COUNTERING JPEG ANTI-FORENSICS

G. Valenzise, V. Nobile, M. Tagliasacchi, S. Tubaro

Dipartimento di Elettronica e Informazione
Politecnico di Milano, Italy

ABSTRACT

JPEG coding leaves characteristic footprints that can be leveraged to reveal doctored images, e.g. providing the evidence for local tampering, copy-move forgery, etc. Recently, it has been shown that a knowledgeable attacker might attempt to remove such footprints by adding a suitable anti-forensic dithering signal to the image in the DCT domain. Such noise-like signal restores the distribution of the DCT coefficients of the original picture, at the cost of affecting image quality. In this paper we show that it is possible to detect this kind of attack by measuring the noisiness of images obtained by re-compressing the forged image at different quality factors. When tested on a large set of images, our method was able to correctly detect forged images in 97% of the cases. In addition, the original quality factor could be accurately estimated.

Index Terms— digital image forensics; anti-forensics; JPEG compression

1. INTRODUCTION

Tampering with digital images frequently entails the use of several pictures as "sources" to create high quality composite forgeries. Most images are coded with the JPEG standard and the doctored images are often re-saved in this format. Therefore, the footprints left by JPEG compression provide valuable clues that can be leveraged by the forensic analyst. For example, when an image is compressed using JPEG, the histogram of the discrete cosine transform (DCT) coefficients features distinctive periodic patterns resembling a comb-like shape. This property has been employed to trace back the compression history of an image by estimating the original quantization matrix [1], and to identify: double JPEG compression [2]; evidence of local tampering [3]; and copy-move forgeries [4].

Recently, it has been shown that a knowledgeable attacker could restore the original distribution of transform coefficients, thus hiding the traces of JPEG compression [5, 6]. In this manner, all forensic techniques that leverage the presence of JPEG footprints can be fooled. The underlying idea consists in adding a suitable dithering signal, whose distribution is such that it fills the gaps of the comb-shaped DCT coefficient distribution in the JPEG-compressed image. However, it can be pointed out that the anti-forensic technique proposed in [5, 6] focuses only on restoring the original distribution of DCT coefficients, neglecting the consequences of this operation in the spatial domain. Indeed, the dithering signal does not recover the image content lost during quantization, but it introduces a visible distortion in the attacked image. In our previous work [7], we formally analyzed this kind of distortion, showing that it could be very difficult to conceal even if more sophisticated, perceptually-aware insertion techniques are adopted.

In this paper we propose a method that can be employed by the forensic analyst to detect whether an image has been attacked with the anti-forensic technique in [5, 6], i.e. with the purpose of hiding the traces of JPEG compression. In addition, the proposed method is able to accurately estimate the original JPEG quality factor. Our method is somewhat inspired by the work of Farid [8]. There, doubly compressed portions of a picture are detected by re-compressing the analyzed image at different quality factors. Briefly, when the quality factor matches one of those used for a certain image region, the differences between the analyzed image and its re-compressed version tend to be locally small, thus indicating double compression. Similarly, we study the mean-square-error distortion between the re-compressed version of the attacked image and the original JPEG-compressed image. We show that the distortion is completely annihilated only if the image is re-compressed at the same quality factor. Conversely, when the quality factor used for re-compression is higher, the dithering signal contributes to a non-negligible distortion, which is clearly visible as noise.

Differently from [8], however, we do not have access to the original JPEG-compressed image when computing the mean-square-error distortion after re-compression. Nevertheless, we observe that the dithering signal introduces noise-like artifacts in the re-compressed image, which are clearly visible when the quality factor is higher than the original. Therefore, we measure the noisiness of re-compressed images by means of the total variation (TV), i.e. the ℓ_1 norm of the spatial first-order derivatives. This metric has been widely used in the field of denoising to quantify the presence of noise in natural images [9]. According to the proposed method, the forensic analyst re-compresses the (possibly) attacked image at different quality factors, and analyzes the TV of the re-compressed images. This analysis enables to automatically detect whether the image has been attacked and, in that case, to estimate the quality factor of the original JPEG-compressed picture. Experiments on a large dataset of natural images demonstrate that it is possible to correctly detect attacked images with an accuracy equal to 97%, and to identify the original quality factor in the 100-points IJG scale [10] with an average error of 0.1 units. The method works with a broad range of quality factors, making it a practical as well as robust forensic tool.

