Universal Counterforensics of Multiple Compressed JPEG Images

Mauro Barni, Marco Fontani, and Benedetta Tondi

Dept. of Information Engineering and Mathematics,
University of Siena, Siena, Italy
barni@dii.unisi.it, marco.fontani@unisi.it, benedettatondi@gmail.com

Abstract. Detection of multiple JPEG compression of digital images has been attracting more and more interest in the field of multimedia forensics. On the other side, techniques to conceal the traces of multiple compression are being proposed as well. Motivated by a recent trend towards the adoption of universal approaches, we propose a counter-forensic technique that makes multiple compression undetectable for any forensic detector based on the analysis of the histograms of quantized DCT coefficients. Experimental results show the effectiveness of our approach in removing the artifacts of double and also triple compression, while maintaining a good quality of the image.

1 Introduction

In the last years the supremacy of images as the most direct and trustful mean of communication has been threatened by the widespread availability of photo manipulation software. The research community has tackled with this problem in several ways: at the beginning, only active approaches like image watermarking and digital signatures were investigated; lately, Multimedia Forensics emerged as the discipline that tries to infer as much information as possible about the processing history of an image without having access to any further information.

Shortly afterwards the birth of Multimedia Forensics, counter-forensic methods started to be investigated as well, whose goal is to conceal the evidence of manipulation when a user alters an image by means of processing tools for malicious purposes. As we will discuss in the paper, most of the state-of-the-art approaches are targeted at deceiving a specific forensic detector, by erasing the traces it searches for. On the contrary, universal approaches exist that, instead of deceiving a fixed detector, attempt at making the doctored image undetectable for *any* detector, at least within a certain class [8]. While the literature is rich of targeted counter forensic schemes, the first universal approaches have been proposed only recently [2, 5]; this fact witnesses the higher complexity of developing universal CF methods. In this paper, we propose a universal counter forensic approach for concealing traces of multiple JPEG compression. From a forensic point of view, JPEG compression is one of the most important stages in the processing chain of a digital image, because it leaves peculiar statistical footprints that can be used as a telltale of tampering. In particular, traces left by *multiple* JPEG compressions are usually a powerful tool in analyzing the authenticity of an image. Therefore, the method proposed in this paper establishes a new challenge for future forensic detectors.

The paper is organized as follows: Sect. 1.1 briefly reviews related works and clarify our contribution; a summary of the effects of multiple JPEG compressions in the frequency domain is given in Sect. 2. In Sect. 3 we introduce the theoretical framework behind our method, while Sect. 4 is devoted to a detailed description of all the phases of the algorithm. Experimental validation is finally reported in Sect. 5.

1.1 Related Works and Contribution

The interest of forensic researchers in the detection of multiple compressions is motivated by the fact that when JPEG images are manipulated by a photo-editing software and later re-saved in JPEG format, artifacts are introduced in the image. Popescu et al. [14] showed that multiple quantization steps introduce periodic artifacts into the histograms of DCT coefficients, which can then be searched for in order to detect a manipulation. Inspired by this work, many techniques for detecting double JPEG compression have been proposed, which analyze the first order statistics of the DCT coefficients, e.g., [13]. Many recently proposed forensic approaches rely on the analysis of the first significant digits (FSD) of the DCT coefficients. Specifically, the distribution of the FSDs in the frequency domain is investigated in order to tell apart single compressed images from double compressed [10] and, more in general, multiple compressed ones [11]. On the other hand, counterforensic schemes have been developed in order to remove or disguise the artifacts of multiple compression in the FSD distributions, like in [12]. A unifying characteristic of these anti-forensic methods is that they are targeted to deceive a specific forensic detector (*targeted approaches*). As such, they do not guarantee that a possible different detector, even based on the analysis of the same statistic, would be defeated in turn; in fact, the analyst may develop a modified version of the detector that is robust to the counter-forensic approach, thus pushing forward the cat-and-mouse game.

To overcome this limitation, that is inherent in the use of targeted couter-forensic techniques, a recent trend has turned to *universal approaches*, see [5] and [2], for which the optimality under certain criteria is discussed. In particular, in [2], a universal technique for concealing manipulations of gray-scale images in the spatial domain is proposed. The general idea behind this method is the following: in order to avoid the introduction of new traces, the attacker should try to make *the statistics* of the image as close as possible to the statistics of an untouched image. In this way, in principle, the tampering would be statistically undetectable for any forensic detector, whatever are the traces it looks for in the image, thus definitively ending the cat and mouse loop. The rigorous theoretical framework behind such an approach is provided in [1], where the general problem of hypothesis testing in presence of an adversary is addressed, thus opening the door to the applicability of the method to many different scenarios, like watermarking, fingerprinting, spam filtering, secure classification, reputation systems, and so on. Of course, it is very hard to devise a universal technique that is capable of “fixing” every statistic of the processed signal; in [1] the first steps have been taken by assuming some limitations on the resources available to the forensic investigator. Specifically, the optimum strategies for the forensic analyst and the attacker are found under the assumption that the analyst relies on first order statistics to perform the decision. Although this limitation might sound restrictive, it holds in many realistic scenarios, like in the image forensic scenario, and permits to cope with an entire class of forensic detectors. Leveraging on the theoretical results in [1], the universal attacking strategy proposed in [2] is optimum against any forensic detector based on the analysis of the histogram of the image.

