Adversary-aware, data-driven detection of double JPEG compression: how to make counter-forensics harder

Mauro Barni[1], Zhipeng Chen [2,3], Benedetta Tondi[1]

[1]*Department of Information Engineering and Mathematics, University of Siena, Siena, Italy*
[2] *Institute of Information Science, Beijing Key Laboratory of Advanced Information Science and Network Technology,*
Beijing Jiaotong University, Beijing, China
[3] *Department of Computer Science, Tangshan Normal University, Tangshan, China*
barni@dii.unisi.it, chzhpeng@hotmail.com, benedettatondi@gmail.com

Abstract—**In the attempt to investigate the final race of arms between the forensic analyst and the adversary in practical scenarios based on data-driven approaches, we introduce the idea of adversary-aware SVM-based forensic detection. By focusing on the problem of double JPEG compression, we first propose an improved universal counter-forensic (C-F) attack which works against any forensic detector based on the first order statistics of block-DCT coefficients and show its good performance against three different forensic detectors. Forensic detectors are commonly designed to distinguish between the absence and the presence of a given processing in a non-adversarial environment. We emphasize how such an evaluation methodology is unfair as, in order to test the real effectiveness of an attack, the forensic detector should take into account the possible presence of the attack. Accordingly, we propose an adversary-aware double JPEG detector which is trained to recognize the universal C-F attack. Experimental results confirm that the adversary-aware detector yields good performance thus suggesting that developing an effective counter-forensic attack is much harder than one could expect.**

I. Introduction

Counter-forensics has become a hot research topic in multimedia forensics due to the importance of understanding the limit of forensic analysis when the analyst must face the presence of an adversary with the explicit goal of impeding the analysis [1]. Early works were rather simplistic, since deleting the traces the forensic analysis relies on is usually a simple task: basic processing like noise addition, re-compression, resampling, median filtering is often enough to prevent a correct analysis. Research has then been focused on the development of anti-counter-forensic techniques that either provide good performance even in the presence of attacks, or at least detect that a certain attack or class of attacks were applied. Quite naturally, this activity has triggered the search for even more powerful counter-forensic strategies. In order to prevent research from entering an endless cat&mouse loop, some attempts have been made to develop attacks that are optimal against an entire class of forensic techniques, like in [2], where a universal attack which is effective against any image forensic technique based on histogram analysis is presented. Researchers have also started working on a general theory of adversarial signal processing [3], wherein the interplay between the attacker and the forensic analyst is cast into a game-theoretic framework to understand the ultimate optimum strategies for the two contenders. Such an approach has already been applied to a number of forensic problems, like source identification [4], video forensics of frame deletion [5], camera identification [6], ENF analysis [7]. Despite the above advances, the final outcome of the race of arms between the forensic analyst and the attacker is still hard to figure out. This is partly due to the difficulty of applying the theoretical findings to practical scenarios, wherein the precise statistical models adopted by theory do not hold, thus opening the way to the adoption of data-driven approaches based on modern machine learning methods [8]. Additionally, computational complexity often prevents the application of the optimum strategies devised by theory.

In this paper, by focusing the attention on the detection of double JPEG compression, we take some steps to fill this gap. To start with, we introduce an improved version of the universal attack proposed in [9]. By adopting an *optimal transportation* approach [10] similar to the one described in [2] to counter histogram-based image forensics, the new attack is able to counterfeit *any* detector based on the analysis of the first order statistics of block-DCT coefficients. The expectedly good performance of the new attack are confirmed experimentally against the double compression detector in [11], and two data-driven double JPEG detectors: a state-of-the-art detector based on the analysis of the DCT First Significant Digits (FSD) [12] and a new detector built by feeding a Support Vector Machine (SVM) classifier directly with the histograms of the block-DCT coefficients, which is similar to that proposed in [13] for steganography applications. As a further contribution, we adopt a data-driven approach to build a detector which is *trained so to take into account the possible presence of the attacker*. Even if the detector is still based on the first order statistics of block-DCT coefficients, its performance under attack is surprisingly good. In order to understand the reason for such a failure of the attacker, we resort to some recent theoretical results, which link the success of the attacker to the *maximum admissible distortion* he can introduce during the attack [14]. As we will see, the distortion that the attacker should introduce to implement a successful universal attack easily leads to unacceptable image degradation, thus showing that attacker's life is harder than one could imagine at first sight. The above conclusion is reached by assuming that the analyst knows all the details about the implementation of the attack, including the reference dataset used by the attacker to identify a histogram vector to be used as the target in the optimal transport attack [9]. As a last contribution, then, we analyse the performance of the adversary-aware detector when such an assumption does not hold, and the analyst must train the detector by relying on a local reference dataset. As we will see, in this case, the performance of the detector drop, hence making the battle between the analyst and the attacker more uncertain.

