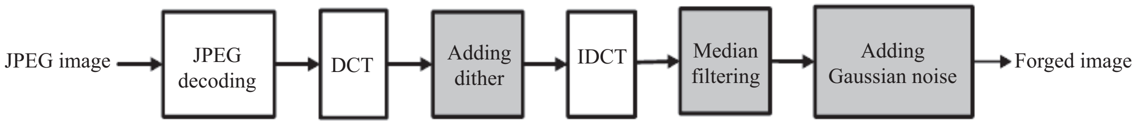

Figure 4 JPEG anti-forensics procedure.

the operations in gray shadow has been explicitly analyzed in this paper.

2.2 Existing countermeasures against JPEG anti-forensics

There exist several studies aiming to detect traces of the above anti-forensic technique. Lai and Bohme [22] proposed a detector based on the calibration feature. A test image is first cropped by 4 rows and 4 columns to obtain its cropped version. Then, the variances of the 28 high frequency DCT subbands coefficients are calculated for both versions of image. Finally, the differences of the 28 variances are averaged to get the calibration feature F_c. The investigator can distinguish an uncompressed image from a forged one by comparing F_c to a threshold t_F. Valenzise et al. [23, 24] observed that the added dither in DCT domain can be removed completely if a forged image is recompressed with the same quantization matrix used in the previous JPEG compression. Let q be the quantization matrix used in the previous JPEG compression and q_A be the quantization matrix used in the recompression. An obvious slope change can be observed in $q_A = q$ when examining the amount of recompression noise varying with q_A. The authors adopt the total variation (TV) metrics [25] to measure the amount of recompression noise present in an image. These countermeasures [22–24] can cope with the anti-forensics method without deblocking [14]. However, they fail in countering against the whole anti-forensic method [15]. Li et al. [26] observed that the anti-forensic operation may destroy the statistical correlations of an image, and regard the counter anti-forensics as a JPEG steganalysis problem. The transition probability matrix of the DCT coefficients is fed into a SVM classifier to identify the forged images from those original ones. This technique can detect the forgery of [15] well.

3 Countering technique based on noise level estimation

In this work, we formulate the detection of JPEG anti-forensics of [15] as the following hypothesis test problem:

$$\begin{aligned} &H_0 : \boldsymbol{y} = \boldsymbol{x}, \\ &H_1 : \boldsymbol{y} = \mathrm{med}_s(\boldsymbol{u}) + \boldsymbol{n} = \boldsymbol{z} + \boldsymbol{n}, \end{aligned} \tag{3}$$

where $\boldsymbol{x}$ corresponds to an unmodified image, $\boldsymbol{y}$ represents a test image. The additive Gaussian noise in (3) raises the noise level of a forged image inevitably and will be used as the feature to detect JPEG anti-forensic operation in this work.

3.1 Noise level estimation of an image

There are some algorithms in image processing area to deal with blind noise level estimation [27–31]. However, these algorithms are designed to estimate the noise level with standard deviation varying in a large range, e.g. usually $\sigma \in [0, 25]$. On the other hand, the noise introduced by JPEG anti-forensic operations is usually much lower. To our best knowledge, the algorithm in [27] performs best when the underlying true noise level to be estimated is very weak, e.g., $\sigma < 2$. Hence we adopt this algorithm to estimate the noise level of a test image in our work.

We first decompose a test image $\boldsymbol{y}$ into overlapping patches $\boldsymbol{y}_i, i = 1, \ldots, N$. Note here $\boldsymbol{y}_i$ corresponds to the vectorization of each patch, i.e., if the patches are with the size of $B \times B$ pixels, $\boldsymbol{y}_i$ is a $B^2 \times 1$ vector. The covariance matrix of image $\boldsymbol{y}$ is obtained as the average of the outer product of each vectorized patch with itself, as

$$\sum\nolimits_{\boldsymbol{y}} = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{y}_i-\mu)(\boldsymbol{y}_i-\mu)^{\mathrm{T}}, \tag{4}$$

where N is the total number of patches and μ is the average of all the patches. Using the Courant-Fisher theorem which states that the variance of the data projected onto the minimum variance direction equals the minimum eigenvalue of its covariance matrix, we can get the following equation:

$$\begin{aligned} &H_0:\lambda_{\min}(\textstyle\sum_{\boldsymbol{y}})=\lambda_{\min}(\textstyle\sum_{\boldsymbol{x}});\\ &H_1:\lambda_{\min}(\textstyle\sum_{\boldsymbol{y}})=\lambda_{\min}(\textstyle\sum_{\boldsymbol{z}})+\sigma_1^2, \end{aligned} \tag{5}$$

where $\lambda_{\min}(\sum)$ represents the minimum eigenvalue of covariance matrix $\sum$.