In this paper, we want to extend the *universal* counterforensic algorithm developed in [2] to the frequency (DCT) domain, at the purpose of countering the detection of multiple JPEG compressions. In order to do so, we will exploit the results which come from the extension of the adversarial hypothesis testing to the case of multiple observations, addressed in [3]. To the best of our knowledge, the proposed techniques is also the first that has been applied for concealing traces left by any number of compression stages. On top of that, the very good results obtained in terms of visual quality of the attacked image are a strength of our method.

2 JPEG-Based Image Forensics and Watson’s model

The JPEG standard is today the most widely used method for storing digital images. Despite its lossy nature, JPEG compression is designed not to introduce annoying artifacts in pixels, at least for reasonable compression ratios. On the other hand, appreciable artifacts are introduced in the Discrete Cosine Transform (DCT) domain, where most of the computation is carried, and this fact fostered the development of a whole branch of image forensic techniques. For this reason, we find it worthy to introduce the basic concepts of JPEG coding and briefly describe how JPEG-based forensic algorithms work.

To begin with, we revisit the procedure of compression of a gray-scale image according to the JPEG standard. As a first operation, the input is divided into blocks of 8×8 pixels each. For each block, the two dimensional DCT is computed. Let $X(i,j)$, $1 \leq i, j \leq 8$, denote the DCT coefficient in position (i,j) of the block. The DCT coefficients are then quantized into integer-valued quantization levels $X_q(i,j)$ as follows:

$$X_q(i,j) = \text{sign}(X(i,j))\text{round}\left(\frac{|X(i,j)|}{q(i,j)}\right), \tag{1}$$

where the quantization steps $q(i,j)$ are given by a predetermined (chosen) quantization matrix $Q = \{q(i,j)\}_{i,j=1}^{8}$. After quantization, the values $X_q(i,j)$ of the block are ordered by zig-zag scanning and finally compressed by a lossless encoder. Viceversa, in the decompression procedure, first the bit stream is decoded, and the integer coefficients $X_q(i,j)$ are rearranged back into blocks. Then,

the de-quantized DCT coefficients are recovered by multiplying the coefficients with the corresponding entry of the quantization matrix, i.e., $X_q(i,j) \cdot q(i,j)$. Due to the quantization step, the compression procedure is not invertible and the dequantized coefficients assume only values which are integer multiples of the corresponding quantization step. Finally, the inverse DCT of each block is computed and the result is rounded and truncated so that the integer values range in $[0,255]$. The quantization factor is the parameter which determines the amount of approximation introduced by the compression, thus affecting both the compression ratio and the quality of the reconstructed image. Typically, the quantization matrix is fixed by selecting a quality factor (QF), in $[0,100]$; a high quality factor corresponds to a high quality of the reconstructed image, which also means lower values for the quantization coefficients.
Now let us suppose that an image is compressed twice. Let $X_{q_1}(i,j)$ denote the quantized value in position (i,j) after the first encoding with quantization step $q_1(i,j)$. When the image goes through a second compression stage, the resulting quantization level is:

$$X_{q_2}(i,j) = \text{sign}(X_{q_1}(i,j))\text{round}\left(\frac{|X_{q_1}(i,j) \cdot q_1(i,j)|}{q_2(i,j)}\right), \tag{2}$$

where $q_2(i,j)$ is the quantization step of the second encoding. Popescu et al. [14] observed that double quantization, and more in general consecutive quantizations, introduce periodic artifacts in the histogram of DCT coefficients. Such a periodic pattern depends on the ratio between the quantization steps, that is, on the ratio between the quality factor of first and second compression. More specifically, when the step size decreases (i.e., QF increases) some bins in the histograms are empty, whereas when it increases (i.e., QF decreases) some bins contain a large number of samples and some other bins only few. It is proper to observe that forensic analysers have usually to deal with the first kind of artifacts, since in many application the goal of the attacker is to pass off a lower quality image as an image of higher quality. For this reason, in this paper we consider the case of multiple compression with increasing quality factors; however it is proper to stress that, being universal, our technique can equivalently be applied in the other case.

Below, we give a brief description of the Watson's model that we will use to characterize the distortion constraint of the DCT coefficients, that is for the estimation of the Just Noticeable Difference (JND) of the block-based DCT coefficients [17].

The Watson's DCT-based Visual Model This model establishes a link between modifications in the (unquantized) DCT domain and their impact in the pixel domain. To account for the sensitivity of the Human Visual System (HVS) to different frequencies, the model defines a *sensitivity table*, which is an 8×8 matrix W whose element $W(i,j)$ gives the amount of modification for coefficient (i,j) that produces a JND in the pixel domain. Lower values in the matrix correspond to higher sensibility for the HVS to that frequency. For our experimental evaluations, we use the matrix of standard values provided in

[6]. The sensitivity table is the simplest estimation of the JND, as it does not take into account the local properties of the image. To obtain a more accurate evaluation of the JND for a DCT coefficient we need to consider two more effects: the *luminance masking* and the *contrast masking*.
The luminance masking effect is due to the fact that, according to the HVS, a bright background hides more noise that a dark background. To account for such an effect, Watson's model modifies the matrix for each block of the image on the basis of the value of the DC coefficient (mean luminance intensity of the block). The refined threshold for the (i, j) DCT coefficient of the k-th block is given by

$$T_L(i, j; k) = W(i, j) \cdot \left[\frac{C(1, 1; k)}{\overline{C}}\right]^{\alpha}, \tag{3}$$

where $C(1, 1; k)$ is the DC value of the k-th block, $\overline{C}$ is the mean intensity of the image, and α is a constant. The value suggested by Watson is $\alpha = 0.649$.
Watson's model further refines the estimation of the JND by considering also the contrast masking effect. This is done by evaluating the influence that the AC energy has in the DCT coefficients. The threshold for the DCT coefficient (i, j) of the k-th block is then given by:

$$T(i, j; k) = \max\{T_L(i, j; k), |C(i, j; k)|^{\eta} \cdot T_L(i, j; k)^{1-\eta}\}, \tag{4}$$

where η is a constant between 0 and 1 (Watson suggests $\eta = 0.7$).