The rest of the paper is organized as follows. Section 2 reviews the fundamentals of JPEG compression and the anti-forensic technique proposed in [5]. In Section 3 we describe the proposed method based on the analysis of the total variation of re-compressed versions of the image available to the forensic analyst. Experiments on a large dataset are reported in Section 4. Finally, Section 5 concludes the paper.

2. JPEG ANTI-FORENSICS

Without loss of generality, we consider JPEG compression on the luminance channel only. In the JPEG standard, the input image is divided into B non-overlapping blocks of size 8×8. For each block,

the two-dimensional DCT transform is computed. Each transform coefficient X_i^b, $1 \leq b \leq B$, $1 \leq i \leq 64$ is then quantized using a uniform quantizer, with quantization step size q_i which depends only on the DCT subband. The set of q_i's forms the quantization matrix $\mathbf{Q}$. While $\mathbf{Q}$ is not standardized, it is customary in many JPEG implementations to define $\mathbf{Q}$ as a scaled version of a template matrix $\mathbf{Q}_t$, by changing a (scalar) quality factor Q. The quantization levels W_i^b are obtained from the original coefficients X_i^b as $W_i^b = \text{round}\left(X_i^b/q_i\right)$, and then entropy coded and written in the JPEG bitstream. When the bitstream is decoded, the DCT values are reconstructed from the quantization levels as $\tilde{X}_i^b = q_i W_i^b$. Then, the inverse DCT is applied to each block, and the result is rounded and truncated in order to take integer values in the $[0, 255]$ range.

The distribution of the unquantized AC DCT coefficient is typically Laplacian. However, after dequantization, the reconstructed values $\tilde{X}_i^b$ can only be integer multiples of the quantization step size q_i. As a result, the distribution of the transform coefficients in a certain DCT subband will be composed by a train of spikes spaced apart by q_i. This comb-like structure reveals that: a) a quantization process has occurred; and, b) which was the original quantization step size. The anti-forensic technique in [5] thwarts the detection of JPEG compression by injecting a noise-like dithering signal into the dequantized DCT coefficients. The distribution of the dithering signal is designed in order to remove the traces left by the quantization process and to reconstruct the original coefficient distribution, i.e. a Laplacian distribution for AC coefficients, and an approximately uniform distribution for DC coefficients. The energy of the dithering signal necessary to restore the original DCT coefficient distribution depends on the initial JPEG quality factor Q, as well as on the image content. Generally speaking, this energy is higher for low quality factors and highly textured images [7]. In the spatial domain, this anti-forensic method introduces noise-like artifacts to the JPEG-compressed picture.

3. DETECTING JPEG COMPRESSION ANTI-FORENSICS

In this section we describe a procedure that enables the detection of attacked images, which have been forged according to the method in [5], with the purpose of hiding the traces of JPEG compression.

The method is based on re-compressing the available image at different quality factors Q_A and analyzing the properties of these re-compressed images as a function of Q_A. First, in Section 3.1 we consider an ideal case in which the original JPEG-compressed image at quality factor Q is available to the forensic analyst. In this scenario, we study the mean-square-error distortion between the attacked image, after re-compression, and the original JPEG-compressed image. We show that the distortion due to the dithering signal is completely suppressed only if $Q_A = Q$. When $Q_A > Q$, the high-frequency components of the dithering signal are retained, causing noise-like artifacts in the re-compressed image. Based on this observation, we address a realistic scenario in Section 3.2, in which the forensic analyst does not have access to the original JPEG-compressed image. In this case we propose a detection algorithm which is based on measuring the noisiness of re-compressed images, as expressed by means of the total variation.

3.1. MSE distortion analysis

First, we analyze the mean-square-error distortion for a single DCT coefficient subband. With reference to the scheme depicted in Figure 1, let X be the value of a DCT coefficient in the original (uncompressed) image. During JPEG compression, X is quantized using a uniform quantizer $\mathcal{Q}$ with quantization step size q, thus producing $\tilde{X}$, which is available to the attacker. In order to remove the traces of quantization, the attacker adds a suitable dithering signal Z, thus producing Y. Indeed, the dithering signal is chosen in such a way that the p.d.f. $p_Y(y)$ is undistinguishable from $p_X(x)$.

