Image counter-forensics based on feature injection

Image counter-forensics based on feature injection

M. Iuliani[a], S. Rossetto[b], T. Bianchi[c], A. De Rosa[a], A. Piva[b] and M. Barni[d]

[a]National Inter-University Consortium for Telecommunications, Florence, Italy;
[b]Dept. of Information Engineering, University of Florence, Italy;
[c]Dept. of Electronic and Telecommunications, Politecnico di Torino, Italy;
[d]Dept. of Information Engineering and Mathematical Sciences, University of Siena, Italy

ABSTRACT

Starting from the concept that many image forensic tools are based on the detection of some features revealing a particular aspect of the history of an image, in this work we model the counter-forensic attack as the injection of a specific fake feature pointing to the same history of an authentic reference image. We propose a general attack strategy that does not rely on a specific detector structure. Given a source image x and a target image y, the adversary processes x in the pixel domain producing an attacked image $\tilde{x}$, perceptually similar to x, whose feature $f(\tilde{x})$ is as close as possible to $f(y)$ computed on y. Our proposed counter-forensic attack consists in the constrained minimization of the feature distance $\Phi(z) =\mid f(z)-f(y) \mid$ through iterative methods based on gradient descent. To solve the intrinsic limit due to the numerical estimation of the gradient on large images, we propose the application of a feature decomposition process, that allows the problem to be reduced into many subproblems on the blocks the image is partitioned into. The proposed strategy has been tested by attacking three different features and its performance has been compared to state-of-the-art counter-forensic methods.

Keywords: image forensics, counter-forensics, feature injection, feature decomposition, constrained minimization, gradient descent, numerical computation, bisection method.

1. INTRODUCTION

Research in multimedia forensics stems from the idea that through the analysis of intrinsic properties of digital contents it is possible to understand their life cycle: in this framework, many tools have been developed and can be used by the forensic analyst.[1] In most cases such tools assume the form of detectors of specific features revealing some information on a particular aspect of the history of the digital content.

Recently, researchers have begun to consider the opposite point of view, that of an adversary aiming at processing the digital content in order to introduce fake features pointing to a fake history.[2] Most of the counter-forensic schemes proposed so far propose to insert/remove a specific feature detectable by means of one particular forensic detector; since it is usually assumed that the detector is known, the design of the attack can exploit the knowledge of its weaknesses.

To hide fingerprints left by image resampling due to geometrical operations like resizing or rotation, Kirchner and Böhme[3] proposed a set of attacks: since the main idea to detect resampling is to look for the presence of periodic linear dependencies between pixels in a close neighborhood, non linear filtering, or small geometrical distortions are applied to distort such condition; this allows to disguise resampling detection schemes like the one proposed by Popescu and Farid.[4] Other anti-forensic operations have been designed to remove or to falsify the photo-response non-uniformity (PRNU) fingerprint left in digital images by sensor imperfections. Gloe *et. al*[5] proposed a removal attack based on the application of flat fielding; next, a fingerprint-copy attack is proposed: a specific camera fingerprint is estimated from a set of acquired images and pasted onto an image from a different camera (where the removal attack has already been carried out) with the

Further author information: (Send correspondence to Alessia De Rosa)
Alessia De Rosa: E-mail: alessia.derosa@unifi.it
Massimo Iuliani: E-mail: massimo.iuliani@gmail.com
Simone Rossetto: E-mail: simros85@gmail.com
Tiziano Bianchi: E-mail: tiziano.bianchi@polito.it
Alessandro Piva: E-mail: alessandro.piva@unifi.it
Mauro Barni: E-mail: barni@dii.unisi.it

aim to introduce a false source camera fingerprint. A countermeasure against such an attack, named Triangle Test, has been introduced by Goljan *et. al*;[6] however, a more sophisticated behavior of the attacker is studied by Caldelli *et. al*[7] allowing to invalidate such new countermeasures. A method to synthetically create or restore a color filter array (CFA) fingerprint in digital images is proposed by Kirchner and Böhme.[8] This attack can be useful to conceal traces of manipulation that disrupted the CFA pattern. A lot of work has been concentrated on the study of methods allowing to hide traces left by a compression operation. Stamm *et. al*[9] proposed a method for removing the quantization artifacts left on DCT coefficients in JPEG-compressed images. The main idea is to modify the comb-shaped distribution of DCT coefficients in JPEG-compressed images, in such a way to restore a Laplacian distribution, which typically arises in uncompressed natural images, by adding a dithering noise signal in the DCT domain. Stamm and Liu[10] extended the above approach to hide quantization footprints left by a wavelet-based coding scheme, like JPEG2000, so as to fool the scheme by Lin *et. al*.[11] However, Valenzise *et. al*[12] demonstrated that this attack induces a loss of perceived image quality, with respect to both the original (uncompressed) and to the JPEG-compressed image. The authors propose then a perceptually modified version of the attack, taking into account the level of just-noticeable distortion (JND) that can be sustained by each DCT coefficient. The same authors[13] 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. Other detectors of the dithering attack on DCT coefficients are proposed by Lai and Böhme,[14] analyzing the magnitude and the number of zeros in high frequency AC coefficients. Stamm *et. al*[15] proposed also a deblocking method to remove blocking artifacts caused by JPEG compression, to disguise the forensic detector proposed by Fan and de Queiroz;[16] the attack consists in smoothing the JPEG-compressed image with a median filter, and then adding a low-power white noise signal to the filtered image.

Recently, counter-forensic methods have been proposed that try to mimic the (statistical) properties of a certain class of contents (e.g. unaltered contents), in order to fool - at least in principle - all forensic algorithms that are based on the analysis of such properties. As an example of this class of methods, Barni *et. al*[17] proposed a counter-forensic technique for hiding traces left on the image histogram by any processing operation, by assuming that the forensic scheme to be deceived is based on first-order statistics only.

Inspired by this latter approach, we model the counter-forensic attack as the injection into a digital image of a specific feature that points to a fake history the adversary wants to be believed, specifically the same history of a given reference image. In this way, when the reference image and the fake-and-attacked image are analysed, the detector will give the same answer on both images. We propose a general attack strategy that does not rely on a particular detector structure: the adversary considers the detector as a black box computing the feature he wants to attack. In this sense, the proposed strategy can be defined as a blind attack; furthermore, our strategy is a very general one, since it works regardless of the specific feature to be attacked and can be applied to a wide class of detectors (assuming that some general properties are satisfied, as specified in the following sections).

Our work is based on the following idea: let us consider a scalar real-valued feature f, and a certain detector that classifies an image by relying on the values assumed by f. The adversary wants that a source image x, when properly modified, exhibits a value of f that points to a given fake history. To do so, the adversary considers a given target image y telling the same history x should tell, and processes the image x in the pixel domain producing an attacked image $\tilde{x}$, so that $\tilde{x}$ is perceptually similar to x and the feature $f(\tilde{x})$ computed by the detector is as close as possible to $f(y)$. This problem is addressed as the minimization of the distance $\Phi(z) = \mid f(\tilde{x}) - f(y) \mid$ under a perceptual distortion constraint on x, by resorting to iterative techniques based on gradient descent.