3 Theoretical Background for the Proposed Work

Before diving into the proposed scheme we need to give a brief overview of the theoretical framework behind it. A universal counter-forensic method is derived by modelling and studying the struggle between the forensic analyst (or defender D) and the attacker (A). In the general case, D wants to tell apart modified and untouched images, while A aims at making the decision fail. More specifically, after generating the manipulated image with the desired properties, A wants to slightly modify it in such a way to prevent D from detecting manipulations, while respecting a distortion constraint. A first theoretical analysis of such an interplay between A and D has been proposed in [1], under the assumption that D only considers first order statistics for the analysis. By formulating the problem as a game and solving it by means of game theoretical tools, the authors derived the optimum strategies for both players (D and A). Interestingly, since there is a dominant strategy for the defender, the optimum attacking strategy results from the resolution of an optimization problem, obtained by assuming that D plays its dominant strategy. These theoretical findings have been put in practice in the image forensic scenario for designing a universal counter-forensic (U-CF) method, able to fool *any* forensic detector based on the image histogram, whatever the trace it looks for (e.g., footprints left by contrast enhancement operations, cut and paste, splicing, and so on) [2]. According to this technique, starting from the processed image $\boldsymbol{y}$ with histogram h_Y, A produces the attacked image $\boldsymbol{z}$ in three steps: *retrieval* of a target histogram from a database of untouched histograms, *computation of the optimum mapping* and *implementation of the mapping* into the image. The overall scheme is preserved in the algorithm proposed in this paper; however, working in the DCT domain poses several new challenges that need to be solved, especially in the second and the third phase.

With a specific reference to the JPEG forensic scenario, the D/A interplay can be described as follows: on one side, D wants to tell apart single compressed from multiple compressed images while, on the other side, A aims at hiding the effect of multiple compressions so that the image looks like a single compressed one. We assume that, as an extension of the previous case, the defender relies the decision on the analysis of the histograms of the DCT coefficients; this hypothesis actually holds for most of the existing forensic tools. The main difference with respect to the previous case is that now the forensic analyst has to combine the information brought by 64 histograms, one for each DCT frequency (i, j). At the same time, A has 64 histograms to act upon in order to fool D, while preserving the constraint on the visual distortion of the image *in the spatial domain.* It should now be evident that, although having similarities with the analogous problem in the pixel domain, the case under analysis cannot be treated with the theoretical tools proposed in [1]. Instead, the detection of JPEG multiple compression in frequency domain finds an appropriate background in [3], where the case of *multiple observations* is considered, and then D bases the decision on a number S of features (or summaries) each one extracted from an observed sequence which describes the status of a system. This is exactly the case with the JPEG forensic methods, since they separately analyse coefficients belonging to different DCT frequencies. Let $\boldsymbol{x}$ be a reference single compressed image on which D bases the decision; we denote by $h_{X_{ij}}$ the histogram of the quantized DCT coefficients in position (i, j) and with $v_{X_{ij}}$ the normalized one. Moreover, we indicate with v_{ij}, for $1 \leq i, j \leq 8$, the normalized histograms of the image under analysis. Because of the decorrelation property of the DCT transform, the dependence among DCT coefficients in different subbands is low (intrablock dependence), and then we can approximately assume them independent. Exploiting this fact in the analysis in [3], it is easy to show that the optimum log-likelihood function of the Neyman-Pearson test performed by D (under the assumption of resources limited to the first order analysis) can be noticeably simplified, becoming

$$\sum_{i,j=1}^{8} \mathcal{D}(v_{ij} || v_{X_{ij}}), \tag{5}$$

where $\mathcal{D}(\cdot||\cdot)$ is the Kullback-Leibler (KL) divergence. Given two probability distributions R and Q defined over the same alphabet, the KL divergence is defined as

$$\mathcal{D}(R||Q) = \sum_{a} R(a) \log \frac{R(a)}{Q(a)}.$$

According to the game theoretical analysis, expression (5) is the optimum objective function that A has to minimize in producing the forgery [3].

4 Universal JPEG Counter-Forensic Algorithm

In this section we describe in detail each phase of the proposed counter-forensic algorithm. The attacker owns an image which has been compressed two or more times with increasing quality factor, i.e., with $QF_k > QF_{k-1}$, where k denotes the number of times that the image has undergone a compression stage. In order to pass off the image as a single compressed image, the attacker runs the universal counter-forensic algorithm schematized in Figure 1. Before entering the details of each step, let us introduce some necessary notation. For simplicity, the capital letter X is used to denote image $\boldsymbol{x}$ in the transformed domain and X_q for the quantized version. In addition, $X_q(i,j)$ indicates the transformed coefficient in position (i,j) of a generic block; when a particular block k is addressed we denote it by $X_q(i,j;k)$.