The rest of the paper is organised as follows: in Section II we introduce our improved version of the universal C-F algorithm in [9] by testing its good performance against three forensic detectors. Then, in Section III we present our adversary-aware double JPEG detector and show experimentally that it achieves good classification performance. The link with the theoretical results on source distinguishability is investigated in Section IV. Finally, in Section V the performance is evaluated in the case in which the analyst does not have perfect knowledge of the parameters of the attack.

978-1-5090-1138-4/16/$31.00 2016 IEEE

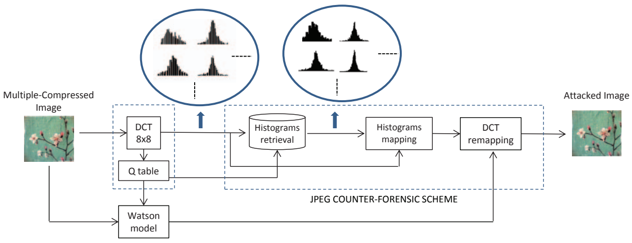

Fig. 1. Block scheme of the JPEG counter-forensic algorithm in [9].

II. Universal counter-forensics of double-JPEG compression

A universal post-processing C-F attack which works against *any* detector based on the analysis of the first order statistics of the image, i.e., the image histogram, was proposed in [2]. The attack is derived by leveraging on the theoretical results in [4], that are adapted to the case in which the statistical model is unknown (as it is in practical applications) and then the analysis is based on a data driven approach.

The scheme is later extended in [9] and applied to the transformed domain for concealing traces of multiple JPEG compression against detectors based on first order statistics of block-DCT coefficients. Specifically, given an image $\mathbf{y}$, which has been compressed two (or more) times with different quality factors, in order to pass off the image as a single compressed one, the attacker runs the universal JPEG C-F algorithm, illustrated in Figure 1. The three basic steps of the universal JPEG C-F algorithm, described below, are: the *retrieval phase*, in which the attacker searches for the best vector of DCT histograms (according to some similarity measure) in a database of single-compressed images; the *mapping phase*, in which the optimum transportation map between forged and target vector of histograms is computed, subject to a constraint on the maximum distortion that can be introduced; and, finally, the phase of *implementation of the mapping*, when the map is implemented into the processed (multiple compressed) image by minimising the perceptual distortion introduced in the pixel domain. A strength of the method in [9] with respect to state-of-the-art approaches, e.g. the ones in [15], [16], is that the final perceptual impact of the attack in the pixel domain is evaluated and the amount of modification introduced by the scheme is constrained so to guarantee imperceptibility of the attack.

It is worth stressing that, although the arguments of this paper are focused on double JPEG compression, the C-F scheme in [9] works for general multiple JPEG compression.

A. Improved double-JPEG counter-forensic scheme

Throughout the paper, capital letters X and X_q are used respectively to denote the transformed image and its quantized version. Specifically, $X_q(i,j)$ indicates the transformed coefficient in subband (i,j) of a generic block; to refer to the coefficient of a specific block h we use the notation $X_q(i,j;h)$. Notation $h_{X(i,j)}$ is used for the histogram of the DCT coefficients at frequency (i,j). Similarly, $\mathbf{h}_X = (h_{X(i,j)})_{i,j=1}^{8}$ denotes the vector of the DCT histograms. We let $n_{ij}(m,r)$ indicate the number of elements in $h_{Y_{ij}}$ which are moved from the m-th to the r-th bin. We call $N_{ij} = \{n_{ij}(m,r)\}_{m,r=1}^{|\mathcal{C}|}$ *transportation map*, where $|\mathcal{C}|$ is the cardinality of the alphabet of the DCT coefficients. Then we have $h_{Y(i,j)}(m) = \sum_r n_{ij}(m,r)$, $\forall m$. Similarly, $\sum_m n_{ij}(m,r) = h_{Z(i,j)}(r)$, $\forall r$, where $h_{Z(i,j)}(r)$ denotes the histogram we get after the application of the map. Finally, given two images Y and Z, the *maximum* (or *infinite*) *distance* between the transformed coefficient in subband (i,j) can be expressed in terms of transportation map as:

$$\max_h |Z(i,j;h) - Y(i,j;h)| = \max_{(m,r):n_{ij}(m,r)\neq 0} |m-r|. \qquad (1)$$

Below, we describe in more detail the three stages of the attack by outlining the modifications with respect to the scheme originally proposed in [9].