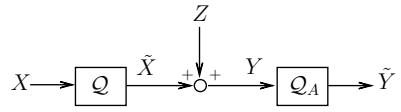

Fig. 1. Scheme of the re-quantization of a DCT coefficient.

The forensic analyst re-quantizes the coefficient Y using a quantizer $\mathcal{Q}_A$, with quantization step size q_A, and produces a dequantized coefficient $\tilde{Y}$. We are interested in measuring the mean-square-error distortion D between $\tilde{X}$ and $\tilde{Y}$, as a function of the quantization step size q_A. That is,

$$\begin{aligned} D(q_A) &= E\left[(\tilde{X} - \tilde{Y})^2\right] \\ &= \sum_{k=-\infty}^{+\infty} p_k \left[\int_{-\frac{q}{2}}^{+\frac{q}{2}} (\tilde{x}_k - \mathcal{Q}_A(\tilde{x}_k + z))^2 f_Z(z)\, dz\right], \quad (1) \end{aligned}$$

where $\tilde{x}_k = kq$, the expectation is taken with respect to the joint distribution of $\tilde{X}$ and Z, $f_Z(z)$ is the probability density function of the dithering noise Z (see [5]), and

$$p_k = \int_{kq-\frac{q}{2}}^{kq+\frac{q}{2}} f_X(x)\, dx \quad (2)$$

is the probability of the original coefficient X to fall in the k-th quantization bin. Notice that the output of $\mathcal{Q}_A$ assumes values at integer multiples of q_A. Therefore, after a change of variables, (1) can be written as

$$D(q_A) = \sum_{k=-\infty}^{+\infty} p_k \left[\sum_{h=-\infty}^{+\infty} (kq - hq_A)^2 p_{h|k}\right], \quad (3)$$

where

$$p_{h|k} = \int_{hq_A-\frac{q_A}{2}}^{hq_A+\frac{q_A}{2}} f_Z(z - kq)\, dz \quad (4)$$

is the probability of quantizing a dithered sample to the h-th bin of $\mathcal{Q}_A$, given that the original sample was quantized to the k-th bin of $\mathcal{Q}$. This is illustrated in Figure 2. Notice that the support of $f_Z(z - kq)$ is the interval $\left(kq - \frac{q}{2}, kq + \frac{q}{2}\right]$.

From (3), we observe that when $q_A \to 0$, i.e. re-quantization is almost lossless, $D(q_A) \to \sigma_Z^2$, the variance of Z. On the other hand, when $q_A \to \infty$, Y is always quantized to zero. Therefore, $D(q_A) \to \sigma_{\tilde{X}}^2$, i.e. the variance of $\tilde{X}$. $D(q_A)$ in (3) is zero when $q_A = \frac{k}{h} q$ *for all the values* h, k for which $p_{h|k} > 0$. This is achieved when $q_A = q$, such that $p_{h|k} = 1$ when $k = h$ and $p_{h|k} = 0$ otherwise. It can be easily seen from Figure 2 that this corresponds to the case when all the noise Z added to the coefficients in the k-th bin is re-absorbed by the quantized values $kq_A = kq$, resulting in $D(q_A) = 0$.

When $q_A \neq q$, re-quantization does not suppress distortion completely, i.e. $D(q_A) > 0$. Specifically, when $q_A > q$, the dithering signal is mostly canceled. Conversely, when $q_A < q$, new non-empty bins in the histogram of $\tilde{Y}$ are created, due to the dithering