The idea of using minimization techniques to derive a counter-forensic attack is not novel; in fact Fan *et. al*[18] proposed a constrained total variation minimization method for preventing the detection of JPEG deblocking. However, in this case the function to be minimized is assumed to be known: since the function is convex, the authors solve the optimization problem through gradient descent finding the analytical solution. Unlike the aforementioned approach, our aim is to generalize the application of minimization techniques also when it is not possible to achieve a closed form solution for the minimization problem, as it happens when the function computing the feature is unknown or too complex. Therefore, we propose a generic numerical computation strategy that does not depend on the considered feature/detector.

An intrinsic limit in the use of numerical techniques for the estimation of the gradient, is the resulting heavy computational effort that depends on the size of the attacked image. In order to make possible the application of the proposed strategy to large images, we turn the solution of the whole problem into the solution of many sub-problems of smaller size. This idea comes from the observation that many forensic features are based on local image properties, thus suggesting the possibility to decompose the image into non overlapping blocks, solve the constrained minimization separately on each

block, and then recompose the attacked blocks to obtain the final attacked image $\tilde{x}$. Unfortunately, in most cases of practical interest, the value of $f(\tilde{x})$ computed on $\tilde{x}$ is not close enough to $f(y)$ because of an intrinsic error in the recomposition process. Such a limit is overcome by considering a modified version of distance function Φ that takes into account the decomposition error.

The details of the proposed strategy are given in the following sections. In Section 2, the addressed problem is described and the feature injection strategy is detailed; in Section 3, the solutions proposed to overcome the practical problems are presented. The validity of this new counter-forensic approach is demonstrated in Section 4, where three different feature detectors are considered and the proposed algorithm is compared to some state-of-the-art counter-forensic attacks. Concluding remarks and future extensions are discussed in Section 5.

2. PROBLEM STATEMENT

Let us assume that we have a forensic detector that classifies an image in $\mathbb{R}^N$ according to the value of a feature $f : \mathbb{R}^N \rightarrow I$, where I is an open subset of $\mathbb{R}$ and f is based on local statistical properties of a particular footprint. The set I is partitioned into two sets A and $\bar{A}$ such that if $f(y) \in A$ we say that the value of f supports a certain hypothesis on the history of the image y. We also assume that f is $C^2(I)$ and that $A = \bigcup_{i=1}^{s} A_i$ with A_i is an open set $\forall i = 1, \ldots, s$. This assumption means that if $f(y) \in A$ then we can find $\delta > 0$ such that $B_\delta(f(y)) \in A$, where $B_\delta(o)$ is a ball of radius δ centered in o.

Let us consider a forensic adversary that, given a generic image x with $f(x) \in \bar{A}$, i.e., the value of f points to a different history for x, wants to produce an attacked image $\tilde{x}$ such that $f(\tilde{x}) \in A$ and $\tilde{x}$ is perceptually indistinguishable from x. Formally, the problem is

$$\tilde{x} : \tilde{x} \in \mathrm{Q}_\Delta(x),\ f(\tilde{x}) \in A \tag{1}$$

where $\mathrm{Q}_\Delta(x)$ is a suitable perceptual distortion constraint with respect to x. In particular, we define as perceptual constraint

$$\mathrm{Q}_\Delta(x) = \{z \in \mathbb{R}^N\ :\ |z_i - x_i| < \Delta_i,\ i = 1, \ldots, N\} \tag{2}$$

Depending on the constraint $\mathrm{Q}_\Delta(x)$ the above problem may not have a solution. In the following, we will assume that there always exists a convenient configuration of Δ_i such that the above problem has solution. The rationale is that it is reasonable to think that for every image x there is always an image $\tilde{x}$ showing the same scene, and hence perceptually close to x, for which the value of f points to a specific history.

Let us assume that the adversary has access to a black box computing f. A possible solution, referred in the watermarking literature as sensitivity attack, is just to find an initial z_0 such that $f(z_0) \in A$ and then minimize the distance between z_0 and x by letting z_0 move on the boundary of the acceptance region A. Unfortunately, the above solution can not be easily applied in the forensic scenario, since it is in general very difficult to find a good starting point z_0.

An alternative strategy is to assume that the adversary has a partial description of A by means of a training set Y composed of images satisying $f(y) \in A$ for each image $y \in Y$. If, given a random image $y \in Y$, the adversary is able to produce a attacked image $\tilde{x}$ such that $f(\tilde{x}) \in B_\delta(f(y))$ and $\tilde{x}$ is close to the target image x according to a suitable perceptual metric, then $\tilde{x}$ is a solution to our problem.

The adversary has basically two ways for finding an useful $\tilde{x}$. The first one is to start from y and try to minimize the perceptual distance between y and x while keeping the variations of $f(y)$ bounded. The second one is to start from x and try to minimize the distance between $f(x)$ and $f(y)$ while keeping the perceptual distortions of x bounded. In the following, we choose the second approach, since the distance between x and an arbitrary authentic image y chosen from the training set can be very large, and hence difficult to optimize.

The problem in (1) can be addressed by minimizing the distance $\Phi(z) = |f(z) - f^*|$, where $f^* \triangleq f(y)$, under a perceptual distortion constraint with respect to the image x.

We propose to solve the above problem through the constrained minimization of $\Phi(z)$ by iterative methods based on gradient descent; in particular, since the detector computing f is considered to be unknown, the gradient has to be numerically estimated. We have no guarantee that the minimization procedure actually achieves the minimum, since: *i)* f may have an arbitrary structure, so it is not possible to define a unique minimization technique that assures to reach

the target minimum for every f; *ii)* the evaluation of the gradient may be inaccurate if f is locally noisy, affecting the convergence of gradient descent methods.

However, in order to solve problem (1) it is not necessary to exactly achieve the minimum for the function $\Phi(z)$, the only requirement is that the solution must be in A; in the following we will describe the proposed methodology.

3. PROPOSED METHOD

The reduction of $\Phi(z)$ consists in the application of a minimization function $\mathcal{M}$ based on gradient descent such that, given a starting point x, it produces an output $\hat{x} = \mathcal{M}(x)$ so that

$$\Phi(\hat{x}) = \tau,\ \hat{x} \in \mathrm{Q}_{\Delta}(x) \tag{3}$$

where τ depends on how successful is the reduction of $\Phi(z)$. If $\tau < \delta_M$, where δ_M is the greatest radius $\delta > 0$ such that $B_\delta(f^*) \in A$, then we have $f(\hat{x}) \in A$ and $\hat{x}$ is a solution to the problem in (1).

3.1 Feature Decomposition

The numerical estimation of the gradient requires a number of evaluations of the function $\Phi(z)$ equal to the image size N. Moreover the computational cost of evaluating f usually increases with respect to the size of the image, so that the complexity of computing a single gradient vector is more than linear with N, thus hindering the application of the suggested approach to large images.