In the retrieval phase, the attacker selects the target vector $\mathbf{h}_X$ from a reference database DB of single compressed images. For any frequency subband (i,j), the similarity between $h_{Y(i,j)}$ and $h_{X(i,j)}$ is measured through a *cross-bin distance*, which is related to the Earth Movers Distance (*EMD*) [17][1]. More specifically, to characterize the similarity, we consider the minimum infinite distortion that we need to introduce in the DCT coefficients to move one histogram into the other. Formally, given two pmf's or, more in general, two mass distributions, as they are $h_{Y(i,j)}$ and $h_{X(i,j)}$, and a cost for unitary mass $d(i,j)$, the *EMD* is defined as the minimum transportation cost to turn one mass into the other, obtained by solving the transportation problem[2]

$$\min_{N_{ij}:\sum_r n_{ij}(m,r)=h_{Y(i,j)},\sum_m n_{ij}(m,r)=h_{X(i,j)}} \sum_{(m,r)} n_{ij}(m,r)d(i,j). \qquad (2)$$

It is known that, when $d(i,j)$ is a convex function of $|i-j|$, the minimization in (2) can be solved through a *greedy algorithm* known as *North West Corner (NWC) rule* [18]. As a result of the analysis in [14], given $h_{Y(i,j)}$ and $h_{X(i,j)}$, the transportation map N_{ij} which *minimizes* the maximum distance between images $Y(i,j)$ and $X(i,j)$ (see equation (1)) is the map obtained by applying the NWC rule to $h_{Y(i,j)}$ and $h_{X(i,j)}$, named N_{ij}^{NWC}.

Then, we measure the similarity between the pair of attacked and candidate target DCT histogram at frequency (i,j) by considering the maximum distance which results from the application of the NWC, i.e., the quantity

$$S(h_{Y(i,j)}, h_{X(i,j)}) = \max_{(m,r):n_{ij}^{NWC}(m,r)\neq 0} |m-r|. \qquad (3)$$

The choice is motivated by the fact that the maximum distance is the same measure which characterizes the distortion constraint for the attack in the mapping phase. Besides, this measure has also some ties with the theoretical concept of Security Margin (see discussion in Section IV).

Distinction is made in [9] between two possible kinds of search: *joint* and *disjoint*. In the joint search case, the most similar vector of DCT histograms $\mathbf{h}_X^*$ is selected among the vectors of the DB; i.e., the one which minimizes the total distance $\sum_{(i,j)} S(h_{Y(i,j)}, h_{X(i,j)})$. In this case, the retrieved vector of DCT histograms belongs to an image stored in the DB. In the disjoint search mode, differently, the retrieval is done separately for each DCT subband. Then, for each (i,j), the attacker searches the DB for the histogram $h_{X(i,j)}^*$ minimizing $S(h_{Y(i,j)}, h_{X(i,j)})$. Hence, in such case, the recovered vector does not belong to an image of the DB. Although, arguably, the *joint* search approach is preferable with respect to the *disjoint* one because it is forensically more secure, it needs significantly large DB for getting good performances. To partially overcome the limitations, as further contribution, we propose to use an *hybrid* approach according to which a separate search is

[1] The use of a cross-bin distance overcomes the limitation of the metrics adopted so far, like the chi-square $\mathcal{X}$ and the divergence $\mathcal{D}$, which do not take into account relationships between adjacent bins.

[2] Strictly speaking, the formulation holds when the histograms have equal mass; however, the extension to the case in which such assumption does not hold is immediate.

done for the DC coefficient (which suffers the most the joint search in limited-sized DB because of its larger variance), whereas the joint search approach is kept for the AC coefficients (*partially joint* approach).

Regarding the mapping phase, from the theoretical results in [19], by exploiting the low intra-block dependence among DCT coefficients, the attacker's strategy consists in finding the histograms $h_{Z(i,j)}$, $1 \leq i,j \leq 8$, which minimizes the quantity $\sum_{(i,j)} \mathcal{D}(h_{Z(i,j)}||h_{X(i,j)})$, subject to a distortion constraint imposed to limit the distortion introduced in the pixel domain in order to maintain the final image visually similar to the initial one. In order to characterize this constraint in the frequency domain, we consider the Just Noticeable Distortion, i.e., the maximum modification for each DCT coefficient which is visually undetectable. Reasonably, this provides a so-to-say maximum value for the distortion that the attacker can introduce in the coefficients of the transformed image. The model used for the JND is Watson's model [20], which provides a 8×8 sensitivity matrix $W = \{W(i,j)\}_{i,j=1}^{8}$. We denote with $W_q = \{\text{round}\,(W(i,j)/q_Y(i,j))\}_{i,j=1}^{8}$ the quantized Watson's matrix (q_Y denotes the quantization table for image Y). Then, entry $W_q(i,j)$ provides the maximum amount of distortion which can be introduced in the quantized DCT coefficients of the subband (i,j) without generating annoying artifacts. The maximum distortion for the (i,j) coefficient is given by $W_q(i,j) \cdot D_{max}$ for some $D_{max} \geq 1$ (larger D_{max} allows to obtain more accurate mapping at the price of a higher visual distortion). Since the distortion constraint is defined subband-wise, the problem can be solved as 64 separate minimizations:

$$\min_{Z(i,j):\max_h |Z(i,j;h)-Y(i,j;h)| \leq \lceil W_q(i,j) \cdot D_{max} \rceil} \mathcal{D}(h_{Z(i,j)}, h_{X(i,j)}), \tag{4}$$

where D is the K-L divergence. Looking at the expression in (4), it is easy to argue that the attacker will exploit all the available distortion to bring each $h_{Y(i,j)}$ as close as possible to the corresponding target histogram $h_{X(i,j)}$; hence, in some sense, the 'best' target histogram is the one that minimizes such final maximum distance, i.e. the histogram to which we are able to get close as much as possible after the mapping, thus motivating the choice of S (see equation (3)) in the retrieval phase.