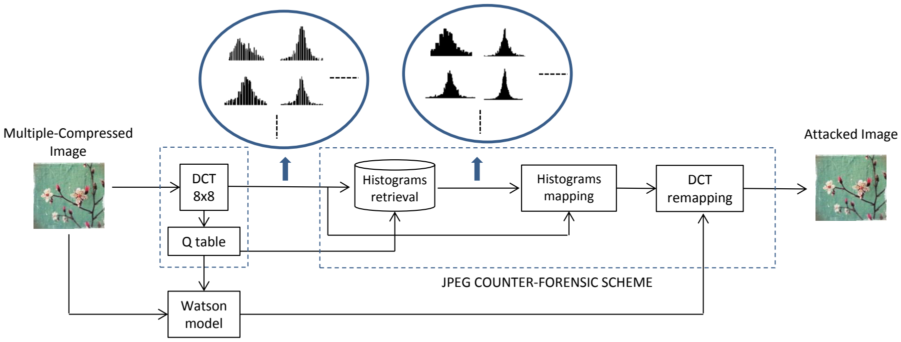


Fig. 1. The block scheme of the proposed universal JPEG Counter-Forensic algorithm.

4.1 Retrieval phase

The proposed counter-forensic scheme assumes that the attacker owns a database (DB) of images that have been JPEG compressed only once. Given the multiple compressed image Y_q with quantization matrix $Q_Y = \{q_Y(i,j)\}_{i,j=1}^{8}$, A searches in the adapted DB of images the one whose vector of DCT histograms is the most similar to the histogram vector $\boldsymbol{h}_Y = (h_{Y_{11}}, h_{Y_{12}}, ..., h_{Y_{88}})$.
For any frequency subband (or block position) (i,j), the similarity between an histogram $h_{X_{ij}}$ and $h_{Y_{ij}}$ is measured by the chi-square distance χ^2, defined as follows:

$$\chi^2(h_{X_{ij}}, h_{Y_{ij}}) = \frac{1}{2} \sum_{m \in \mathcal{C}} \frac{(h_{X_{ij}}(m) - h_{Y_{ij}}(m))^2}{(h_{X_{ij}}(m) + h_{Y_{ij}}(m))},$$

where $\mathcal{C}$ denotes the set of all the values taken by the DCT coefficients [1]. While in the spatial domain these values range from 0 to 255 (pixel values), in the frequency domain the DCT coefficients vary in $[-1024, 1023]$.

[1] Experiments show that using χ^2 in place of $\mathcal{D}$ in this phase lightens the computation without significantly affecting the results.

We can distinguish between two methods for performing the choice of the 64 DCT target histograms, depending on how the overall χ^2 distance is computed:

- *average distance*: for each image X in the DB the attacker sums each contribution $\chi^2(h_{X_{ij}}, h_{Y_{ij}})$ provided by each couple of histograms[2] and chooses the vector of 64 DCT histograms minimizing the overall distance. That is, the attacker looks for the vector $\boldsymbol{h}_X$ which minimizes $\sum_{(i,j)} \chi^2(h_{X_{ij}}, h_{Y_{ij}})$;
- *separate distance for each subband*: for each DCT subband, A searches the DB for the histogram associated with the minimum of the χ^2 distance; i.e., for each (i, j), the A chooses the $h_{X_{ij}}$ which minimizes $\chi^2(h_{X_{ij}}, h_{Y_{ij}})$.

It is evident that, in the second case, the target DCT histograms retrieved from the DB probably belong to different images, i.e., different histogram vectors of the DB. However, in our model, which is confined to the analysis of first-order statistics, this fact does not arise any contradiction, being consistent with the optimum strategy for A. Besides, it must be stressed that performing the choice in the second way allows to find, for each subband, target histograms which are closer to the processed ones with respect to those found in the first way.

There are some important considerations we need to do about the retrieval phase. First, we notice that the attacker can hardly resort directly to a DB of single compressed images, since the corresponding quantization matrices would be probably different from the input quantization matrix Q_Y. Instead of storing thousands versions of the same image quantized with all possible tables, the attacker can more practically consider a DB of never-compressed images and, depending on the quantization matrix of the Y_q under analysis, adapt the DB "on-the-fly". This means that, for a given input image Y_q, the attacker *simulates* the (single) JPEG compression by quantizing the DCT coefficients according to the input quantization matrix Q_Y. The second observation still concerns practicality: since the search is conducted on the vector of DCT histograms and not on images, only the histograms of unquantized DCT coefficients need to be stored in the DB. This allows to reduce both the size of the dataset and the execution time.