The weak textured areas of an image are known to span only low dimensional subspace. Hence, $\lambda_{\min}(\sum_{\boldsymbol{x}})$ and $\lambda_{\min}(\sum_{\boldsymbol{z}})$ are supposed to be approximately zero in these areas. Then, the above hypothesis can be simplified as

$$\begin{aligned} &H_0:\lambda_{\min}(\textstyle\sum_{\boldsymbol{y}'})\approx 0;\\ &H_1:\lambda_{\min}(\textstyle\sum_{\boldsymbol{y}'})\approx \sigma_1^2, \end{aligned} \tag{6}$$

where $\boldsymbol{y}'$ is the set of weak textured patches in image $\boldsymbol{y}$. From (6), it is observed that we can distinguish H_0 from H_1 by examining the estimated noise level $\sigma=\sqrt{\lambda_{\min}(\sum_{\boldsymbol{y}'})}$. The noise level estimation can be summarized as:

(1) The test image is decomposed into overlapping patches with the size of $B\times B$ pixels. $B=7$ in our work.

(2) An initial noise level $\sigma^{(0)}$ is estimated using all patches in the test image. That is, $\sigma^{(0)}=\sqrt{\lambda_{\min}(\sum_{\boldsymbol{y}})}$. The patches being over bright (average gray value $>255-\Delta_1$) or over dark (average gray value $<\Delta_2$) are ignored in order to avoid the overflow effect [28]. Δ_1 and Δ_2 are chosen to be 1 in our work.

(3) The patches whose texture strength less than a threshold $\mathit{\Gamma}^{(k)}$ are selected as the weak textured patches. For a given patch $\boldsymbol{y}_i$, its texture strength ξ_i is measured by the sum of squares of the entries in the difference matrices of $\boldsymbol{y}_i$:

$$\xi_i=[\mathbf{D}_h\boldsymbol{y}_i\ \mathbf{D}_v\boldsymbol{y}_i][\mathbf{D}_h\boldsymbol{y}_i\ \mathbf{D}_v\boldsymbol{y}_i]^{\mathrm{T}}, \tag{7}$$

where the $1\times B^2$ vectors $\mathbf{D}_h\boldsymbol{y}_i$ and $\mathbf{D}_v\boldsymbol{y}_i$ represent the horizontal and vertical difference of $\boldsymbol{y}_i$, respectively. $\mathit{\Gamma}^{(k)}$ is directly proportional to $\sigma^{(k)}$, i.e., $\mathit{\Gamma}^{(k)}=c\times\sigma^{(k)}$. In our work, we adopt $c=81.8$ as that in [27].

(4) A new noise level $\sigma^{(k+1)}$ is estimated using the selected patches $\boldsymbol{y}'$, that is, $\sigma^{(k+1)}=\sqrt{\lambda_{\min}(\sum_{\boldsymbol{y}'})}$. The process of steps 3 and 4 is iterated until σ is stable. According to our experiments, most of this iteration converges after three iterations and we let the algorithm stop after three iterations.

Figure 5(a) is an image from UCID. Figure 5(b) shows the selected weak textured areas (white parts). This noise level algorithm is effective for most images in our experiments. However, it will be failed when the test image contains little weak textured area. This is because the assumption of (6) becomes invalid in such situation. Figure 6 shows that some images containing little weak textured area. It is observed that almost no homogeneous areas can be found in those images. For these images, we simply set the estimated noise level $\sigma=0$ in [16]. Obviously, this setting cannot help in differentiating original images and forged images. To reduce the wrong detections due to the algorithm failure, we resort to the noise level estimation of the compression residual CR as illustrated in the next subsection.

3.2 Noise level estimation of the compression residual

It is well known that JPEG compression has the nature of removing high frequency component and reserving low frequency component in the image. Hence, for the weak texture areas of an original image, JPEG compression would not introduce much difference, i.e., the value of compression residual CR tends to be zero in these areas. However, for a forged image, the added Gaussian noise would contaminate these

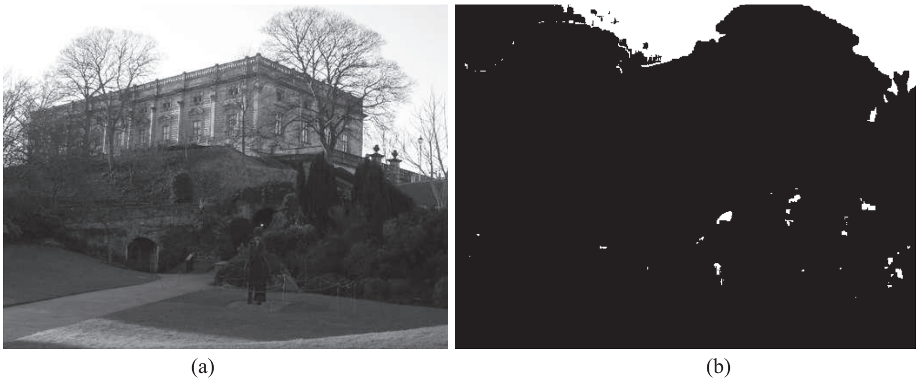

Figure 5 Noise level estimation of a test image from UCID. (a) An uncompressed image; (b) selected weak textured areas, where some overexposure areas are excluded.