In the final phase of the algorithm, the mapping is mapping in a perceptually convenient way. The DCT coefficients in different blocks of the images are modified according to a refined sensitivity model, that takes into account luminance and contrast masking effects to get an accurate evaluation of the modifications that can be performed (in un imperceptible manner) on each block. With respect to the remapping algorithm in [9], in order to improve the imperceptibility of the modifications, the mean intensity of the local area is used here in place of the mean intensity of the image for calculating the blockwise threshold of JND in the refined Watson's model. An example of image before and after the application of the refined universal C-F scheme is provided in Figure 2: the images are visually identical.

B. Effectiveness of the attack

The performance of the refined universal attack is tested first against a simple double compression detector based on the so-called calibration technique [11] and then against two data-driven detectors: the first one is based on the analysis of the FSD features [12], while the second one is our proposed SVM-based detector fed with all the histograms of the block-DCT coefficients, inspired by the detector in [13]. The idea behind the new detection is simple: rather than considering specific features derived from the first order statistics of DCT coefficients, we can directly feed the SVM with a feature vector formed by the histograms of block-DCT coefficients. Being the C-F attack scheme universal (within the class of forensic detectors based on first-order statistics), it is expected to work even against such a detector. To build the vector, before concatenating the DCT histograms, each of them is arranged on a reference support which is determined so to be large enough to accommodate the histogram content, whatever the quality factor of the JPEG image is. For the DC histogram, we consider the range determined by the JPEG 100% (4096 bins), whereas, to save the length of the feature vector, a worst-case range extent for the histograms of the AC coefficients is determined experimentally.

(a) double-compressed

(b) attacked

Fig. 2. Comparison between a double compressed image, with quality factors $(70, 85)$, (a) and its attacked version (b).

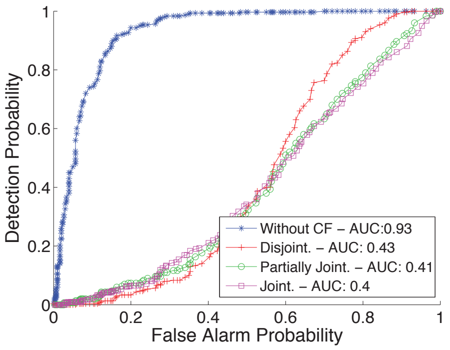

Fig. 3. Performance of the C-F method against the detector in [11].

To assess the validity of the proposed modifications, for our experiments, we used uncompressed (camera-native) grayscale images from the RAISE dataset [21]. Specifically: 300 images are selected to build the test set $\mathcal{S}_t$; besides, a set of 2000 images is used as database for the attacker (DB). In our experiments, we considered the following pairs of quality factors for the first and second compression: $(QF_1, QF_2) \in \{(65; 85); (70; 85); (70, 90); (75; 90)\}$. Then, for any QF_2, the images in set $\mathcal{S}_t$ are compressed once with QF_2 to build the set of single compressed $\mathcal{S}_{t,s}$, and twice with (QF_1, QF_2) for the various QF_1, to build the double compressed set $\mathcal{S}_{t,d}$. For the tests with the two data-driven detectors, we additionally select 700 uncompressed images to build the training set $\mathcal{S}_T$. Similarly, from this set, we build the set $\mathcal{S}_{T,s}$ and $\mathcal{S}_{T,d}$ of single and double compressed images which are used to train a SVM with Gaussian kernel. In both cases, we considered 15 DCT coefficients taken in zig-zag order to build the feature vector. Then, for the images in $\mathcal{S}_{t,d}$, we run the C-F scheme with $D_{max} = 4$, to build the set of the attacked images $\mathcal{S}_{t,a}$.

Figure 3 shows the good performance of the improved universal C-F method against the detector in [11]. The visual performance of the attack in the case of joint, disjoint and partially joint search is provided in Table I in terms of Structural Similarity Index (SSIM)

	Mean SSIM	Std. dev. SSIM	Mean PSNR
Disjoint	0.92	0.052	34.5 dB
Joint	0.87	0.056	30.3 dB
Partially Joint	0.91	0.056	32.5 dB

TABLE I
PERFORMANCE OF THE C-F ATTACK IN TERM OF PERCEPTUAL QUALITY.

[22] and Peak-Signal-to-Noise ratio (PSNR). Experiments show that the method is also able to fool the detector in [12], although not tailored for this purpose: in fact, whereas the single and double compressed images are correctly classified (100% true positives and 100% true negatives), more than 98% attacked images are deemed as single-compressed. This is expected due to the universality of the method (when first-order statistics are considered). Even the proposed, more general detector based on the block-DCT histograms is fooled by the universal CF attack, as 99% of the attacked images are classified as single compressed.