4.2 Mapping phase

According to our previous discussion, in this phase the attacker has to determine the histograms $v_{Z_{i,j}}$ which minimizes the quantity $\sum_{(i,j)} \mathcal{D}(v_{Z_{ij}} || v_{X_{ij}})$, subject to a distortion constraint imposed in order to maintain the final image visually similar to the initial one. In order to characterize this constraint in the frequency domain we rely on the concept of Just Noticeable Distortion (JND), defined as the maximum modification of the DCT coefficients which is visually undetectable. Then, it is reasonable to take the JND as maximum value for the distortion that A can introduce in the coefficients of the transformed image Y. A commonly used model for the JND is Watson's model [17], described in Sect. 2, which provides a 8×8 sensitivity matrix $W = \{W(i,j)\}_{i,j=1}^{8}$. Each entry

[2] Each contribution may be possibly weighted by some coefficients in order to give more importance to the low frequency coefficients.

of the matrix $W_q(i,j)$ provides the maximum amount of distortion which can be introduced in the quantized DCT coefficients of the subband (i,j) without generating annoying artifacts. Let $W_q = \{\text{round}\,(W(i,j)/q_Y(i,j))\}_{i,j=1}^{8}$ denote the quantized Watson's matrix, approximated to integer values[3]. The maximum distortion for the (i,j) coefficient is given by $K(i,j) = W_q(i,j) \cdot D_{\max}$ for some $D_{\max} \geq 1$ (larger $D_{\max}$ allow to obtain more accurate mapping at the price of a higher visual distortion). Interestingly, since distortion constraints are defined subband-wise, the problem can be solved as 64 separate minimizations:

$$\min_{|Z(i,j)-Y(i,j)| \leq K(i,j)} \mathcal{D}(v_{Z_{ij}} || v_{X_{ij}}), \quad \forall (i,j), 1 \leq i,j \leq 8. \tag{6}$$

Let us focus on a single DCT subband and analyze the corresponding problem. It is useful to introduce the *transportation matrix* $N_{ij} = \{n_{ij}(m \to r)\}_{m,r=1}^{|\mathcal{C}|}$, where each term $n_{ij}(m \to r)$ indicates the number of elements in $h_{Y_{ij}}$ which must be moved from the m-th to the r-th bin. Let n_{ij} be the total number of blocks in the image (i.e., the number of DCT coefficients for each position (i,j)). Each constrained optimization problem in (6) is quite similar to the one in [2] and, similarly, can be rephrased in function of the $n_{ij}(m \to n)$ variables as follows:

$$\min_{n_{ij}(m \to r)} \sum_{r=1}^{|\mathcal{C}|} \frac{(\sum_m n_{ij}(m \to r))}{n} \cdot \log \frac{(\sum_m n_{ij}(m \to r))}{n v_{X_{ij}}(r)}, \tag{7}$$

subject to

$$\begin{cases} \sum_r n_{ij}(m \to r) = h_{Y_{ij}}(m) \; \forall i \\ n_{ij}(m \to r) = 0, \; \forall (m,r) \in \mathcal{I} : |m-r| > K(i,j) \\ n_{ij}(m \to r) \geq 0 \quad \forall m, r \\ n_{ij}(m \to r) \; \in \mathbb{N} \end{cases} \tag{8}$$

where the histogram $h_{Y_{ij}}$ and the distortion constraint were rewritten in terms of $n_{ij}(m \to r)$ variables. Solving problem (7)-(8) provides the optimum map N_{ij}^{*}, from which we obtain the final attacked histogram $h_{Z_{ij}}$ by computing $\sum_m n_{ij}^{*}(m \to r)$ for each r. Problem (7)-(8) is a convex mixed integer non-linear problem (MINLP) [4] for which a global optimum solution exists and efficient solvers are available for the resolution. It is worth observing that the number of optimization variables is given by $|\mathcal{C}|$, that is the cardinality of the alphabet of the DCT coefficients ($|\mathcal{C}| = 2048$), and it does not depend on the size of the image. This value seems to be significantly larger compared to the one in the pixel domain (i.e., 256); however, since the statistics of the DCT coefficients are usually peaked around the mean value [9], the number of variables can be noticeably reduced by cutting off the bins below $m_{\min}$ (where $m_{\min}$ is s.t. $h_{Y_{ij}}(m) = 0$ $\forall m < m_{\min}$) and above $m_{\max}$ (where $m_{\max}$ is s.t. $h_{Y_{ij}}(m) = 0 \; \forall m > m_{\min}$). Let

[3] Performing the rounding for computing W_q may cause a slight violation of the JND constraint, but it is preferable for the remapping operation.

$\mathcal{E}$ be the set of the empty bins within the interval $[m_{\min}, m_{\max}]$. It is easy to argue that the actual complexity/number of variables of the (i, j)-th minimization is $2K(i, j) \cdot ((m_{\max} - m_{\min}) - |\mathcal{E}|)$, which is usually much lower than $|\mathcal{C}|$. Moreover, since the JPEG compression quantizes more strongly the high-frequency DCT coefficients, the complexity of the minimizations will decrease at higher frequencies, because histograms will tend to cluster around zero. Experiments showed that, except for the very low-frequency subbands, the complexity of the minimizations is often smaller than the one for the spatial domain.
It is interesting to note that problem (7)-(8) has very close ties with the *transportation problem* (TP) [15]. The difference with the classical TP is that, according to the definition of the attacker's strategy, the attacker is satisfied with any distortion less than $K(i, j)$, that is, he/she is not concerned about minimizing the distortion provided that it is less that $K(i, j)$.[4]. In this way, the optimum attacking strategy in (7)-(8) provides the optimum map even when the classical transportation problem, which moves $v_{Y_{ij}}$ exactly into $v_{Z_{ij}}$, would introduce too much distortion into the image (i.e., more than $K(i, j)$).