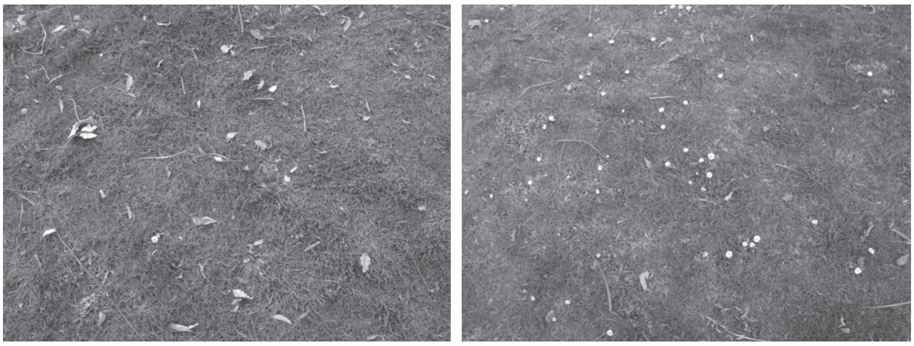

Figure 6 Images containing only textures. There are no enough homogeneous areas which can be selected from those images for noise level estimation.

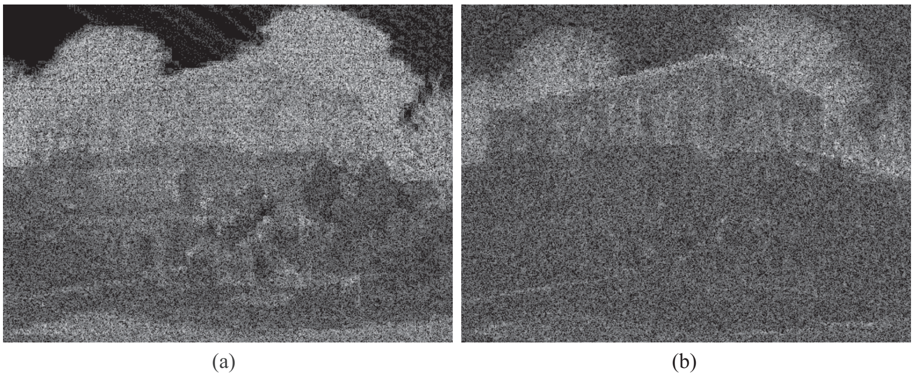

Figure 7 Compression residual of (a) an original image and (b) a forged image.

areas. This difference can be observed from Figure 7. Figure 7(a) shows the CR of an original image. It is observed that some regions that are close to zero on the top which correspond to weak textured regions of the original image. Figure 7(b) shows the CR of the forged image. Due to the anti-forensic dither and Gaussian noise, even the weak textured regions become noisy now. Hence, the noise level of CR is also effective in revealing the traces of JPEG anti-forensics. Moreover, since the texture strength of CR is usually much weaker than that of the image itself, more patches will be regarded as weak textured patches according to step 3 of the noise level estimation algorithm. Hence, the chance that the algorithm fails due to no patches can be founded is significantly reduced, which makes up the limitation of noise level estimation of the image mentioned in the last subsection.

To choose a proper quality factor Q_2 in producing compression residual, we compress an image with different Q_2 and obtain different CRs. Then, the noise level of CR is estimated with the same method

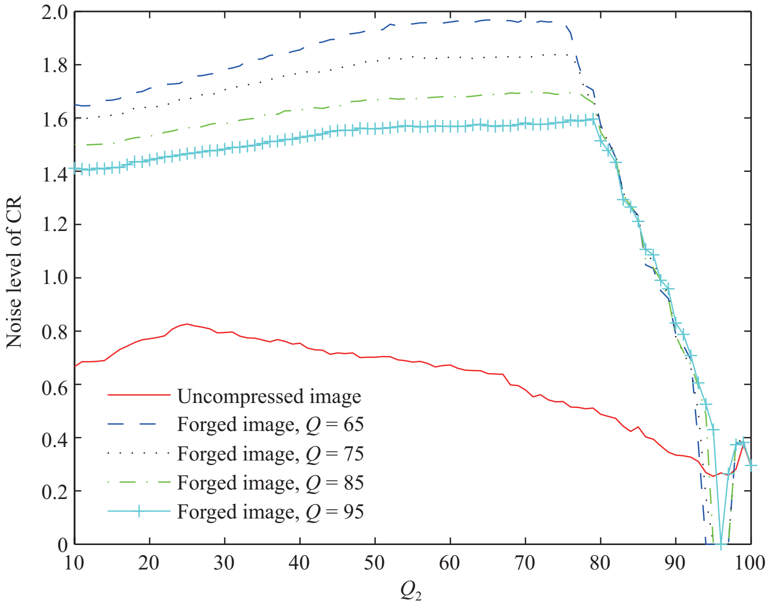

Figure 8 (Color online) Noise level of the CR of an original image and a forged image under different Q_2.

mentioned in last subsection except that neither over bright nor over dark patches need to be ignored. Figure 8 shows noise level of CR for an original image and four forged images, respectively, under different Q_2. The four forged images are created by first JPEG compress the original image with four different quality factors, $Q = [65, 75, 85, 95]$, and forged with the anti-forensic method in [15]. It is observed that the noise level of CR remains in a relative low level for an original image and obviously higher for the forged images when $Q_2 < 80$. When a high quality factor Q_2 is used, e.g., $Q_2 > 80$, the difference between the image and its JPEG compressed version will be trivial, and it is hard to distinguish a forged image from an original image by the noise level of CR. It is also observed the results are similar for different Q. Based on the analysis above, we average the estimated noise level of CRs under $Q_2 = \{55, 65, 75\}$ in order to obtain a robust estimation.