We also verified that, quite expectedly, the new SVM-based detector makes ineffective the attacks in [16], [23] focused on FSD features (as they are features extracted from the histograms of the DCT coefficients), by labeling 100% of attacked images as double compressed. The concealment of the traces in the FSD domain, in fact, leaves traces back in the DCT histograms that our SVM-based classifier is able to recognize.

III. ATTACKER-AWARE DOUBLE-JPEG DETECTION

We have shown that the universal C-F method is very effective, being able to fool several double JPEG forensic detectors. However, in hindsight, at the point we are now, the struggle between the analyst and the adversary is not fair. In fact, while the adversary is aware of the forensic analysis and plays its best strategy (according to the theory) in order to mislead the detection, on the other side, we do not consider the case in which the forensic analyst, aware of the presence of the adversary, takes countermeasures. This is a common problem in Forensics, where the effectiveness of counter-forensic techniques is evaluated against detectors designed for distinguishing between the absence or presence of the processing in a so called 'licit' scenario, i.e. by assuming that no attempt has been made to conceal it.

It is easy to understand that, in order to properly evaluate the powerfulness of an attack, *the performance must be checked against an attacker-aware detector*. The issue of proper performance evaluations of security-oriented systems is also pointed in [24], together with an extensive discussion on the evaluation methodologies; in that paper, the attention is drawn to the fact that, biometric systems should be designed to work properly both in the normal operating scenario with no adversary and in the presence of spoofing attacks. A forensic detector should then be designed so to be able to tell apart single compressed images from both double and attacked images. About the possible decision strategies, it makes sense to consider both a 3-class classifier (which tells apart original, processed and attacked images) and a 'pseudo-ternary classification', where the *positive* class corresponds to the original/untouched images whereas processed and attacked images constitutes the *extended negative* class [24]. Note that resorting to a pseudo-ternary classification makes sense, as the final goal of the analyst is to reliably tell apart original images from the other ones (attacked and processed images). On the other hand, an attacker who is not able to make the attacked image look like an original one is failing, no matter if the attacked images are classified as processed or manipulated. Clearly, when the ternary classification works, the 'pseudo-ternary classification' works as well.

A. Adversary-aware SVM-based double-JPEG detector

We refine our data-driven double JPEG forensic detector based on block-DCT histograms by taking into account the possible presence of an attacker performing the universal C-F algorithm. We do so by training the SVM with samples of attacked images, so to make the detector able to 'recognize' the attack. For the case of ternary classification, the feature vectors made of the histograms of the DCT coefficients are computed for all the images in $\mathcal{S}_{T,s}$, $\mathcal{S}_{T,d}$ and $\mathcal{S}_{T,a}$ (built from the set $\mathcal{S}_{T,d}$ by running for each image the C-F attack with the chosen D_{max}) and used to feed a 3-class SVM. When instead a binary or pseudo-ternary classification is chosen, the images in $\mathcal{S}_{T,d}$ and $\mathcal{S}_{T,a}$ are selected in some percentage to build the extended negative class, which is used, together with the positive class $S_{T,s}$, to train the 2-class SVM.

TABLE II
RESULTS FOR $D_{max} = 4$ (TOP ROW) AND $D_{max} = 8$ (BOTTOM ROW).

	S	D	A
S	86.67	0.33	13
D	0	100	0
A	25	0	75

(a) Joint

	S	D	A
S	92	0.33	7.67
D	0	100	0
A	6	0	94

(b) Partially joint

	S	D	A
S	99.67	0.33	0
D	0	99.67	0.33
A	0.33	0	99.67

(c) Disjoint

	S	D	A
S	80	0.33	19.67
D	0	100	0
A	29.67	0	70.33

(d) Joint

	S	D	A
S	88.67	0.33	11
D	0	100	0
A	12	0	88

(e) Partially joint

	S	D	A
S	99.67	0.33	0
D	0	99.67	0.33
A	0.33	0	99.67

(f) Disjoint

It is worth observing that, from the analyst's perspective, the attacker-aware detector should be trained to recognize all the possible attacks, as, in principle, it may not be able to classify attacks he is not made aware of. However, due to the universality of the C-F method, it is reasonable to expect that the detector trained to recognize this attack, will keep working even against other manipulations of the first order statistics.

B. Experimental results

For the experiments here and in the following sections, we focus on the case in which $(QF_1, QF_2) = (70, 85)$ The 700 images in $\mathcal{S}_T$ are first compressed once with $QF_2 = 85$ to form $\mathcal{S}_{T,s}$, compressed twice with $(QF_1, QF_2) = (70, 85)$ ($\mathcal{S}_{T,d}$) and attacked with D_{max} ($\mathcal{S}_{T,a}$); these sets are used for training the 3-class SVM. The SVM is then tested with the single, double compressed and attacked versions of the 300 images of the test set $\mathcal{S}_t$.