To sum up, the mapping phase provides the attacker with the 64 matrixes N^*_{ij}, $1 \leq i, j \leq 8$; each matrix N^*_{ij} defines the modifications that must be made on the DCT coefficients in position (i, j) in order to obtain the optimum attacked histogram $h_{Z_{ij}}$.

4.3 Implementation of the mapping

After obtaining the transportation matrixes, it is necessary for the attacker to implement the mapping in such a way that reduces as much as possible the visual distortion introduced in the image. Notice that, since the forensic detector relies on the histograms of the DCT coefficients, the result of the attack in terms of detectability of the produced forgery only depends on the results of the mapping phase, and it is not affected by the modifications performed in this phase. In the following, we describe an approach that allows the attacker to implement the modifications set by the matrixes N^*_{ij}'s in a perceptually convenient way. The basic idea is to exploit the different sensitivity of the Human Visual System to the DCT coefficients of the different blocks in order to first modify the coefficients in those blocks where the HVS is less sensitive. To do so, we exploit the threshold (refined) values of the JND provided by Watson's model which, as described in Sect. 2, are indeed block-dependent. Again, modifications are implemented separately on the DCT coefficients of each frequency subband.
Below, we describe the main steps of the proposed scheme for the implementation of the transportation matrix N^*_{ij} in the generic subband (i, j):

1. Set all the coefficients as "admissible";
2. Rank the blocks based on the value of the threshold $T(i, j)$ in decreasing order: block k such that $T(i, j; k)$ is maximum is ranked first, and so on;

[4] In the transportation problem the objective function of the minimization problem would be the distortion (cost of the transportation), which in our formulation is instead a constraint.

3. For each couple of values (m, r) such that $n_{ij}(m \to r) \neq 0$ proceed as follows:
   (a) find the blocks with admissible DCT coefficients having value m;
   (b) select the first $n_{ij}(m \to r)$ according to the order established by the ranking;
   (c) substitute them with r;
   (d) remove selected coefficients from the admissible ones[5];

The procedure is applied to all the 64 DCT subbands.
Notice that, according to the above scheme, the attacker computes the thresholds of the JND only once, without updating them to account for the variations caused by incremental modifications. In principle, lower distortion can be introduced by iteratively updating the thresholds. However, since Watson's model is mainly concerned about average luminance and energy of each block, the benefit obtained by iterative updating is not relevant enough to justify the increased computational complexity, and for this reason this feature was not implemented.
At the end of the procedure, the adversary gets the transformed image Z_q with the quantized 'remapped' DCT coefficients, whose DCT histograms are, by construction, the 64 target histograms $h_{Z_{ij}}$, $1 \leq i, j \leq 8$, obtained in the mapping phase. Computing the de-quantized coefficients and applying the inverse DCT transform yields the final attacked image $\boldsymbol{z}$ in the pixel domain. The image will appear visually close to the input one, but its histograms will show traces of just one compression step.

5 Experimental validation

In this section we put the proposed technique to work, in order to show that it actually conceals the traces of multiple compression in all the histograms of the DCT coefficients. Besides, we evaluate the perceptual similarity between the input image and the one obtained after the implementation of the mapping.

To generate the database for the attacker, we computed the histograms of each DCT coefficient from 2000 grayscale uncompressed images, obtaining 64 histograms per image. Then, 25 grayscale uncompressed images were chosen from different sources for performing tests. Both the database and the test images are available on our research group website[6]. For a multiple compressed image, consistently with the notation introduced in Sect. 4, we denote by $\{QF_1,QF_2,\ldots,QF_k\}$ the quality factor used for the first, second, . . . , k-th compression step. Each test image was used to generate, using the `imwrite` function of Matlab, the following images: three double-compressed versions of the image, with quality couples $\{65, 85\}$, $\{75, 90\}$ and $\{85, 95\}$; five triple-compressed versions, with quality triplets $\{65, 85, 90\}$, $\{70, 75, 95\}$, $\{70, 80, 95\}$, $\{75, 85, 95\}$, $\{80, 85, 95\}$; for each of the above multiple-compressed images, one single-compressed image with quality given by QF_1 (these images serve to test the discrimination capability of a forensic detector).

[5] This avoids multiple substitutions of the same coefficients.

[6] `http://clem.dii.unisi.it/~vipp/index.php/download/imagerepository`

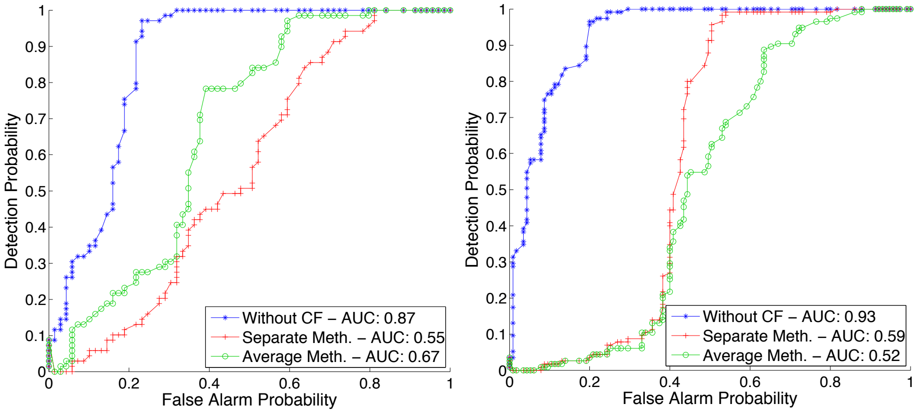
We applied the JPEG counter-forensic scheme to each of the above images, using $D_{max} = 4$; the experiment was performed using both the *separate* and *average* search (as defined in Sect. 4.1) in order to compare performance. To test the effectiveness of the proposed scheme, we implemented a simple double compression detector based on the so-called calibration technique [7]. Calibration is a procedure allowing to estimate the original distribution of a quantized signal by removing a small number of rows/column to disrupt the structure of JPEG blocks. The calibration-based detector simply works by calculating the "expected" histograms for quantized DCT coefficients and comparing them to the histograms of observed DCT coefficients in the given image.[7] If the image was compressed only once, the expected histogram is quite similar to the observed one (the χ^2 distance is used to compare histograms); on the other hand, if multiple compressions were performed, the expected histogram differs significantly from the observed one. We limit the detector to consider the first 12 DCT coefficients (in the JPEG zig-zag ordering), because higher frequency coefficients are not reliable for this kind of analysis, due to the sparsity of histograms induced by quantization. It is proper to stress again that, since the proposed scheme is *universal*, it is not tailored for deceiving this specific detector.

