Performance Analysis of Fridrich-Goljan Self-Embedding Authentication Method

Gabriel Domínguez-Conde† *, Pedro Comesaña† and Fernando Pérez-González†

Abstract

This paper analyzes the performance of the image authentication method based on robust hashing proposed by J. Fridrich and M. Goljan. In this method both the embedder and the detector generate the watermark from a perceptual digest of the image. Therefore, an accurate performance analysis requires assessing the relation between noise and hash bit errors. Our approach first derives the probability of hash bit error due to watermark embedding and/or the attack, and then uses such probability to derive the probabilities of false positive and false negative.

Index Terms

Content-based authentication, robust hash, performance analysis.

I. Introduction

In the last years, a number of powerful and user-friendly multimedia editing tools for digital video, image and audio have proliferated, allowing an unskilled person to easily modify contents and produce forgeries. To regain confidence on digital objects, several authentication tools have been proposed, some of them relying on digital watermarking techniques. In watermarking-based authentication techniques a low-power signal (watermark) is embedded in the digital content to be protected (host signal); then, the authenticity of the received signal is determined by verifying the presence of the correct watermark. Concerning the generation of such watermark, sometimes it is just pseudorandomly produced using a secret key shared by the embedder and the detector, whereas in other cases a perceptual digest of the digital content is also used. This low-dimensional summary of a digital content is usually referred to as a *robust hash*, *perceptual hash* or *soft hash*.

In contrast to a cryptographic hash, a robust hash would be ideally sensitive only to perceptual changes, meaning that two perceptually identical digital objects should yield the same result. However, this requirement is difficult to be fully met in practice, as the design of a tractable mathematical model of human perception that accurately quantifies perceptual similarity is still an unsolved task. On the other hand, given that only authorized users should be able to generate a valid robust hash, the chosen hash functions usually depend on a secret key.

So far, several authentication techniques based on watermarking have been presented, most of them aimed at images. Schneider et al. [1] proposed, based on the ideas of Friedman [2], one of the earliest methods which uses a robust hash to aid the verification of the authenticity of digital contents. Bhattacharjee and Kutter [3], proposed a robust hashing authentication scheme by generating a set of feature-points with a set of Mexican-Hat wavelets. In [4] Kundur and Hatzinakos described a fragile method for tamper localization by using a quantization technique to embed the watermark in a transform domain. Venkatesan

† Dept. Teoría do Sinal e Comunicacións, ETSE Telecom., Universidade de Vigo, 36310 Vigo, Spain. Phone: +34 986 812683. Fax: +34 986 812116. E-mails: gdomin@gts.tsc.uvigo.es, pcomesan@gts.tsc.uvigo.es and fperez@gts.tsc.uvigo.es

* Corresponding author. E-mail: gdomin@gts.tsc.uvigo.es

This work was supported in part by *Xunta de Galicia* under Projects 07TIC012322PR (FACTICA), 2007/149 (REGACOM), 2006/150 ("Consolidation of Research Units"), and by the Spanish Ministry of Science and Innovation under projects COMONSENS (ref. CSD2008-00010) of the CONSOLIDER-INGENIO 2010 Program and the SPROACTIVE (ref. TEC2007-68094-C02-01/TCM).

et al. [5] developed a robust hashing method which divides an image into random non-overlapping blocks. The statistics of each block are used to generate the robust hash. Lin and Chang [6] introduced a robust authentication watermarking scheme, designed for JPEG compression, based on the relation between the discrete cosine transform (DCT) coefficients. In [7], Cannons and Moulin proposed a non-blind robust authentication method based on robust hashing and watermarking, and used a statistical description of the host image to analyze its performance. More recently, Swaminathan et al. have proposed a robust hashing method for image authentication, based on invariant characteristics of the images, e.g., the resilience to RST (rotation, scaling and translation) of Fourier-Mellin transform [8]. Finally, in [9] Monga and Mıhçak have presented a robust hashing technique where a content-based binary vector is extracted by applying a Non-Negative Matrix Factorization (NMF) to the image.