3.3 Counter anti-forensics based on noise level estimation

Now for an image $\boldsymbol{I}$ we have obtained two noise level estimations. One is the noise level estimation of itself, which is denoted as

$$F_1 = \mathrm{NL}(\boldsymbol{I}) = \sqrt{\lambda_{\min}(\textstyle\sum_{\boldsymbol{y}'})}, \tag{8}$$

where $\mathrm{NL}(\boldsymbol{I})$ denotes the estimated noise level of $\boldsymbol{I}$, $\boldsymbol{y}'$ is the set of weak textured patches in image $\boldsymbol{I}$. Another is the noise level estimation of its CR, which is denoted as

$$F_2 = \frac{1}{3}[\mathrm{NL}(\mathrm{CR}_{55}(\boldsymbol{I})) + \mathrm{NL}(\mathrm{CR}_{65}(\boldsymbol{I})) + \mathrm{NL}(\mathrm{CR}_{75}(\boldsymbol{I}))], \tag{9}$$

where $\mathrm{CR}_{55}(\boldsymbol{I})$, $\mathrm{CR}_{65}(\boldsymbol{I})$, and $\mathrm{CR}_{75}(\boldsymbol{I})$ denote the CRs under compression quality factor $Q_2 = \{55, 65, 75\}$, respectively.

The two-dimension scatter plots of F_1 and F_2 of images in UCID and BOSSbase are shown in Figure 9. The blue point corresponds to an original image and the red star corresponds to a forged image. The points that located in the axes ($F_1 = 0$ or $F_2 = 0$) correspond to the cases of that the algorithm fails to obtain a real value due to no selected weak-textured patches. By observing the scatter plots in Figure 9, we fuse F_1 and F_2 to obtain the final feature $F = \sqrt{F_1^2 + F_2^2}$ for classification. The decision rule δ^n is defined as follow:

$$\delta^n = \begin{cases} H_0, & F < t_n; \\ H_1, & \text{otherwise}, \end{cases} \tag{10}$$

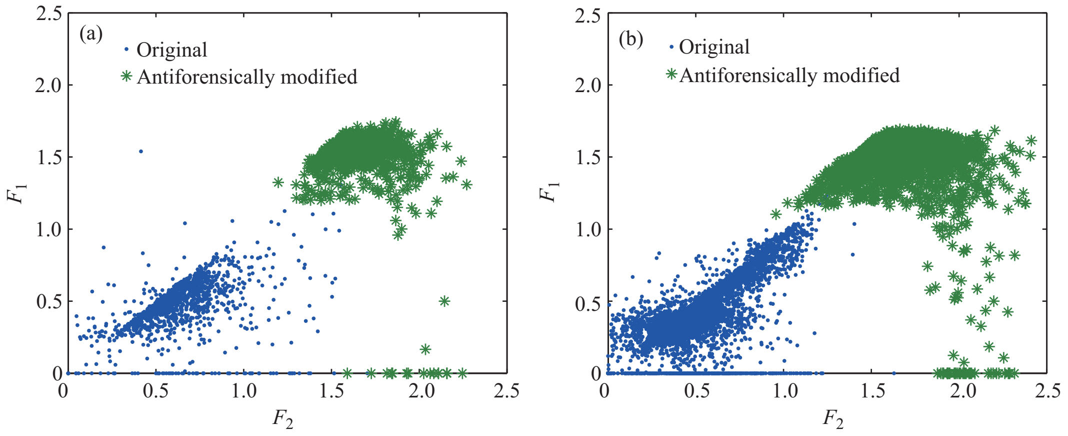

Figure 9 (Color online) Two-dimension scatter plots of features F_1 and F_2, where F_1 is the noise level of the test image itself and F_2 is the noise level of CR. The blue points denote the original images and the green stars denote the forged ones. (a) UCID; (b) BOSSbase.

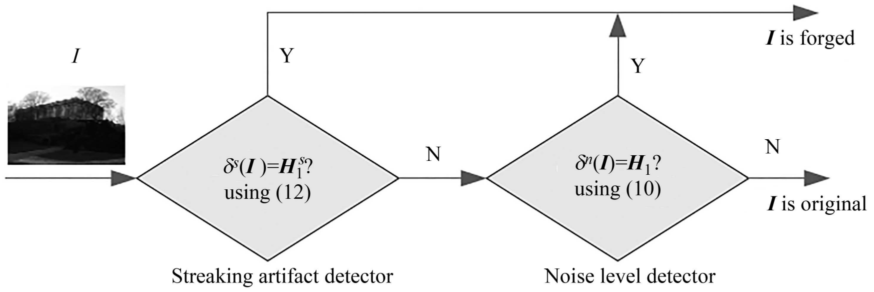
I
Y
$\delta^s(\boldsymbol{I})=\boldsymbol{H}_1^s$?
using (12)
N
Y
$\delta^n(\boldsymbol{I})=\boldsymbol{H}_1$?
using (10)
$\boldsymbol{I}$ is forged
N
$\boldsymbol{I}$ is original
Streaking artifact detector
Noise level detector

Figure 10 The diagram of the forgery detector. We use the metric ρ [21] to measure the streaking artifact and the proposed method to measure the noise level.

where the decision threshold t_n is chosen according to a given false alarm rate P_{fa}^n.