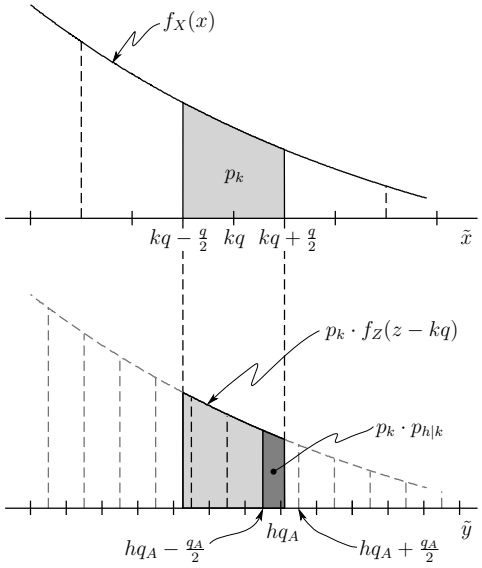

Fig. 2. Re-quantization of a dithered DCT coefficient (originally quantized with a quantization step size q), with a quantization step q_A. The anti-forensic dither recovers the original distribution of the coefficient $f_X(x)$. When re-quantized, the noise in the k-th quantization bin is redistributed into several bins.

signal Z leaking to neighboring bins (see Figure 2). In other words, the noise Z is more accurately reproduced in $\tilde{Y}$.

Due to the additive nature of MSE distortion, it is possible to generalize the analysis above from individual DCT subbands to the whole image. Hereafter we consider the widely used JPEG quantization tables suggested by the Independent JPEG Group [10]. In the IJG scheme, quantization matrices $\mathbf{Q}$ are obtained by properly scaling a template table $\mathbf{Q}_t$ by a 100-points quality factor $Q = 1, \ldots, 100$. Therefore, re-compression is driven by a quality factor Q_A. For each of them, a corresponding quantization step $q_{A,i}$ is determined for the i-th DCT subband. Figure 3 illustrates the mean-square-error distortion between the re-compressed image (at quality factor Q_A) and the JPEG-compressed one, when the latter was originally compressed at $Q = 35, 60, 85$. We notice a trend similar to the one predicted for each DCT coefficient subband, where the distortion is minimized when $Q = Q_A$ (thus $q_i = q_{A,i}$ for all i). The distortion is not exactly zero due to rounding and truncation of pixel values.

In the spatial domain, for $Q_A < Q$ the re-compressed image does not reveal the traces of the dithering signal, which is mostly suppressed together with the high frequency components of the underlying image. On the other hand, for $Q_A > Q$, the dithering signal is somewhat preserved, and transformed back to the spatial domain, thus resulting in a noisier image. This observation triggers the intuition for the detection method illustrated in the next section.

3.2. JPEG anti-forensics detection based on TV

The above analysis suggests that it is possible to identify an attacked image by quantifying how "noisy" the image is after re-quantization. We employ the total variation (TV) of the image (the ℓ_1 norm of the spatial first-order derivatives) as a blind measure of noisiness [9].

The anti-forensic detection algorithm re-compresses the doubted image using different quality factors Q_A. For each re-compressed

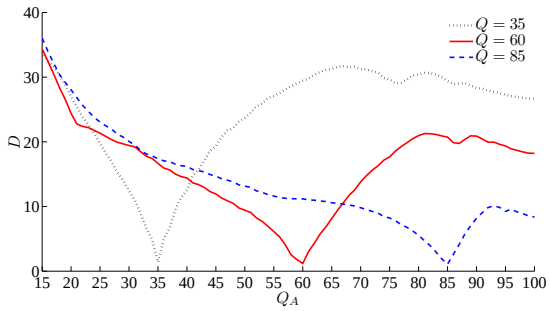

Fig. 3. MSE distortion between a JPEG image quantized with quality Q, and its re-compressed version at quality Q_A. The distortion is zero when $Q_A = Q$.

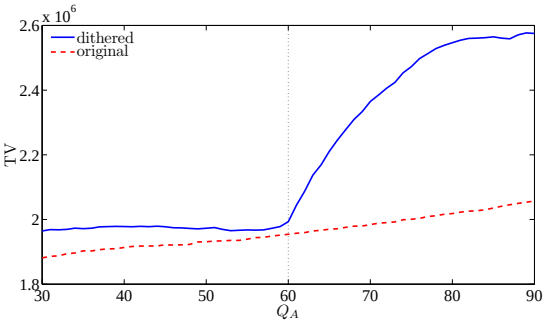

Fig. 4. Total variation as a function of the re-compression quality factor, for two versions of the *Lenna* image.

image, the TV is computed. Figure 4 shows the TV as a function of Q_A for two versions of the *Lenna* image. The dashed line corresponds to the genuine, uncompressed image. Not surprisingly, the TV increases smoothly with Q_A. To generate the solid line, instead, the *Lenna* images has been compressed (with $Q = 60$) and a dithering signal has been added to restore the original distribution of the DCT coefficients. The apparent slope change in $Q_A = 60$ is due to the fact that noise starts being visible right after Q_A exceeds Q. We claim that, analyzing the $\mathrm{TV}(Q_A)$ curve, it is possible to detect whether an image has been attacked and, in this case, to find the original Q.