The resulting heavy computational effort can be faced by turning the solution of the whole problem into the solution of many sub-problems of smaller size. This idea comes from the observation that many forensic features are based on local image properties. We then propose the application of a feature decomposition process: the image x is partitioned into non overlapping blocks and the feature $f(x)$ is evaluated as a combination of the features computed on the single blocks. Hence, the proposed counter-forensic attack, i.e. the constrained minimization process, is applied separately on each block, and the final attacked image is obtained by recomposing the attacked blocks.

This decomposition can be formally expressed as follows: let an image $z \in \mathbb{R}^N$, $\mathcal{P} = \{b_1, \ldots, b_m\}$ a block-based partition of the image and $a = (a_1, \ldots,\ a_m)$ with $a_i \geq 0$ such that $\sum_{i=1}^m a_i = 1$. Then f can be decomposed with respect to $\mathcal{P}$ and a as follows

$$f(z) = \sum_{i=1}^{m} a_i f(z_{b_i}) + \epsilon(f, \mathcal{N}, z), \tag{4}$$

where z_{b_i} is the restriction of z to the components of the $i-$th block, $f(z_{b_i})$ is the feature evaluated on the $i-$th block, and $\epsilon(f, \mathcal{N}, z)$ is a bounded error term intrinsic in the recomposition process. The recomposition error depends on the feature, on the particular block-based partition adopted for the feature decomposition and on the image itself. From now on we will refer to this error as $\epsilon(z)$ for simplicity.

Then the distance function Φ can be expressed as

$$\begin{aligned} \Phi(z) &= \left| \sum_{i=1}^{m} a_i f(z_{b_i}) + \epsilon - \sum_{i=1}^{m} a_i f^* \right| \leqslant \\ &\leqslant \sum_{i=1}^{m} a_i |f(z_{b_i}) - f^*| + |\epsilon(z)| = \sum_{i=1}^{m} a_i \Phi(z_{b_i}) + |\epsilon(z)| \end{aligned} \tag{5}$$

In practice, the application of the minimization function $\mathcal{M}$ on the m blocks leads to m solutions $\hat{x}_{b_i}$ such that $\Phi(\hat{x}_{b_i}) = \tau_i,\ i = 1, \ldots, m$. Hence, for the recomposed image $\hat{x}_r$ the distance from the target f^* results

$$\Phi(\hat{x}_r) \leqslant \sum_{i=1}^{m} a_i \tau_i + |\epsilon(\hat{x}_r)| \tag{6}$$

The recomposition error $\epsilon(\hat{x}_r)$ affects the final distance $\Phi(\hat{x}_r)$: if $\epsilon(\hat{x}_r)$ could be compensated and $\sum_{i=1}^m a_i \tau_i < \delta_M$, we would achieve the solution of problem (1): in particular, the error could be removed by choosing as target $f^* + \epsilon(\hat{x}_r)$.

On the contrary, if we are not able to compensate the term $\epsilon(\hat{x}_r)$, we can face two possible errors: i) $f(\hat{x}_r)$ falls between $f(x)$ and f^*, ii) $f(\hat{x}_r)$ falls beyond $f(x)$ and f^*. In the former case we still have to solve the same problem, having as starting point $f(\hat{x}_r)$ instead of $f(x)$. But in the latter case we have found a new point so that f^* is between $f(x)$ and $f(\hat{x}_r)$, thus suggesting the application of a bisection method in order to achieve the target f^*: in this case we can move from a difficult problem to a simple one.
So our proposed strategy is to consider as target $f^* + R$, where R is an over-estimation of the recomposition error, so that we have the second kind of error and we can then apply the bisection method.

In detail, let us suppose, without loss of generality, that $f(x) < f^*$: the proposed technique consists in two steps: 1) change the distance function Φ considering $f^* + R$ instead of f^* in order to find as output of $\mathcal{M}$ an image whose feature f is greater than f^*; 2) find the solution of the problem in (1) through a bisection algorithm. In particular, for each of the m blocks the image is partitioned into, we aim at injecting as reference feature $f^* + R$. The corresponding solutions are then recomposed leading to an image $\hat{x}_r$ that when tested by the detector will output the value $f(\hat{x}_r) > f^*$. Then, by applying the bisection method we find along the convex combination of the images x and $\hat{x}_r$ the final attacked image x^* so that $f(x^*) = f^*$.

3.2 Bisection Method

If the injection procedure achieves the goal of producing an output image $\hat{x}$ such that $f(\hat{x}) > f^*$, we can easily exploit a bisection algorithm to find a solution to problem (1). Indeed, let us consider the target function $F(z) = f(z) - f^*$ and observe that $F(x) < 0$ and $F(\hat{x}) > 0$. Then, thanks to the regularity of $F(x)$, it is possible to find a value $\lambda \in (0, 1)$ such that $F(x + \lambda(\hat{x} - x)) = 0$, i.e. the image $z^* = x + \lambda(\hat{x} - x)$ is such that $f(z^*) = f^*$. Moreover, it is easy to show that $z^* \in \mathrm{Q}_\Delta(x)$, since both $x \in \mathrm{Q}_\Delta(x)$ and $\hat{x} \in \mathrm{Q}_\Delta(x)$, the latter being a solution to the constrained problem, and z^* is found along the convex combination of x and $\hat{x}$.
In the next section we propose an improvement of this standard bisection procedure in order to optimise the resulting image under a perceptual point of view.

3.2.1 Perceptual Bisection Method based on Local Variance Map

The solution $z^* = x + \lambda(\hat{x} - x)$ consists in the image x distorted toward $\hat{x}$ with the same weight λ on each pixel. On the other hand, it is well known that the same variation on a single pixel is much more perceptible when the local variance of the content is low. Consequently we propose to take into account such a property exploiting the decomposition of the target image x in different layers $\{x^{L_1}, \dots, x^{L_{s+1}}\}$ depending on the local variance of the content. Let $0 < \sigma_1 < \cdots < \sigma_s$ be suitable real thresholds: we consider $\mathcal{L} = \{L_1, \dots, L_{s+1}\}$ a partition of the indexes $\mathcal{N} = \{1, 2, 3, \dots, N\}$ such that

$$
\begin{aligned}
&x_j \in L_1, \text{ if } \sigma(x_j) < \sigma_1, \\
&x_j \in L_i, \text{ if } \sigma(x_j) \in [\sigma_{i-1}, \sigma_i), i = 2, \dots, s \\
&x_j \in L_{s+1}, \text{ if } \sigma(x_j) \geqslant \sigma_s
\end{aligned}
$$

where $\sigma(x_j)$ is the local variance of the pixel x_j. Then the target image x can be decomposed in $x^{L_1} + x^{L_2} + \cdots + x^{L_{s+1}}$ where $x^{L_i} \in \mathbb{R}^N$ and

$$
x_j^{L_i} = \begin{cases} x_j & \text{if } j \in L_i \\ 0 & \text{otherwise} \end{cases} \tag{7}
$$