Let us describe the experiments we conducted. In the first experiment, the detector was used to discriminate between double-compressed and single-compressed images, generated according to points 1. and 3. of the above list. To test the performance of the proposed scheme, we computed the Receiver Operating Characteristic (ROC) curve of the detector before and after the application of our JPEG counter-forensic attack, along with the Area Under the Curve (AUC). As we can see in Figure 2(a), the detector behaves reasonably well in absence of counter-forensic schemes, while its performance dramatically drop after application of the proposed scheme. Moreover, we see that both the *separate* and *average* search methods lead to reasonably good performance in terms of deceiving the calibration-based detector, with the former slightly favored at small probabilities of false alarm, that counts the most in forensic scenarios.

In the second experiment we tried to discriminate between single- vs. triple-compressed images, both before and after application of the CF method. Results are plotted in Figure 2(b): we see that good CF performance are obtained in the leftmost part of the ROC, corresponding to low false alarm probability. For false alarm probabilities over 0.4, the detector manage to distinguish between single- and triple- compressed images even in presence of counter forensic. This fact is mainly due to the different distribution of quality factors between triple compressed and single compressed images in the considered experiments; from a forensic point of view, however, false alarm probabilities as high as 0.4 are not of interest.

Let us now turn to consider the perceptual quality of produced images. We evaluated the quality by means of the Structural Similarity (SSIM) index [16], computed between the input and the output to the proposed scheme. Results

[7] The expected histogram is obtained by estimating the histograms of unquantized coefficients (using calibration), then quantizing them according to the quantization factors available in the JPEG header of the file.



(a) (b)

Fig. 2. ROC curve for the calibration-based detector for single-vs-double (a) and single-vs-triple (b), before and after application of the proposed method.

are given in Table 1. We can confirm that using the separate search on the

Experiment	**Mean SSIM**	**Std. dev. SSIM**
Double compr - Separate	0.920	0.033
Double compr - Average	0.903	0.046
Triple compr - Separate	0.945	0.025
Triple compr - Average	0.935	0.027

Table 1. Performance of the proposed method in term of perceptual quality. Each row shows the mean and the standard deviation of the SSIM obtained for a given experiment. For double compression a total of 75 images were processed, while the number raises to 125 for triple compression.

database (as defined in Sect. 4.1) allows the attacker to obtain better results in terms of perceptual quality of the produced image. It may seem counter-intuitive to the reader that a better similarity was obtained in the case of triple compression: this is actually not surprising if we keep in mind that the similarity is computed between the input and the output of the CF scheme, and it is easier to keep fidelity to an image whose quality was not so high from the beginning (as it is a triple compressed images). A practical comparison between a multiple-compressed image and the counter-forensic version is shown in Figure 3.

6 Conclusions

In this paper we have presented a universal counter forensic technique for hiding traces of multiple compression in JPEG images. The described method is proposed as an extension of the counter forensic algorithm developed in [2] for concealing the manipulation of tampered image in the spatial domain. With