4 Game theory evaluation

To avoid being detected by our method, an image forger can use weaker noise perturbation in (2). Also note in practice, an image forger may not need to mislead the forensic tool completely, but rather just need to reduce the accuracy of it. On the other hand, using a very weak noise cannot avoid the streaking artifact mentioned in Subsection 2.1. Therefore the image forger is faced with a tradeoff in choosing the noise strength balancing the risk of being detected by the noise level detector and the streaking artifact detector.

From the perspective of a forensic investigator, whose goal is to reveal such image manipulations, the requirement is to combine the results of streaking artifact detector and noise level detector together, i.e., making OR operation with the binary outputs from the two detectors. The diagram of the whole forgery detector is shown in Figure 10. A query image $\boldsymbol{I}$ will be labeled as forged if H_1 is accepted either the streaking artifact detector or the noise level detector.

Here we use the metric ρ to measure the streaking artifact [21]:

$$\rho = h_D(0)/h_D(1), \tag{11}$$

where h_D is the histogram of pixel difference of a test image. ρ is expected to be approximately 1 for original images and will be much greater for median filtered ones. Hence, the streaking artifact detector δ^s can be defined as

$$\delta^s = \begin{cases} H_0^s, & \rho < t_s; \\ H_1^s, & \text{otherwise,} \end{cases} \tag{12}$$

where the decision threshold t_s is chosen according to a given false alarm rate P_{fa}^s. The total detection rate is the probability of a forged image either being detected by streaking artifact detector or being detected by noise level detector.

$$P_d = P(\delta^s(\boldsymbol{I}) = H_1^s \cup \delta^f(\boldsymbol{I}) = H_1^f | \boldsymbol{I} \text{ is forged}). \tag{13}$$

Note both the two detectors would introduce false alarm to the final decision, and the total allowed false alarm rate is often limited in practical scenario. Hence, the forensic investigator needs to make a tradeoff in allocating the allowed false alarm rate between the two detectors. The total allowed false alarm rate is defined as

$$P_{\text{fa}} = P_{\text{fa}}^s + P_{\text{fa}}^n. \tag{14}$$

These tradeoffs made by the forger and the forensic investigator lead to the following questions. With the existence of this counter anti-forensics, what is the optimal σ_1 in (2) that the forger should adopt? For the investigator, how to allocate the allowed false alarm rate P_{fa} between the two detectors? What is the ultimate performance if both sides adopt their optimal strategies? To answer these questions, the interplay between the forger and the investigator is modeled as a zero-sum game [32, 33].

Definition 1 (The JPEG forensic game)**.** The JPEG(S_I, S_F, u) game is a zero sum game played by an investigator and a forger, defined by the following strategies and payoff.

S_I: the investigator's strategy is the false alarm rate P_{fa}^s that can be allocated to the streaking artifact detector δ^s.

S_F: the forger's strategy is the strength σ_1 of the added Gaussian noise.

u: the payoff matrix is defined in terms of the detection rate of revealing the anti-forensic attack,

$$u(P_{\text{fa}}^s, \sigma_1) = P_d(P_{\text{fa}}^s, \sigma_1). \tag{15}$$

The optimal strategies of both players can be obtained by solving the JPEG forensic game, e.g., finding its mixed strategy Nash equilibrium. The investigator's mixed strategy $\boldsymbol{P}_{\text{fa}}^s = [x_1, x_2, \ldots, x_m]$ is a probability distribution over different P_{fa}^ss, and the forger's mixed strategy $\boldsymbol{\sigma}_1 = [y_1, y_2, \ldots, y_n]$ is a probability distribution over different σ_1s.

To solve the JPEG forensic game, we formulate it as a linear programming problem [33], i.e., to find the minimum v which is subject to the following constraints:

$$\begin{array}{ll} y_i \geqslant 0, & i = 1, 2, \ldots, n; \\ \sum_i y_i = 1; & \\ \sum_i u_{ij} y_i - v \leqslant 0, & j = 1, 2, \ldots, m, \end{array} \tag{16}$$

where $u_{ij} = P_d(P_{\text{fa}i}^s, \sigma_{1j})$ is the total detection rate when the investigator adopts $P_{\text{fa}i}^s$ and the forger adopts σ_{1j}. v is the objective function of the linear programming problem. The solution v^* to the JPEG forensic game and the optimal strategy $\boldsymbol{\sigma}_1^*$ for the forger can be obtained by solving the optimization over $n+1$ parameters $(v, y_1, y_2, \ldots, y_n)$. The optimal strategy $\boldsymbol{P}_{\text{fa}}^{s*}$ for the investigator can be obtained by solving a dual problem of (16). These optimizations can be solved with the linear programming method [34].