and the same decomposition can be applied to $\hat{x}$, referring to the same layers obtained for x, so that $\hat{x} = \hat{x}^{L_1} + \hat{x}^{L_2} + \cdots + \hat{x}^{L_{s+1}}$. Let us consider the sequence of images

$$
\begin{aligned}
&w_0 = \hat{x}, \\
&w_k = \hat{x} + \sum_{j=1}^{k} (x^{L_j} - \hat{x}^{L_j}), \quad k = 1, \dots, s+1
\end{aligned}
$$

composed by the first k layers of x and the last $s-k+1$ layers of $\hat{x}$. Obviously, we have $F(w_0) > 0$ and $F(w_{s+1}) < 0$, so that it must exist at least an index $t \leqslant s$ such that $F(w_t) > 0$ and $F(w_{t+1}) < 0$. Let t be the greatest index such that $F(w_t) > 0$ and $F(w_{t+1}) < 0$: since the images w_t and w_{t+1} differ only by the layer L_{t+1}, i.e.

$$w_t = x^{L_1} + \cdots + x^{L_t} + \hat{x}^{L_{t+1}} + \hat{x}^{L_{t+2}} + \cdots + \hat{x}^{L_{s+1}}$$
$$w_{t+1} = x^{L_1} + \cdots + x^{L_t} + x^{L_{t+1}} + \hat{x}^{L_{t+2}} + \cdots + \hat{x}^{L_{s+1}}.$$

the standard bisection method can be applied using w_t and w_{t+1} as starting points. In this way, the first t layers of the solution are exactly the first t layers of x, reducing the possibility of introducing distortion on those areas of the image where such artefacts would be more perceptible.

4. EXPERIMENTAL RESULTS

The proposed strategy has been practically tested by attacking some state-of-the-art detectors: in particular, to demonstrate its applicability to different features, we chose three different image footprints, namely Color Filter Array (CFA) artefacts, photo response non-uniformity (PRNU) noise, and JPEG compression traces. For each of these traces, we selected one feature detector, and we applied the proposed attack. Furthermore, the method has been compared with some state-of-the-art counter-forensic methods in order to evaluate the progress with respect to prior works.

4.1 Considered Features

In the following, we summarize the chosen feature detectors, as well as the corresponding attacks chosen to deceive each detector.

4.1.1 PRNU

PRNU is a high-frequency multiplicative noise, generally stable throughout the camera's lifetime in normal operating conditions, that is unique to each camera. The following simplified model for the image signal can be assumed[19]

$$\mathbf{I} = \mathbf{I}^{(0)} + \mathbf{K}\mathbf{I}^{(0)} + \Psi \tag{8}$$

where $\mathbf{I}$ is the signal in a selected color channel, $\mathbf{I}^{(0)}$ denotes the captured light in absence of any noise or imperfections, $\mathbf{K}$ is a zero-mean noise-like signal responsible for PRNU, and Ψ is a combination of random noise components.

To extract the PRNU, a filtered version of $\mathbf{I}$, $F(\mathbf{I})$, obtained through Mihcak denoising filter[20] F is subtracted from $\mathbf{I}$

$$\mathbf{W} = \mathbf{I} - F(\mathbf{I}) = \mathbf{I}\mathbf{K} + \Phi \tag{9}$$

where Φ is the sum of Ψ and two additional terms introduced by the denoising filter. By assuming to have a set of N images $\mathbf{I}_k$ acquired by the same camera, and to apply the previous procedure to these images to obtain the terms $\mathbf{W}_k$, the maximum likelihood predictor for $\mathbf{K}$ is then formulated as

$$\mathbf{K} = \frac{\sum_{k=1}^{N} \mathbf{W}_k \mathbf{I}_k}{\sum_{k=1}^{N} (\mathbf{I}_k)^2} \tag{10}$$

Now, to assess whether an image $\mathbf{I}$ has been taken with a given camera, it is requested to extract its noise term $\mathbf{W} = \mathbf{I} - F(\mathbf{I})$, and then to compute a correlation between $\mathbf{W}$ and the PRNU characterizing that camera[21]

$$\rho = \mathbf{I}\mathbf{K}_i \otimes \mathbf{W} \tag{11}$$

where $\otimes$ denotes normalized correlation. If the correlation is higher than a given threshold, the camera is identified as the source of the image.

Corresponding Attack A fake camera fingerprint can be estimated from a set of acquired images and then superimposed to an image from a different camera with the aim to introduce a false source camera identification. A practical attack in this sense is described by Gloe *et al.*[5]

4.1.2 CFA

Along with PRNU, another important artifact left by cameras during acquisition is due to the presence of the Color Filter Array (CFA), filtering the incoming light before reaching the sensor, so that for each pixel only one particular color is gathered. As a consequence, for each color channel only one third of pixels of the image are directly sensed, while the others are interpolated.

Ferrara *et al.*,[22] by means of a local analysis of the CFA, identify image forgeries whenever the presence of CFA interpolation is not present. Starting from such an assumption, a new feature is proposed, that measures the presence/absence of these artifacts even at the smallest 2×2 block level, thus providing as final output a forgery map indicating with fine localization the probability of the image to be manipulated.

Given a $N \times N$ image, its green channel is extracted and analyzed. In particular the local weighted variance of the prediction error is computed as

$$\sigma_e^2(x,y) = \frac{1}{c}\Big[\Big(\sum_{i,j=-K}^{K} \alpha_{ij}\; e^2(x+i,y+j)\Big) - (\mu_e)^2\Big] \tag{12}$$

where α_{ij} are suitable weights

$$\mu_e = \sum_{i,j=-K}^{K} \alpha_{ij}\; e(x+i,y+j) \tag{13}$$

is a local weighted mean of the prediction error and

$$c = 1 - \sum_{i,j=-K}^{K} \alpha_{ij}^2 \tag{14}$$

is a scale factor that makes the estimator unbiased, i.e., $E[\sigma_e^2(x,y)] =$var$[e(x,y)]$, for each pixel class.

By considering $b \times b$ non-overlapping blocks $\mathcal{B}_{k,l}$ with $k,l = 0,\ldots,\frac{N}{b}-1$, and where b is 2, or its multiples and noting that each block is composed by disjoint sets of acquired and interpolated pixels, indicated as $\mathcal{B}_{A_{k,l}}$ and $\mathcal{B}_{I_{k,l}}$ respectively, the following feature $\mathbf{L}$ is computed

$$\mathbf{L}(k,l) = \log\left[\frac{GM_A(k,l)}{GM_I(k,l)}\right] \tag{15}$$

where $GM_A(k,l)$ is the *geometric mean* of the acquired pixels defined as

$$GM_A(k,l) = \left[\prod_{i,j\in\mathcal{B}_{A_{k,l}}} \sigma_e^2(i,j)\right]^{\frac{1}{|\mathcal{B}_{A_{k,l}}|}} \tag{16}$$

whereas $GM_I(k,l)$ is similarly defined for the interpolated pixels. The proposed feature $\mathbf{L}$ allows to detect the imbalance between the local variance of prediction errors when an image is demosaicked: indeed, in this case the local variance of the prediction error of acquired pixels is higher than that of interpolated pixels and thus the expected value of $\mathbf{L}(k,l)$ is a nonzero positive amount. On the other hand, if an image is not demosaicked, this difference between the variance of prediction errors of acquired an interpolated pixels disappears, since the content can be assumed locally uniform, and the expected value of $\mathbf{L}(k,l)$ is zero. The feature

$$\kappa = \left(\frac{b}{N}\right)^2 \sum_{k,l} \mathbf{L}(k,l) \tag{17}$$

i.e. the mean of $\mathbf{L}$ on all blocks is adopted to distinguish between the presence or absence of demosaicking artefacts on the whole image.