Table II(a) through (f) show the performance of the 3-class classifier in the case of $D_{max} = 4$ and 8. The experiments were performed by using the joint, partially joint and disjoint search. The results show that, now that the forensic analyst is aware that counter-forensic measures are taken and react accordingly, the attack is not as powerful as before. Besides, they confirm that the joint search approach is forensically safer. Quite expectedly, increasing D_{max} from 4 to 8 improves the performance of the attack; however, the visual degradation of the attacked images is significant (resulting in an average SSIM of 0.82 and a PSNR of 29 dB) and annoying artifacts become clearly visible.

Experiments were conducted even for the case of pseudo-ternary classification. We denote with α the percentage of attacked images which contributes to the negative class, $0 \leq \alpha \leq 1$. For the training, the SVM is fed with the 700 images from the same set $S_{T,1}$ (for the positive class) and 700 images, randomly chosen in $\mathcal{S}_{T,d}$ and $\mathcal{S}_{T,a}$, with the fixed percentage α (for the negative class). For the test, the same sets $\mathcal{S}_{t,s}$, $\mathcal{S}_{t,d}$ and $\mathcal{S}_{t,a}$ are used. The classification performance is shown in Figure 4 for the various search modes and different values of α. In all cases, the classification performance improve by increasing α. Hence, we learn the following interesting and general lesson: when dealing with the training of binary classification with extended negative classes,we should take into more consideration the

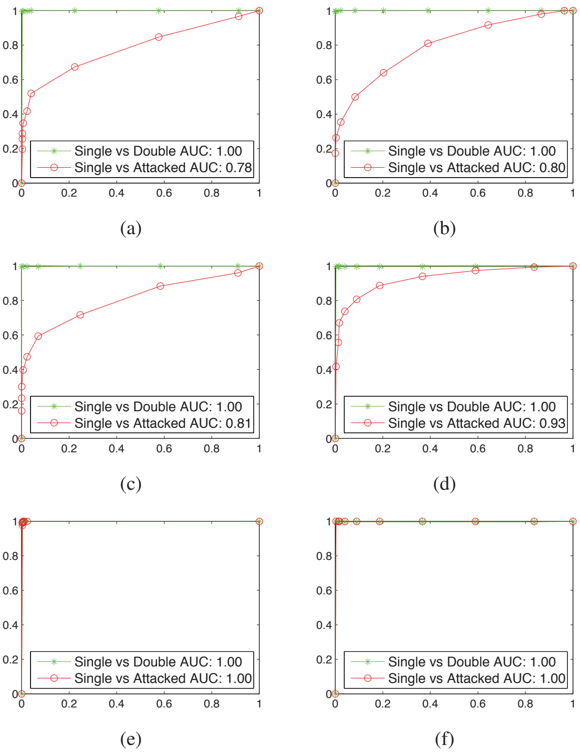

Fig. 4. Detection performance of the aware 2-class detector against the universal C-F attack with $D_{max} = 4$. Joint search case: $\alpha = 0.5$ (a) and $\alpha = 1$ (b); partially joint search case: $\alpha = 0.5$ (c) and $\alpha = 1$ (d); disjoint search with $\alpha = 1$ (e). Performance of the aware detector against the attack in [16] for $\alpha = 1$ (f) (similar results hold for any value of α in $[0, 1]$).

subclass which deviates less from the positive class; in adversarial applications, this corresponds to the class of the signals intentionally modified to be passed off as positive. We also see that the new 2-class detector keeps working well with respect to the counter-forensic attack in [16] (see Figure 4(f)). This is not surprising since the aware detector simply refines the classification of the single compressed images, based on the attacked samples, with respect to the unaware detector, thus actually enlarging the negative class.

IV. Security-margin perspective

In this section, we provide a qualitative motivation of the results presented in the previous sections by the light of theoretical findings in [14], where the distinguishability of two sources under adversarial condition is summarized by a single quantity named Security Margin ($\mathcal{SM}$). When the analyst-attacker interplay is casted in a game theoretic framework, by exploiting the parallelism with Optimal Transport Theory [10], the concept of Security Margin is defined as the maximum distortion introduced by the attacker for which the two sources can be distinguished by the defender. Then, in our practical experiments, looking at the values of the maximum distance we need to move the attacked histogram $\mathbf{h}_Y$ into the retrieved histogram $\mathbf{h}^*_X$, i.e. at the values $S(h_{Y(i,j)}, h^*_{X(i,j)})$, $\forall (i,j)$, may help to understand how difficult is for the adversary to completely delete the traces left by the double JPEG compression, by making the attacked histogram identical to an uncompressed one. By introducing a large distortion D_{max} (i.e., such that $\lceil W_q(i,j) \cdot D_{max} \rceil > S(h_{Y(i,j)}, h_{X(i,j)})$, $\forall i,j$), in fact, the attacker is sure to make the forensic analysis fail ($0.3 \leq W_q(i,j) \leq 1.5$ for the first 15 DCT coefficients). However, the values of S we experimentally get with our dataset DB are rather large, thus possibly compromising the visual quality of the forgery. Clearly, by resorting to a larger dataset, we could be able to slightly decrease this value.