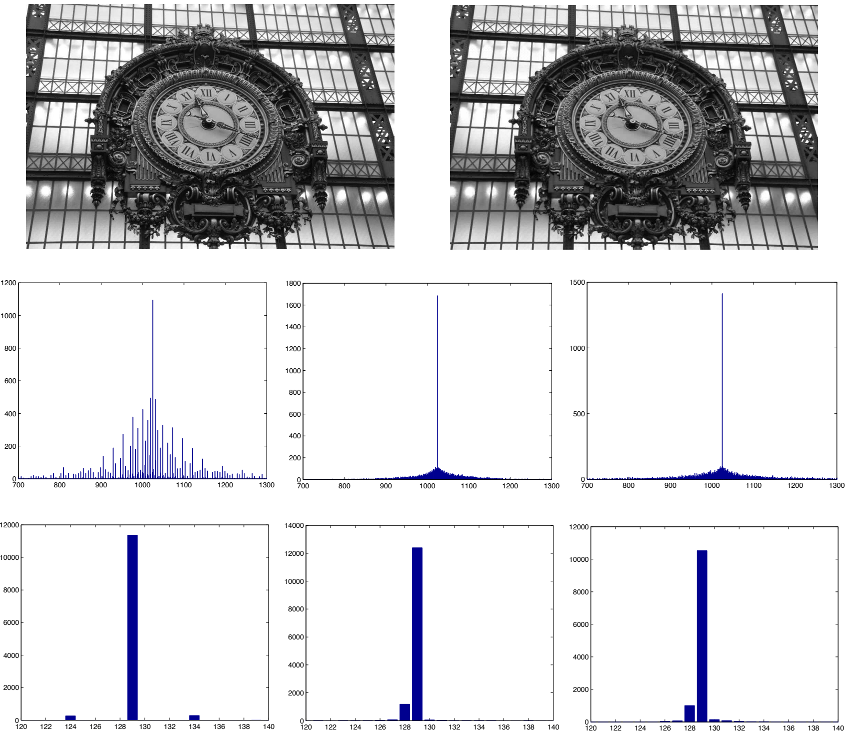


Fig. 3. Upper row: comparison between a triple-compressed image with qualities {70, 80, 95} (left) and its counter-forensic version (right). Middle row: the histogram of DCT coefficients in position (1,2) coming from the above multiple-compressed image (left), the target from the DB (middle), and the remapped version obtained with the proposed scheme (right). Bottom row: histograms of DCT coefficients in position (3,4), ordered as in the previous line.

respect to most of the state-of-the-art techniques, which aim at deceiving a targeted detector by removing the specific artifacts of multiple compression it searches for, the proposed method attempts to make the image look like a single compressed one working on the first order statistics of the image. The main strength of this method is that it can be applied for concealing any number of compression stages that the image may have undergone and with any quality factor. We showed that our method provides good results in terms of quality of the attacked image and degradation of detection performances. As a challenge for the future, it would be interesting to devise universal counter-forensic methods against detectors whose analysis is based on higher order statistics (e.g., the joint RGB histograms for color images).

7 Acknowledgments

This work was partially supported by the European Office of Aerospace Research and Development under Grant FA8655-12-1- 2138: AMULET - A multi-clue approach to image forensics.

References

1. Barni, M., Tondi, B.: The source identification game: an information-theoretic perspective. IEEE Transactions on Information Forensics and Security 8(3), 450–463 (March 2013)
2. Barni, M., Fontani, M., Tondi, B.: A universal technique to hide traces of histogram-based image manipulations. In: Proc. of MM&Sec 2012, 14th ACM workshop on Multimedia & Security. pp. 97–104. ACM, New York, NY, USA (2012)
3. Barni, M., Tondi, B.: Multiple-observation hypothesis testing under adversarial conditions. In: Proc. of WIFS 2013, IEEE International Workshop on Information Forensics and Security. pp. 91–96 (Nov 2013)
4. Bonami, P., Kilinc, M., Linderoth, J., et al.: Algorithms and software for convex mixed integer nonlinear programs. Tech. rep., Computer Sciences Department, University of Wisconsin-Madison (2009)
5. Comesana-Alfaro, P., Perez-Gonzalez, F.: Optimal counterforensics for histogram-based forensics. In: Proc. of ICASSP 2013, IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 3048–3052 (May 2013)
6. Cox, I., Miller, M., Bloom, J., Fridrich, J., Kalker, T.: Digital Watermarking and Steganography. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2 edn. (2008)
7. Fridrich, J., Goljan, M., Hogea, D.: Steganalysis of JPEG images: Breaking the F5 algorithm. In: Proc. of IH 2003, Int. Conference on Information Hiding. pp. 310–323. Springer (2003)
8. Kirchner, M., Böhme, R.: Tamper hiding: Defeating image forensics. In: Proc. of IH 2007, Int. Conference on Information Hiding. pp. 326–341 (2007)
9. Lam, E.Y., Goodman, J.W.: A mathematical analysis of the DCT coefficient distributions for images. IEEE Transactions on Image Processing 9(10), 1661–1666 (2000)
10. Li, B., Shi, Y., Huang, J.: Detecting doubly compressed JPEG images by using mode based first digit features. In: Proc. of MMSP 2008, IEEE Workshop on Multimedia Signal Processing,. pp. 730–735 (Oct 2008)
11. Milani, S., Tagliasacchi, M., Tubaro, S.: Discriminating multiple JPEG compression using first digit features. In: Proc. of ICASSP 2012, IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 2253–2256 (March 2012)
12. Milani, S., Tagliasacchi, M., Tubaro, S.: Antiforensics attacks to benford's law for the detection of double compressed images. In: Proc. of ICASSP 2013, IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 3053–3057 (May 2013)
13. Pevný, T., Fridrich, J.: Estimation of primary quantization matrix for steganalysis of double-compressed JPEG images. In: in Proc. SPIE. vol. 6819, pp. 681911–681911–13 (2008), `http://dx.doi.org/10.1117/12.759155`
14. Popescu, A.C., Farid, H.: Statistical tools for digital forensics. In: Proc. of IH 2005, Int. Conference on Information Hiding. pp. 128–147. Springer (2005)
15. Rachev, S.T.: Mass Transportation Problems: Volume I: Theory, vol. 1. Springer (1998)
16. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing 13(4), 600–612 (2004)
17. Watson, A.B.: DCT quantization matrices visually optimized for individual images. In: Proc. of IS&T/SPIE's Symposium on Electronic Imaging: Science and Technology. pp. 202–216. International Society for Optics and Photonics (1993)