Corresponding Attack Kirchner and Böhme[8] proposed a method to synthetically create or restore the characteristic local correlation pattern that originates from the CFA interpolation in typical image acquisition devices. Given the CFA pattern, interpolated pixels value are estimated from acquired pixels through the application of an interpolation filter to restore the demosaicing artefacts.

4.1.3 JPEG

In[23] the authors, observing that in natural images AC DCT coefficients follow an approximate Laplacian distribution, concentrate their attention on the integral of the AC DCT coefficient histogram p_{AC} in the range $R_1 = (-1, +1)$, and in the range $R_2 = (-2, -1] \cup [+1, +2)$. Then, they compute the feature

$$\eta = \frac{\int_{R_2} p_{AC}(y)dy}{\int_{R_1} p_{AC}(y)dy} \tag{18}$$

By observing that in a JPEG compressed image, given the quantization and dequantization processes, the value of the integral in R_1 will increase, while the value of the integral in R_2 will decrease, the proposed feature on a JPEG image will be close to zero and its value would be much smaller than that obtained on an uncompressed image. Hence, they propose to set a threshold on this ratio, so that JPEG compression is detected when the ratio is smaller than the threshold.

Corresponding Attack A variational approach to remove JPEG traces is proposed by Fan *et al.*[18] The main idea consists is the constrained minimization of an energy function based on total variation to improve visual quality of compressed images and reduce the statistical difference between the pixel value variation along the block borders and within the blocks. Even if this kind of attack has not been tested on the feature described above, it is one of the most recent counter forensic techniques to hide traces of JPEG compression and the authors have shown its effectiveness against many known detectors proposed by Valenzise *et al.*,[13] Lai *et al.*,[14] and Fan *et al.*[16] for the identification of JPEG compressed images.

4.2 Results

The results presented in this section have been obtained using as reference a Dataset of 100 uncompressed RGB images from a Nikon D90 camera. Each image has been cropped to 512×512 pixels in its centre. We will denote such a dataset* with Y and we will refer to it as the original dataset because each image $y \in Y$ contains CFA demosaicking artefacts and PRNU fingerprint, both belonging to the specific Nikon source. Furthermore, no compression traces can be found on y.

The experiments have been carried out in two parts: first of all we show the decomposition property of the features ρ, κ and η estimating the decomposition error ϵ for each feature; then we test the effectiveness of our counter forensics technique on the three features through the comparison with the state-of-the-art counter forensics attacks.

4.2.1 Decomposition Error

As shown in section 3.1 the distribution of the decomposition error ϵ depends on many different factors (feature f, block size of partition $\mathcal{P}$, image z and weights a_i). We have estimated mean and variance of ϵ on the three features fixing Y as reference set, the same weight a_i for all i and testing blocks b_i with different sizes in order to understand the relations between decomposition error and size of the blocks. With reference to (4) we evaluated for each feature

$$\epsilon(y) = f(y) - \sum_{i=1}^{m} \frac{1}{m} f(y_{b_i}) \tag{19}$$

for all $y \in Y$ and several block sizes (128×128, 64×64, 32×32, 16×16, 8×8). In tables 1, 2 and 3 we report mean and standard deviation of the relative error $|\epsilon_r(z)| = \left|\frac{\epsilon(z)}{f(z)}\right|$.

The results shows that it is reasonable to employ the decomposition process on these features because the relative error committed in the decomposition is bounded, being lower than 50% with large probability even when block size is 8×8. As we expected, to a larger block size corresponds a smaller decomposition error, meaning that a larger block size usually

*Available at `http://lesc.det.unifi.it/en/datasets`

blk size	8×8	16×16	32×32	64×64	128×128
Mean	0.153	0.130	0.107	0.082	0.055
Std	0.131	0.127	0.111	0.091	0.066

Table 1. $|\epsilon_r|$ distribution for ρ

blk size	8×8	16×16	32×32	64×64	128×128
Mean	0.270	0.142	0.071	0.033	0.014
Std	0.033	0.018	0.010	0.005	0.003

Table 2. $|\epsilon_r|$ distribution for κ

blk size	8×8	16×16	32×32	64×64	128×128
Mean	0.197	0.107	0.072	0.053	0.039
Std	0.156	0.131	0.118	0.100	0.091

Table 3. $|\epsilon_r|$ distribution for η

yields a better decomposition. On the other hand, the computational effort for evaluating the gradient strongly increases with the size of the blocks. Looking for a tradeoff between decomposition error and computational cost, we chose the block size 16×16 as the best compromise for all features.

4.2.2 Counter Forensics Effectiveness

The injection procedure has been carried out independently on each of the considered features. First of all, we give a short description of the experimental procedure we employed to inject a generic feature f, then we will give more details for each specific attack.
Given a forensic detector based on feature f and a dataset $X = \{x^{(1)}, \ldots, x^{(m)}\}$ such that $f(x^{(i)}) \in \bar{A}, \forall i = 1, \ldots, m$, we consider a target dataset $Y = \{y^{(1)}, \ldots, y^{(m)}\}$ such that $f(y^{(i)}) \in A, \forall i = 1, \ldots, m$. Then, we inject the feature $f(y^{(i)})$ into each image $x^{(i)}$ through the following two-steps procedure[†]:

- constrained minimization of $\Phi_R(z) = |f(z) - (f(y^{(i)}) + R)|$, using $x^{(i)}$ as starting point and R the over-estimation of the error introduced by feature decomposition for f;
- bisection procedure as described in Section 3.2.1.

We reduce $\Phi_R(z)$ on each block of the attacked image through a modified BFGS (Broyden-Fletcher-Goldfarb-Shanno) method[24] estimating the gradient through the forward differences

$$\nabla_i \Phi_R(z) \simeq \frac{\Phi_R(z + he_i) - \Phi_R(z)}{h} \tag{20}$$

where $e_i > 0$ is a vector of the standard basis[‡] and $h > 0$ is a conveniently chosen increment value. For the details on hessian matrix update we referred to Tim Kelley implementation[§].