In order to decide whether an image has been attacked, we consider the first order backward finite difference signal $\mathrm{TV}^{(n)}(Q_A)$ with lag n, obtained from the total variation curve as:

$$\mathrm{TV}^{(n)}(Q_A) = \frac{1}{n}\left(\mathrm{TV}(Q_A) - \mathrm{TV}(Q_A - n)\right) \qquad (5)$$

We deem an image to have been anti-forensically attacked if $\max(\mathrm{TV}^{(n)}(Q_A)) > \tau_n$ for a lag n and for some value of the threshold τ_n. In this case, we also estimate the quality factor $\hat{Q}$ of the JPEG-compressed image as:

$$\hat{Q} = \left(\arg\max_{Q_A} \mathrm{TV}^{(n)}(Q_A)\right) - n. \qquad (6)$$

4. EXPERIMENTAL RESULTS

We carried out a large-scale test of the algorithm described in Section 3.2 on 1338 images of the Uncompressed Color Image Database

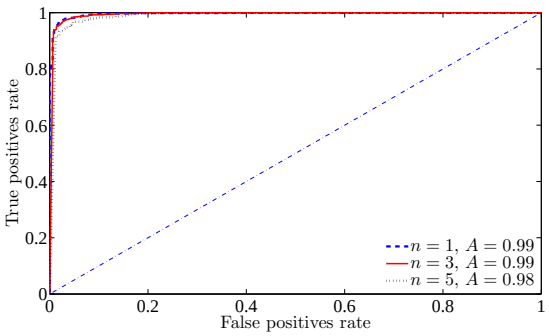

Fig. 5. ROC curve for different values of the difference lag n in (5), with the corresponding areas under the curve A.

(UCID) [11]. All the pictures in this dataset have a resolution of 512×384. Without loss of generality, we considered the luma component only.

We carried out two kinds of experiments. In the first one, we evaluated the detectability of attacked imaged. That is, we wanted to verify whether the proposed method was able to detect in which cases a dithering signal has been added to conceal the traces of JPEG compression. To this end, we compressed half of the UCID images at a quality factor chosen at random in the interval $[30, 95]$. In order to restore the original statistics of the DCT coefficients, we added an anti-forensic dithering signal according to the method in [5]. Then, we used our algorithm to label each test image either as attacked (positive case) or as an uncompressed (negative case) picture. The performance of the detector depends on the choice of n in (5) and on the threshold τ_n. By changing the value of τ_n, it is possible to trace a receiver operating characteristic (ROC) curve for a given lag n. These curves are reported in Figure 5. To avoid cluttering the picture, we show only three values of n, with the corresponding areas under the curve A. Notice that in all the cases $A \geq 0.98$. As an example, using the specific parameter setting $n = 3$ and $\tau_3 = 36 \cdot 10^3$, the detection accuracy (the percentage of correct decisions taken by our algorithm) is 97%. Using these parameters, we are able to find, for images deemed to be attacked, an estimate $\hat{Q}$ of the original quality factor Q with an average error $\mathcal{E}_{\text{avg}} = E[\hat{Q} - Q]$ equal to -0.11 units. Notice that only in 0.78% of the cases of true positive detection we had an absolute error $\mathcal{E} = |Q - \hat{Q}| > 5$ units. The negative bias in $\hat{Q}$ is due to the noisiness of the n-lag difference in (5), which most of the times reaches its maximum only one or two units after the actual Q.