TABLE III
Classification results for $D_{max} = D_s$.

	S	D	A
S	57.67	0	42.33
D	0	100	0
A	37	0	63

(a) Joint

	S	D	A
S	68.33	0.33	31.33
D	0	100	0
A	31.33	0	68.67

(b) Partially joint

	S	D	A
S	99.67	0.33	0
D	0	99.67	0.33
A	0.33	0	99.67

(c) Disjoint

In Table III, we report the performance of the aware detector when the value of D_{max} is such that, for all the images Y in test and training set, the upper bound in the constraint in (4) is larger than $S(h_{Y(i,j)}, h^*_{X(i,j)})$, for any (i,j), and then, as a result of the mapping, each attacked histogram is exactly mapped into a single compressed one (we let D_s denote this value[3]). Although the price to pay in terms of visual distortion of the attacked image, is unacceptable, such kind of analysis allows to make some interesting considerations: first, as we guessed, the disjoint search approach for the attack is not forensically safe, resulting in discrepancies among the histograms of the DCT coefficients at the various frequencies that the adversary-aware detector is able to learn, no matter how large the allowed distortion is (that is, even when $D_{max} = D_s$). This behavior is theoretically supported: since by performing the disjoint search, the retrieved vector of 64 DCT histograms $\mathbf{h}^*_X$ does not belong to a single compressed image stored in the DB, it is not properly an instance of the single compressed class. This makes still possible a reliable distinction between the attacked and the single compressed class. As a second notice, contrarily to what one could expect, even when the forensically safer joint search approach is considered, the detector is not completely fooled by the attacker, and more than 60% of the attacked images are still correctly classified. The reason for this apparently strange behavior will be clear from the analysis in the next section. Finally, by adopting the opposite defender's perspective, it is worth pointing that the analysis of this section provides also a qualitative measure of the goodness of the performance of a detector: a detector which is able to distinguish well between two classes, should not be fooled if the adversary introduces a distortion less than the Security Margin (namely, the distortion which makes the attacked histogram identical to a single compressed one on the average). Experiments could also be done in this sense to validate the goodness of the proposed attacker-aware detector.

V. A further ingredient: database mismatch

So far, we considered the case in which the reference database used by attacker to forge the image and analyst to reproduce the attacked samples is the same. In this section, we consider the more realistic scenario in which the analyst does not know the exact reference database used by the attacker and thus the *local* reference database that the defender relies on to generate the attacked samples does not corresponds to the one used by the attacked to produce the forgeries (*mismatched* database case). To do so, we consider another set of 2000 images from the RAISE dataset to form the local reference database for the analyst (DB'). Then, the images in $\mathcal{S}_{T,d}$ are attacked with the universal C-F scheme by using DB' as reference database to build $\mathcal{S}_{T,a}$. The results of the tests are shown in Table IV. Quite expectedly, the database mismatch plays in favour of the attacker. However, by looking at the results for $D_{max} = 4$ we see that, the

[3] In our experiments $D_s \approx 40$; the exact value depends on the search mode.

TABLE IV

RESULTS FOR $D_{max} = 4$ JOINT (LEFT) AND PARTIALLY JOINT (CENTER) AND $D_{max} = D_s$, JOINT (RIGHT), WITH DATABASE MISMATCH.

	S	**D**	**A**		**S**	**D**	**A**		**S**	**D**	**A**
S	86.67	0.33	13	**S**	92	0.33	7.67	**S**	57.67	0	42.33
D	0	100	0	**D**	0	100	0	**D**	0	100	0
A	25.67	0	74.33	**A**	16.67	0	83.33	**A**	46	0	54

performance drop in the detection is not sufficient to tip the scale again in favour of the attacker. Then, performing successful counter-forensics is really much harder than one could imagine at first glance. Finally we observe that, with the mismatch of the databases, in the case of attack with $D_{max} = D_s$, as expected, the decision is similar to guessing by flipping a coin.

VI. CONCLUSIONS

In this paper, we first refine the universal C-F attack proposed in [9] for concealing traces of double JPEG compression and evaluate its effectiveness against three forensic detectors. Then, as a major contribution, we play the role of the analyst and refine the detection by proposing an adversary-aware, data driven, detector. The performance of the new detector is then tested under various settings against the universal C-F attack (which is the attack the detector is made aware of), as well as under another attack based on first-order statistics of the DCT coefficients. Evaluating the performance of the attack against an attacker-aware detector allows to understand its real effectiveness. However, the detection of double JPEG compression considered here is only a case-study and similar considerations can be done in other forensic applications: whenever an analyst wants to detect the presence of a certain processing and a counterattack may take place, both 'types of negatives' should play a role in performance evaluation. As a future work, we plan to extend our analysis by considering also more powerful machine learning methods. The extension to the case of higher-order statistics is another interesting direction.