Following the described procedure we injected the three features ρ, κ and η extracted by the original dataset Y separately. Therefore we built three different datasets[¶]: a set P without traces of the Nikon D90 sensor noise, a set C without demosaicking artefacts and a set $J^{(q)}$ of JPEG compressed images with quality factor q. Each dataset has been built as follows:

- set P: 100 RGB uncompressed images taken by a Canon camera and cropped to 512×512 pixels, thus no traces of Nikon sensor noise are contained on such a dataset.

[†]The decision to choose a different target for each image is to avoid the generation of statistic anomalies of the features values on attacked images.

[‡]A more generic approach may consider any basis.

[§]`http://www4.ncsu.edu/~ctk/`

[¶]Available at `http://lesc.det.unifi.it/en/datasets`

- set C: starting from set P each color channel of each image has been upsampled by a factor two, blurred with a 7×7 median filter, and downsampled by a factor two, thus removing all CFA artifacts.
- set $J^{(q)}$: starting from P the luminance component of each image has been compressed in JPEG format. Three different quality factors $(80, 60, 40)$ have been tested to highlight the dependence of the effectiveness of the attack with respect to the compression level. We will refer to such datasets as $J^{(80)}$, $J^{(60)}$, $J^{(40)}$, in that order.

Then the sets P, C and $J^{(q)}$ have been injected with the features ρ, κ and η, in that order. Many common parameters have been used for the injection of each feature:

- number of images for each dataset $m = 100$;
- blocks size of 16×16 for feature decomposition;
- perceptive constrain Δ fixed to 5 for each pixel;
- local variance σ evaluated for each pixel on a 7×7 gaussian window (centred on the pixel);
- number of layers $s = 5$;
- thresholds $\sigma_1, \ldots, \sigma_5$ chosen on each image as the quantiles $q_{0.20}, q_{0.40}, q_{0.60}, q_{0.80}, q_{1.00}$ of local variance distribution of the same image;

On the contrary, some parameters have been experimentally tuned for a specific feature. Namely, the step length h in equation (20) has been fixed to:

- $h = \frac{1}{255}$ for the injection of ρ in P;
- $h = \frac{1}{255}$ for the injection of κ in C;
- $h = \frac{3}{255}, \frac{4}{255}, \frac{5}{255}$ for the injection of η in $J^{(80)}$, $J^{(60)}$, $J^{(40)}$ respectively.

Different values of R have been experimentally evaluated overestimating the recomposition error on each feature and the best performances have been obtained with $R = 0,22$ for feature ρ, $R = 2,71$ and $R = 0,72$ for feature κ and η respectively.

In the next subsections we will show the obtained results on each feature through a comparison between the proposed injection method and the three attacks described in the beginning of the section. The effectiveness of the counter-forensic attacks has been evaluated in terms of receiver operating characteristic (ROC) curve and area under the ROC curve (AUC), while the distortion introduced by the attacks has been evaluated in terms of peak signal to noise ratio (PSNR).

4.2.3 Results on feature ρ

Starting from P we produced the attacked sets $\widehat{P}$ and $\widetilde{P}$ obtained through our injection procedure and the attack proposed by Gloe *et al.*[5] respectively.
Finally, the PRNU detector[21] has been applied to both sets of attacked images and to the original set Y. In Figure 1 we present the ROC curves obtained on the two attacked image sets and in Table 4 we show the differences between P and each attacked sets in terms of average PSNR and AUC. Since the adversary aims at obtaining a random classification, the achieved results demonstrate that the proposed injection strategy obtains good performance, improving the result of AUC with respect to the method proposed by Gloe *et al.*, while maintaining a low distortion of the contents.

	PSNR	AUC
$\widehat{P}$	54,30	0,51
$\widetilde{P}$	63,36	0,56

Table 4. Average PSNR and AUC values for feature ρ comparing P with the attacked sets $\widehat{P}$ and $\widetilde{P}$ obtained through our injection procedure and the attack proposed by Gloe *et al.*[5] respectively.

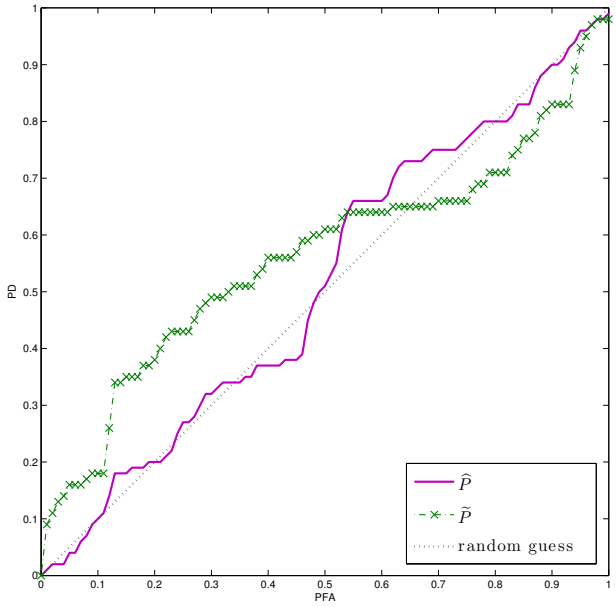

Figure 1. ROC curves for feature ρ on the attacked sets $\widehat{P}$ and $\widetilde{P}$ obtained through our injection procedure and the attack proposed by Gloe *et al.*[5] respectively.

4.2.4 Results on feature κ

Starting from C we produced the attacked set $\widehat{C}$, where feature κ has been injected through our procedure. In addition, the method to synthetically create a CFA fingerprint[8] has been applied on C exploiting a bilinear, a bicubic, and a gradient based interpolation filter producing attacked sets $\widetilde{C}^{(bl)}$, $\widetilde{C}^{(bc)}$, $\widetilde{C}^{(gb)}$ respectively. Finally, the CFA detector[22] has been applied to each datasets of attacked images and to the original set Y. In Figure 2 we present the ROC curve obtained on the four attacked image sets: since the adversary aims at obtaining a curve corresponding to random classification, it is evident that the proposed attack is more effective than the others in deceiving the detector. In addition, in Table 5 we show the differences between C and each attacked set in terms of average PSNR and AUC: the proposed approach outperforms the other attacks, in terms of both low distortion of the attacked images and success rate.

4.2.5 Results on feature η

Starting from $J^{(q)}$ we produced, for each quality factor $q = 80, 60, 40$, an attacked set $\widehat{J}^{(q)}$ obtained through the injection procedure. In addition, the attack proposed by Fan *et al.*[18] has been employed generating three attacked sets $\widetilde{J}^{(q)}$. Finally, the JPEG detector[23] has been applied to the attacked images and the original set Y. In Figure 3 we compare the results through the ROC curves obtained on the two attacked image sets for each quality factor; in Table 6 it is also shown that the feature injection offers better results in terms of AUC, and similar results in terms of distortions, with the only exception of the set heavily compressed, where the distortion becomes higher.