In the second experiment, we aimed at analyzing more precisely how the detection performance varies as a function of the original Q. To this end, we performed an experiment similar to the one described above, but at each round all the UCID images were compressed with a single quality factor $Q \in [30, 45, 60, 75, 85, 90, 95]$ and then dithered. Table 1 reports the values of recall (i.e. the fraction of attacked images that were detected by the proposed method), the average error ($\mathcal{E}_{\text{avg}}$), the maximum absolute error ($\mathcal{E}_{\max}$) and the error standard deviation ($\mathcal{E}_{\text{sd}}$). We also report the fraction of absolute errors larger than 5 ($\mathcal{E}_5$). These values are obtained with the same setting of n and τ_n as in the first experiment. Notice that the recall is above 96% and the errors larger than 5 are below 5% when $Q \in [30, 90]$. We observe that the performance deteriorates when the original JPEG quality is very high. Indeed, when $Q \geq 95$, JPEG compression is almost lossless, and the amount of dithering required to restore the original statistics is very low, resulting in a nearly invisible anti-forensic noise.

Q	Recall [%]	$\mathcal{E}_{\text{avg}}$	$\mathcal{E}_{\max}$	$\mathcal{E}_{\text{sd}}$	$\mathcal{E}_5$ [%]
30	99.85	0.02	10	0.42	0.07
45	98.36	-0.29	25	2.86	1.45
60	98.58	-1.51	40	7.61	3.89
75	98.66	-0.88	55	6.82	1.64
85	98.06	-1.32	65	9.02	2.09
90	96.26	-3.34	70	14.70	4.86
95	9.19	-41.94	75	32.45	90.66

Table 1. Performance of the anti-forensics detection for different quality factors Q of the JPEG-compressed image.

5. DISCUSSION

The statistical traces left by JPEG compression can be removed by inserting a proper dithering signal into the DCT coefficients. However, this process leaves noise-like artifacts in the spatial domain which can be detected. We have proposed a detector of anti-forensically attacked images, which also estimates the original JPEG quality factor Q. We have found that our algorithm is effective over a broad range of quality factors. The proposed method could be extended to estimate the original JPEG quantization matrix, by working on each DCT subband separately. Also, this work provides interesting insights on the design of alternative anti-forensic techniques that would not be detected by the proposed method.

6. REFERENCES

[1] Z. Fan and R. L. de Queiroz, "Identification of bitmap compression history: JPEG detection and quantizer estimation," *IEEE Trans. Image Process.*, vol. 12, no. 2, pp. 230–235, February 2003.

[2] J. Lukáš and J. Fridrich, "Estimation of primary quantization matrix in double compressed JPEG images," *Proc. of Digital Forensic Research Workshop*, 2003.

[3] J. He, Z. Lin, L. Wang, and X. Tang, "Detecting doctored JPEG images via DCT coefficient analysis," in *European Conf. on Computer Vision*, Graz, Austria, May 2006, pp. 423–435.

[4] J. Fridrich, D. Soukal, and J. Lukás, "Detection of Copy-Move Forgery in Digital Images," in *Proc. of Digital Forensic Research Workshop*, Cleveland, USA, August 2003.

[5] M.C. Stamm, S.K. Tjoa, W.S. Lin, and K.J.R. Liu, "Anti-forensics of JPEG compression," in *Proc. of the Int. Conf. on Acoustics, Speech, and Signal Processing*, Dallas, TX, USA, April 2010.

[6] M.C. Stamm, S.K. Tjoa, W.S. Lin, and K.J.R. Liu, "Undetectable image tampering through JPEG compression anti-forensics," in *Proc. of the Int. Conf. on Image Process.*, Hong Kong, September 2010, pp. 2109 –2112.

[7] G. Valenzise, M. Tagliasacchi, and S. Tubaro, "The cost of JPEG compression anti-forensics," in *Proc. of the Int. Conf. on Acoustics, Speech, and Signal Processing*, Prague, Czech Republic, May 2011.

[8] H. Farid, "Exposing digital forgeries from JPEG ghosts," *IEEE Trans. Inf. Forensics Security*, vol. 4, no. 1, pp. 154 –160, March 2009.

[9] L.I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, no. 1-4, pp. 259–268, 1992.

[10] "The independent JPEG group," http://www.ijg.org/.

[11] G. Schaefer and M. Stich, "UCID: an uncompressed colour image database," in *Proc. SPIE: Storage and Retrieval Methods and Applications for Multimedia*, 2004, vol. 5307, pp. 472–480.