VII. ACKNOWLEDGEMENT

This material is based on research sponsored by DARPA and Air Force Research Laboratory (AFRL) under agreement number FA8750-16-2-0173. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA and Air Force Research Laboratory (AFRL) or the U.S. Government. The paper is partially supported by the national key research and development program of China under grant 2016YFB0800404, NSFC under grant 61332012 and Beijing Nova Program Interdisciplinary Cooperation Project under grant Z1511000003150146. We thank Prof. Yao Zhao and Prof. Rongrong Ni for useful discussions and helpful comments on the earlier drafts.

REFERENCES

[1] R. Böhme and M. Kirchner, "Counter-forensics: Attacking image forensics," in *Digital Image Forensics*, H. T. Sencar and N. Memon, Eds. Springer Berlin / Heidelberg, 2012.

[2] M. Barni, M. Fontani, and B. Tondi, "A universal technique to hide traces of histogram-based image manipulations," in *Proc. of the ACM Multimedia and Security Workshop*, Coventry, UK, 6-7 September 2012, pp. 97–104.

[3] M. Barni and F. Pérez-González, "Coping with the enemy: advances in adversary-aware signal processing," in *ICASSP 2013, IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Vancouver, Canada, 26-31 May 2013, pp. 8682–8686.

[4] M. Barni and B. Tondi, "The source identification game: an information-theoretic perspective," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 3, pp. 450–463, March 2013.

[5] M. C. Stamm, W. S. Lin, and K. J. R. Liu, "Forensics vs anti-forensics: a decision and game theoretic framework," in *Proc. of ICASSP 2012, IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Kyoto, Japan, 25-30 March 2012.

[6] H. Zeng, X. Kang, and J. Huang, "Mixed-strategy Nash equilibrium in the camera source identification game," in *Proc. of 20th IEEE International Conference on Image Processing (ICIP)*, 2013, pp. 4472–4476.

[7] W. H. Chuang, R. Garg, and M. Wu, "Anti-forensics and countermeasures of electrical network frequency analysis," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 12, pp. 2073–2088, Dec 2013.

[8] F. Marra, G. Poggi, F. Roli, C. Sansone, and L. Verdoliva, "Counter-forensics in machine learning based forgery detection," in *SPIE/IS&T Electronic Imaging*. International Society for Optics and Photonics, 2015, pp. 94 090L–94 090L.

[9] M. Barni, M. Fontani, and B. Tondi, "Universal counterforensics of multiple compressed jpeg images," in *Digital-Forensics and Watermarking*. Springer, 2014, pp. 31–46.

[10] C. Villani, *Optimal Transport: Old and New*. Berlin: Springer-Verlag, 2009.

[11] J. Fridrich, M. Goljan, and D. Hogea, "Steganalysis of jpeg images: Breaking the f5 algorithm," in *Information Hiding*. Springer, 2003, pp. 310–323.

[12] S. Milani, M. Tagliasacchi, and S. Tubaro, "Discriminating multiple jpeg compression using first digit features," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 2253–2256.

[13] T. Pevny and J. Fridrich, "Detection of double-compression in jpeg images for applications in steganography," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 2, pp. 247–258, June 2008.

[14] M. Barni and B. Tondi, "Source distinguishability under distortion-limited attack: An optimal transport perspective," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 10, pp. 2145–2159, Oct 2016.

[15] P. Comesana-Alfaro and F. Pérez-González, "Optimal counterforensics for histogram-based forensics," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 3048–3052.

[16] C. Pasquini, P. Comesaa-Alfaro, F. Prez-Gonzlez, and G. Boato, "Transportation-theoretic image counterforensics to first significant digit histogram forensics," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 2699–2703.

[17] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *Int. J. Comput. Vision*, vol. 40, no. 2, pp. 99–121, November 2000.

[18] A. Hoffman, "On simple linear programming problems," in *Proceedings of Symposia in Pure Mathematics*, vol. 7. World Scientific, 1963, pp. 317–327.

[19] M. Barni and B. Tondi, "Multiple-observation hypothesis testing under adversarial conditions," in *Information Forensics and Security (WIFS), 2013 IEEE International Workshop on*, Nov 2013, pp. 91–96.

[20] A. B. Watson, "DCT quantization matrices visually optimized for individual images," in *Proc. of IS&T/SPIE's Symposium on Electronic Imaging: Science and Technology*. International Society for Optics and Photonics, 1993, pp. 202–216.

[21] D.-T. Dang-Nguyen, C. Pasquini, V. Conotter, and G. Boato, "Raise: A raw images dataset for digital image forensics," in *Proceedings of the 6th ACM Multimedia Systems Conference*, ser. MMSys '15. New York, NY, USA: ACM, 2015, pp. 219–224. [Online]. Available: http://doi.acm.org/10.1145/2713168.2713194

[22] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[23] S. Milani, M. Tagliasacchi, and S. Tubaro, "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*, May 2013, pp. 3053–3057.

[24] I. Chingovska, A. Anjos, and S. Marcel, "Anti-spoofing in action: Joint operation with a verification system," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, June 2013, pp. 98–104.