	PSNR	AUC
$\widehat{C}$	51,70	0,51
$\widetilde{C}^{(bl)}$	45,91	0,25
$\widetilde{C}^{(bc)}$	47,89	0,98
$\widetilde{C}^{(gb)}$	49,66	0,98

Table 5. Average PSNR and AUC values for feature κ comparing C with the attacked sets $\widehat{C}$, $\widetilde{C}^{(bl)}$, $\widetilde{C}^{(bc)}$, $\widetilde{C}^{(gb)}$, where $\widehat{C}$ is obtained through our injection procedure and $\widetilde{C}^{(bl)}$, $\widetilde{C}^{(bc)}$, $\widetilde{C}^{(gb)}$ through the attack proposed in[8] exploiting a bilinear, a bicubic and a gradient based interpolation filter respectively.

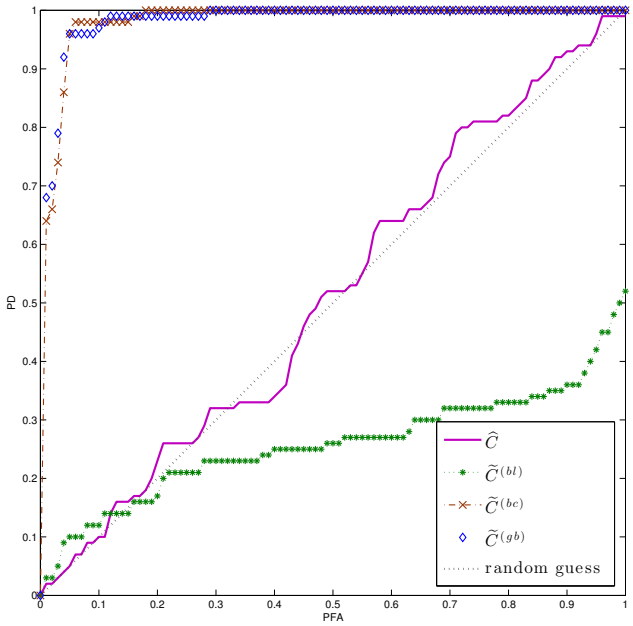

Figure 2. ROC curves for feature κ on the attacked sets $\widehat{C}$, $\widetilde{C}^{(bl)}$, $\widetilde{C}^{(bc)}$, $\widetilde{C}^{(gb)}$, where $\widehat{C}$ is obtained through our injection procedure and $\widetilde{C}^{(bl)}$, $\widetilde{C}^{(bc)}$, $\widetilde{C}^{(gb)}$ through the attack proposed in[8] exploiting a bilinear, a bicubic and a gradient based interpolation filter respectively.

	PSNR	AUC
$\widehat{J}^{(80)}$	42,77	0,59
$\widetilde{J}^{(80)}$	41,69	0,78

	PSNR	AUC
$\widehat{J}^{(60)}$	41,31	0,62
$\widetilde{J}^{(60)}$	40,88	0,79

	PSNR	AUC
$\widehat{J}^{(40)}$	38,13	0,69
$\widetilde{J}^{(40)}$	40,63	0,80

Table 6. Average PSNR and AUC values for feature η comparing $J^{(q)}$ $(q = 80, 60, 40)$ with the attacked sets $\widehat{J}^{(q)}$ and $\widetilde{J}^{(q)}$ obtained through our injection procedure and the attack proposed by Fan *et al.*[18] respectively.

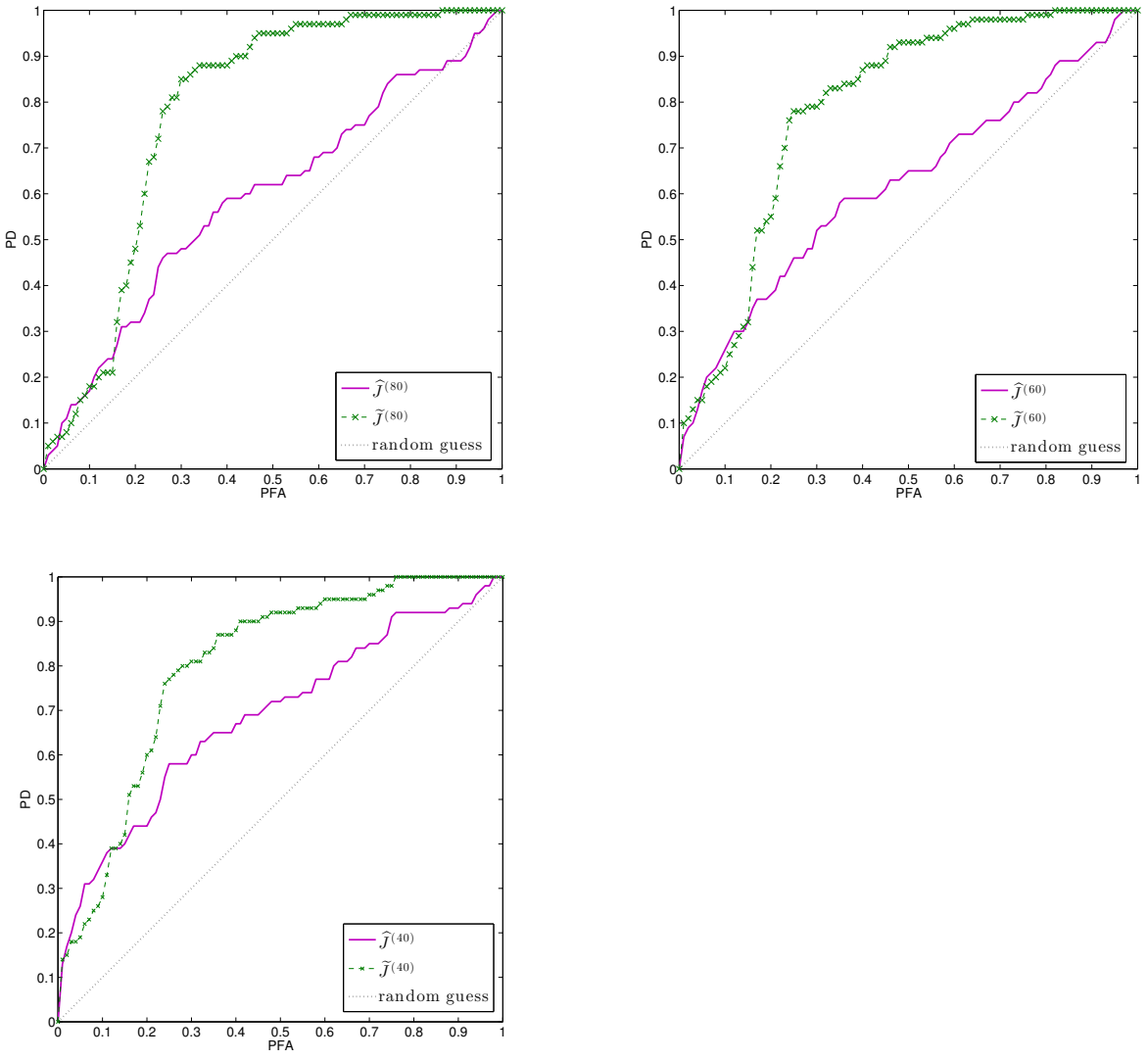

Figure 3. ROC for feature η on the attacked sets $\widehat{J}^{(q)}$ and $\widetilde{J}^{(q)}$ (for $q = 80, 60, 40$) obtained through our injection procedure and the attack proposed by Fan *et al.*[18] respectively..