For a given allowed false alarm rate P_{fa}, the Nash equilibrium of the JPEG forensic game can be derived and the corresponding outcome $u(\boldsymbol{P}_{\text{fa}}^{s*}, \boldsymbol{\sigma}_1^*)$ is the ultimate detection rate. A receiver operating characteristic (ROC) curve can be constructed to show the detection rates under the Nash equilibrium varying with different P_{fa}.

5 Experimental results

5.1 The performance of counter anti-forensics

In order to test the counter JPEG anti-forensic method described in Section 3, we performed experiments on the images from UCID and BOSSbase. There are 1338 images of size 384 × 512 or 512 × 384 pixels in

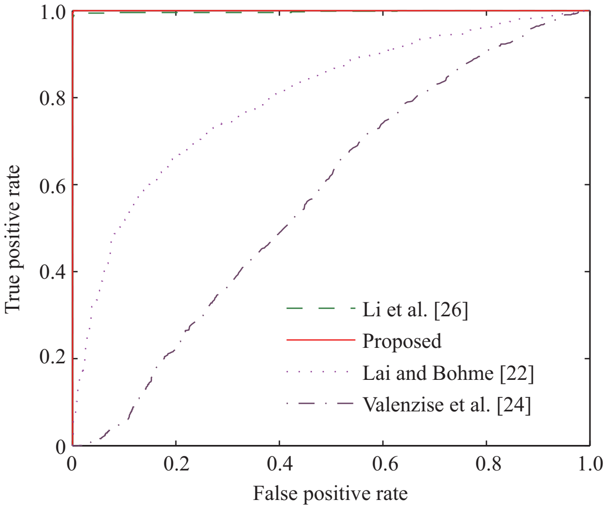

Figure 11 (Color online) ROC comparison in countering JPEG anti-forensics on UCID, $Q = 75$.

Table 1 Detection accuracy compared with other methods on UCID (%)$^{a)}$

	Feature dimension	$Q = 95$	$Q = 85$	$Q = 75$	$Q = 65$
F_1	1	99.1	99.1	99.1	99.0
F_2	1	98.6	99.0	99.1	98.9
F	1	**99.6**	**99.7**	**99.8**	**99.8**
Ref. [26]	100	99.4	99.6	99.7	99.7

a) The number in bold type denotes the best performance.

Table 2 Detection accuracy compared with other methods on BOSSbase (%)$^{a)}$

	Feature dimension	$Q = 95$	$Q = 85$	$Q = 75$	$Q = 65$
F_1	1	99.6	99.6	99.6	99.6
F_2	1	99.7	99.7	99.8	99.7
F	1	**99.9**	**99.9**	**99.9**	**99.9**
Ref. [26]	100	99.8	**99.9**	**99.9**	99.8

a) The same as in Table 1.

the UCID, and 10000 images of size 512 × 512 pixels in the BOSSbase. Both image datasets have been widely used to evaluate forensic and anti-forensic methods in the research community. All the images are converted to grayscale images. The uncompressed images are first JPEG compressed with quality factors $Q = \{95, 85, 75, 65\}$. Then these JPEG images are anti-forensically modified with the method [15] using the suggested parameters, i.e., $s = 3$ and $\sigma_1^2 = 2$. Experimental results show that the detection performance of our method is robust with regards to Q or s.

From the forensic investigator's point of view, the main purpose is to differentiate the original images from the forged ones. Figure 11 shows the ROC curves on UCID when different methods are adopted to counter the JPEG anti-forensics. For the SVM-based method [26], we use half of the images as training dataset and the remaining half of the images as test dataset. It is observed that both the method [26] and our proposed method achieve nearly perfect performance, whereas the methods in [22, 24] fail to differentiate the original images from the forged ones.

The performance comparisons for different quality factor Q are shown in Tables 1 and 2. The accuracy is defined as

$$\mathrm{Acc} = \frac{\sharp\text{correctly classified images}}{\sharp\text{total testing images}}. \tag{17}$$

It is observed that the proposed method with feature F achieves better performance than that with feature F_1 or F_2, which means that it is necessary to combine the two features for better distinguishing original images from the forged ones. Even compared to the SVM-based method [26], the proposed

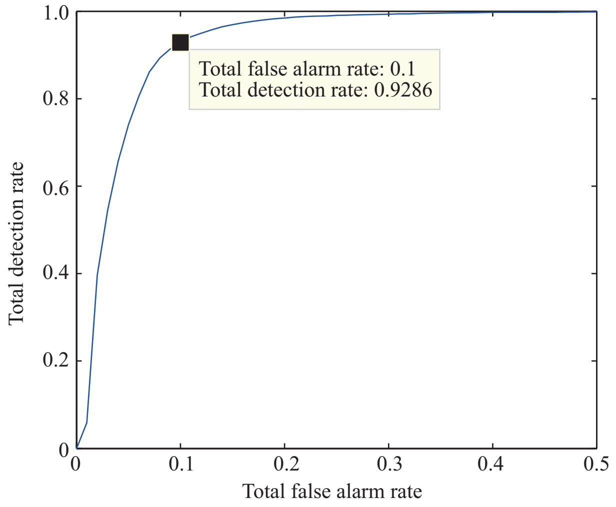

Figure 12 (Color online) Nash equilibrium ROC on UCID.

method achieves comparable or better performance. In fact, as discussed in [24], it is usually difficult to make a fair comparison between a threshold-based method and a SVM-based method. The SVM-based method uses 100-dimension features, which originally designed for steganalysis, requires the investigator owning enough forged samples to train a classifier and having a full knowledge of the anti-forensics. The proposed method, which is specifically tailored to detect the JPEG anti-forensics, merely requires the investigator examining the noise level of a test image. For this point, the proposed method is much easier to be implemented in practical scenario than the SVM-based method [26].