5. CONCLUSIONS

In this work, we modelled the counter-forensic attack as the injection into a digital image of a specific feature that points to a given history the adversary wants to be believed for the attacked image. Specifically, by considering a target image telling the desired history, the to-be-attacked image is processed in the pixel domain producing an attacked image whose feature points to the same history of the reference image. In this way, when the target image and the fake-and-attacked image are analysed, the detector will give the same answer on both images. The proposed attack consists in a general strategy based on minimization techniques, which can be applied irrespective of the specific feature and the considered detector. Computational problems due to numerical optimization have been effectively faced through a feature decomposition process, while an improved bisection procedure has been proposed to obtain a better control on the distortion introduced by the proposed attack. We tested our strategy by applying the proposed feature injection attack to three different features and comparing the achieved performance with the results obtained by some state-of-the-art counter-forensic methods. The approach in principle outperforms the considered attacks in terms of success rate and reduced distortion of the attacked images. Furthermore, a significant benefit of the proposed strategy is that it only needs the soft output of the considered detector, without resorting to any detailed information on the considered feature/detector. On the contrary, most of state-of-the-art attacks rely on specific knowledge, e.g., the technique using interpolation filter to restore demosaicking artefact requires the knowledge of the CFA pattern; or the technique simulating the PRNU of a specific sensor requires a good estimation of the sensor noise. The natural evolution of this work is the application of the feature injection attack to other features, that can be decomposed with bounded relative error. Furthermore, we will explore the possibility to inject more than one feature at the same time, in order to make the image telling a more complicated history (e.g. an image coming from a given camera with a specific PRNU and a specific CFA interpolation).

ACKNOWLEDGMENTS

This work was partially supported by the REWIND Project funded by the Future and Emerging Technologies (FET) programme within the 7FP of the European Commission, under FET-Open grant number 268478.

REFERENCES

[1] Piva, A., "An overview on image forensics," *ISRN Signal Processing* **2013**, Article ID 496701, 22 pages (2013).

[2] Böhme, R. and Kirchner, M., [*Counter-Forensics: Attacking Image Forensics*], 327–366, Springer-Verlag (2013).

[3] Kirchner, M. and Böhme, R., "Hiding traces of resampling in digital images," *IEEE Transactions on Information Forensics and Security* **3**(4), 582–592 (2008).

[4] Popescu, A. C. and Farid, H., "Exposing digital forgeries by detecting traces of re-sampling," *IEEE Transactions on Signal Processing* **53**(2), 758–767 (2005).

[5] Gloe, T., Kirchner, M., Winkler, A., and Böhme, R., "Can we trust digital image forensics?," in [*International Conference on Multimedia*], 78–86 (2007).

[6] Goljan, M., Fridrich, J., and Chen, M., "Defending against fingerprint-copy attack in sensor-based camera identification," *IEEE Transactions on Information Forensics and Security* **6**(1), 227–236 (2011).

[7] Caldelli, R., Amerini, I., and Novi, A., "An analysis on attacker actions in fingerprint-copy attack in source camera identification," in [*IEEE International Workshop on Information Forensics and Security*], (2011).

[8] Kirchner, M. and Böhme, R., "Synthesis of color filter array pattern in digital images," in [*SPIE Conference on Media Forensics and Security*], Delp, E. J., Dittmann, J., Memon, N., and Wong, P. W., eds. (2009).

[9] Stamm, M., Tjoa, S., Lin, W. S., and Liu, K. J. R., "Anti-forensics of JPEG compression," in [*International Conference on Acoustics, Speech, and Signal Processing*], 1694–1697 (2010).

[10] Stamm, M. and Liu, K. J. R., "Anti-forensics of digital image compression," *IEEE Transactions on Information Forensics and Security* **6**(3), 1050–1065 (2011).

[11] Lin, W. S., Tjoa, S., Zhao, H. V., and Liu, K. J. R., "Digital image source coder forensics via intrinsic fingerprints," *IEEE Transactions on Information Forensics and Security* **4**(3), 460–475 (2009).

[12] Valenzise, G., Tagliasacchi, M., and Tubaro, S., "The cost of JPEG compression anti-forensics," in [*International Conference on Acoustics, Speech and Signal Processing*], (2011).

[13] Valenzise, G., Nobile, V., Tagliasacchi, M., and Tubaro, S., "Countering JPEG anti-forensics," in [*International Conference on Image Processing*], (2011).

[14] Lai, S. and Böhme, R., "Countering counter-forensics: The case of JPEG compression," in [*, 13th International Conference on Information Hiding*], Filler, T., Pevný, T., Craver, S., and Ker, A., eds., **6958**, 285–298, Springer-Verlag (2011).

[15] Stamm, M., Tjoa, S., Lin, W. S., and Liu, K. J. R., "Undetectable image tampering through JPEG compression anti-forensics," in [*International Conference on Image Processing*], (2010).

[16] Fan, Z. and de Queiroz, R., "Identification of bitmap compression history: JPEG detection and quantizer estimation," *IEEE Transactions on Image Processing* **12**(2), 230–235 (2003).

[17] Barni, M., Fontani, M., and Tondi, B., "A universal technique to hide traces of histogram-based image manipulations," in [*ACM Multimedia and Security Workshop*], (2012).

[18] Fan, W., Wang, K., Cayre, F., and Xiong, Z., "A variational approach to JPEG anti-forensics," in [*Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*], 3058–3062 (2013).

[19] Fridrich, J., "Digital image forensic using sensor noise," *IEEE Signal Processing Magazine* **26**(2), 26–37 (2009).

[20] Kivanc Mihcak, M., Kozintsev, I., and Ramchandran, K., "Spatially adaptive statistical modeling of wavelet image coefficients and its application to denoising," in [*Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*], **6**, 3253 –3256 vol.6 (mar 1999).

[21] Chen, M., Fridrich, J., Goljan, M., and Lukas, J., "Determining image origin and integrity using sensor noise," *IEEE Transactions on Information Forensics and Security* **3**(1), 74–90 (2008).

[22] Ferrara, P., Bianchi, T., De Rosa, A., and Piva, A., "Image forgery localization via fine-grained analysis of CFA artifacts," *IEEE Transactions on Information Forensics and Security* **7**(5), 1566 –1577 (2012).

[23] Luo, W., Huang, J., and Qiu, G., "JPEG error analysis and its applications to digital image forensics," *IEEE Transactions on Information Forensics and Security* **5**(3), 480–491 (2010).

[24] Kelley, T., [*Iterative Methods for Optimization*], SIAM (1999).