5.2 The performance under Nash equilibrium

The result showed in the Subsection 5.1 corresponds to the case that the forger adds a specific strength noise, i.e., $\sigma_1^2 = 2$. In this subsection, we show the ultimate performance of the interplay between the forensic investigator and the forger by finding the Nash equilibrium of the JPEG forensic game. Here we allow σ_1 varies from 0.1 to 2 with a step of 0.1. Figure 12 shows the Nash equilibrium ROC on UCID. Note the performance of other countermeasures [22–24, 26] are irrelevant to the strength σ_1 used by the forger, i.e., no meaningful Nash equilibrium can be found when these methods are adopted for countering JPEG anti-forensics in the JPEG forensic game. Hence, we only show the Nash equilibrium ROC for our method.

It can be observed the chance of the JPEG anti-forensics being uncovered is obviously reduced if the forger adopts the Nash equilibrium strategy. Take $P_{\rm fa} = 0.1$ for example, the true positive rate equals 1 in Figure 11, which means that all forged images are correctly detected by the proposed method. If the forger adopt the Nash equilibrium strategy, the detection rate $P_d = 0.9286$ as shown in Figure 12. This Nash equilibrium corresponds to the case that the investigator chooses $\boldsymbol{P}_{\rm fa}^s = [0.03, 0.04]$ with probability combination of [0.6, 0.4], whereas the forger chooses $\boldsymbol{\sigma}_1 = [0.4, 0.5]$ with probability combination of [0.24, 0.76].

By examining the Nash equilibrium strategy, we find that the forger tends to choose a much weaker noise compared to previous work [14, 15, 22–24, 26]. In all of our experiments, the strength of added noise $\sigma_1 < 0.8$ according to the Nash equilibrium. This choice affords an intuitive explanation, a noise strength $\sigma_1 = 0.8$ is sufficient to hide streaking artifact, and using a stronger strength merely increases the chance of that the forgery is detected by the proposed counter anti-forensic method. Therefore a rational forger is more likely to choose a relative weaker strength.

6 Conclusion

The development of image forensic techniques fosters the development of anti-forensics, by which image forgers can use to counter forensic analysis. In this work, we provide new weapons to the arsenal of forensic investigators that can expose traces of JPEG anti-forensic operations [14, 15]. We find that a key step in JPEG anti-forensics, namely deblocking, increases the noise level of the forged images, and we use a noise level based feature to reveal traces of JPEG anti-forensics. Compared to [26] with 100-D features, the proposed method yields improved performance with only one-dimensional feature and has more explicit physical meaning. It is also worth emphasizing that the proposed method does not require the forensic investigator owning a large number of forged samples to train a classifier, making it much easier to be implemented in practice.

To our best knowledge, this is the first time that the interplay between the forger and the investigator in JPEG forensics is modeled as a game, for the case that both sides are aware of the existence of each other. Game theory analysis shows that the forger would choose a weaker strength compared with former work [14, 15, 22–24, 26]. Applying the game theory framework to analyze the interplay between forensics and anti-forensics in more scenarios is our future work.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. U1536204, 61379155, 61332012), National Natural Science Foundation of Guangdong Province (Grant No. s2013020012788), and Special Funding for Basic Scientific Research of Sun Yat-sen University (Grant No. 6177060230).

Conflict of interest The authors declare that they have no conflict of interest.

References

1 Kwok C W, Au O C, Chui S H. Alternative anti-forensics method for contrast enhancement. In: Proceedings of the 10th International Conference on Digital-Forensics and Watermarking, Atlantic, 2011. 398–410

2 Milani S, Tagliasacchi M, Tubaro S. Antiforensics attacks to Benford's law for the detection of double compressed images. In: Proceedings of International Conference on Acoustics, Speech, and Signal Processing, Vancouver, 2013. 3053–3057

3 Qian Z X, Zhang X P. Improved anti-forensics of JPEG compression. J Syst Softw, 2014, 91: 100–108

4 Fan W, Wang K, Cayre F, et al. JPEG anti-forensics using non-parametric DCT quantization noise estimation and natural image statistics. In: Proceedings of the 1st ACM Workshop on Information Hiding and Multimedia Security, Montpellier, 2013. 117–122

5 Barni M, Tondi B. The source identification game: an information-theoretic perspective. IEEE Trans Inf Forens Secur, 2013, 8: 450–463

6 Stamm M C, Lin W S, Liu K J R. Forensics vs. anti-forensics: a decision and game theoretic framework. In: Proceedings of International Conference on Acoustics, Speech, and Signal Processing, Kyoto, 2012. 1749–1752

7 Pevny T, Fridrich J. Detection of double-compression in JPEG images for applications in steganography. IEEE Trans Inf Forens Secur, 2008, 3: 247–258

8 Lukas J, Fridrich J. Estimation of primary quantization matrix in double compressed JPEG images. In: Proceedings of Digital Forensic Research Workshop, Cleveland, 2003. 1–17

9 Fu D D, Shi Y Q, Su W. A generalized Benford's law for JPEG coefficients and its applications in image forensics. In: Proceedings of SPIE 6505, Security, Steganography, and Watermarking of Multimedia Contents, San Jose, 2007. 65051L

10 He J F, Lin Z C, Wang L F, et al. Detecting doctored JPEG images via DCT coefficient analysis. In: Proceedings of 9th European Conference on Computer Vision, Graz, 2006. 423–435

11 Bianchi T, de Rosa A, Piva A. Improved DCT coefficient analysis for forgery localization in JPEG images. In: Proceedings of International Conference on Acoustics, Speech, and Signal Processing, Prague, 2011. 2444–2447

12 Farid H. Exposing digital forgeries from JPEG ghosts. IEEE Trans Inf Forens Secur, 2009, 4: 154–160

13 Farid H. Digital Image Ballistics from JPEG Quantization. TR2006–583. 2008

14 Stamm M C, Tjoa S K, Lin W S, et al. Anti-forensics of JPEG compression. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, 2010. 1694–1697

15 Stamm M C, Liu K J R. Anti-forensics of digital image compression. IEEE Trans Inf Forens Secur, 2011, 6: 1050–1065

16 Jiang Y W, Zeng H, Kang X G, et al. The game of countering JPEG anti-forensics based on the noise level estimation. In: Proceedings of Asian-Pacific Signal and Information Processing Association Annual Submit Conference, Taiwan, 2013. 1–9

17 Schaefer G, Stich M. UCID: an uncompressed color image database. In: Proceedings of SPIE 5307, Storage and Retrieval Methods and Applications for Multimedia, San Jose, 2004. 472–480

18 Bas P, Filler T, Pevny T. Break our steganographic system: the ins and outs of organizing BOSS. In: Proceedings of International Conference on Information Hiding, Prague, 2011. 59–70

19 Lam E Y, Goodman J W. A mathematical analysis of the DCT coefficient distributions for images. IEEE Trans Image Process, 2000, 9: 1661–1666

20 Fan Z G, de Queiroz R L. Identification of bitmap compression history: JPEG detection and quantizer estimation. IEEE Trans Image Process, 2003, 12: 230–235

21 Kirchner M, Fridrich J. On detection of median filtering in digital images. In: Proceedings of SPIE, Electronic Imaging, Media Forensics and Security II, San Jose, 2010. 1–12

22 Lai S Y, Bohme R. Countering counter-forensics: the case of JPEG compression. In: Proceedings of International Conference on Information Hiding, Prague, 2011. 285–298

23 Valenzise G, Nobile V, Tagliasacchi M, et al. Countering JPEG anti-forensics. In: Proceedings of IEEE International Conference on Image Processing, Brussels, 2011. 1949–1952

24 Valenzise G, Tagliasacchi M, Tubaro S. Revealing the traces of JPEG compression anti-forensics. IEEE Trans Inf Forens Secur, 2013, 8: 335–349

25 Rudin L I, Osher S, Fatemi E. Nonlinear total variation based noise removal algorithms. Physica D: Nonl Phenom, 1992, 60: 259–268

26 Li H D, Luo W Q, Huang J W. Countering anti-JPEG compression forensics. In: Proceedings of IEEE International Conference on Image Processing, Orlando, 2012. 241–244

27 Liu X H, Tanaka M, Okutomi M. Noise level estimation using weak textured patches of a single noisy image. In: Proceedings of IEEE International Conference on Image Processing, Orlando, 2012. 665–668

28 Shin D H, Park R H, Yang S, et al. Block-based noise estimation using adaptive Gaussian filtering. IEEE Trans Consum Electr, 2005, 51: 218–226

29 Pyatykh S, Hesser J, Zheng L. Image noise level estimation by principal component analysis. IEEE Trans Image Process, 2013, 22: 687–699

30 Liu W, Lin W S. Additive white Gaussian noise level estimation in SVD domain for images. IEEE Trans Image Process, 2013, 22: 872–883

31 Lyu S W, Pan X Y, Zhang X. Exposing region splicing forgeries with blind local noise estimation. Int J Comput Vis, 2014, 110: 202–221

32 Osborne M J, Rubinstein A. A Course in Game Theory. Cambridge: MIT Press. 1994

33 Hespanha J P. An Introductory Course in Noncooperative Game Theory. http://www.ece.ucsb. edu/∼hespanha/

34 Grant M C, Boyd S P. Graph Implementations for Non-smooth Convex Programs. Recent Advances in Learning and Control. Heidelberg: Springer-Verlag, 2008. 95–110