From the set of authentication techniques which embed a watermark that depends on an image robust hash (also known as watermarking self-embedding authentication), the algorithm proposed by Fridrich and Goljan [10]–[12] is one of the most prominent, and has been extensively adopted as a reference in many other works and comparisons (e.g., [7], [8], [13]). In addition, this method uses a watermark synthesis function which successfully fills the gap between robust hashing and watermarking-based authentication; in fact, the watermark generation has the particular feature that it can be tuned to account for the error sensitivity of the overall authentication technique.

However, a performance analysis of this widely-referenced self-embedding authentication algorithm is still lacking. This means that for comparison purposes Monte Carlo techniques need be used, with the obvious drawback of requiring long simulation runs whenever small probabilities are to be estimated. In addition, this lack makes it very difficult to determine the influence of the various parameters of the Fridrich-Goljan method on performance. This paper aims at filling this gap by presenting a novel performance analysis and drawing some conclusions that can be extended to general hash-based authentication schemes.

In our analysis, we first compute the hash bit error probability due to the watermark embedding and noise, and then we show its impact on the Receiver Operating Characteristic (ROC) of the overall scheme. Notice that this methodology could be used to analyze other authentication schemes based on robust hashing, as the same principles would apply. To the best of our knowledge, this is the first time that the performance of a watermarking authentication scheme based on robust hashing is analyzed in this way.

In the next section we give a brief introduction to the robust authentication method proposed by Fridrich and Goljan, and the embedding and detection processes, whereas performance is addressed in Sect. III. In Sect. IV simulations are carried out to validate our approach; in addition, the modifications on the introduced performance analysis necessary for dealing with image compression, intensity change and linear filtering are outlined. Finally, Sect. V presents the main conclusions and discusses some future lines.

A. Notation

We will denote scalar random variables with capital letters (e.g., X) and their outcomes with lowercase letters (e.g. x). The same notation criterion applies to random vectors and their outcomes, denoted in this case by bold letters (e.g. $\mathbf{X}$, $\mathbf{x}$), with transposes denoted by the superindex T. The ith component of a vector $\mathbf{X}$ is denoted as X_i. Images in the pixel domain will be partitioned in N_b blocks and arranged as vectors.

II. Description of Fridrich and Goljan Method

In this section a description of the method by Fridrich and Goljan [11], that will be analyzed in Sect. III, is given; furthermore, a correlation-based detector for that method is proposed.

A. Hash Computation

The original host signal in the pixel domain is block-wise partitioned and arranged as N_b vectors $\mathbf{x}^i$, $1 \leq i \leq N_b$, each of size M.[1] For the sake of notational simplicity, we will avoid the block superindex; however, it must be clear that the authentication method operates at the block level. From each $\mathbf{x}$, the variance of which will be denoted by σ_X^2, and depending on a set of N_h length-M pseudo-random sequences $\mathbf{s}^j$ (generated irrespectively of $\mathbf{x}$), $1 \leq j \leq N_h$, an N_h bits hash vector $\mathbf{h}$ is computed as

$$h_j = \begin{cases} 0 & \text{if } \frac{1}{M}|\mathbf{x}^T \cdot \mathbf{s}^j| < T_e \\ 1 & \text{otherwise} \end{cases},$$

where T_e is a quantization threshold (constant along the complete image) computed to comply with the constraint that the total number of 0's over all the N_b hash vectors $\mathbf{h}$ of the image must be equal to the total number of 1's. In our analysis this threshold will be approximated by the median of the absolute value of the coefficients obtained by projecting the host image blocks onto the pseudorandom sequences, i.e. $\frac{1}{M}\mathbf{X}^T \cdot \mathbf{S}^j$. Noticing that the method operates at the block level, for our purposes we will rely on a block-wise characterization of the host image. Therefore, for the derivations contained in this paper the probability density function (pdf) of the absolute value of the projected host is obtained by averaging over the projected host blocks. Using this approximation, the number of zeros and ones cannot be guaranteed to be equal for every hash originated from a given image; however, their relative frequency will be asymptotically identical as $N_b \cdot N_h$ is increased.

Concerning the generation of the projection sequences $\mathbf{s}^j$, each of them is produced from a $\sqrt{M} \times \sqrt{M}$ pseudo-random matrix (obtained depending on the system secret key), whose components take values uniformly in $[0, 1]$. Each of these matrices is then low-pass filtered, mean adjusted, and rearranged as the length-M vector $\mathbf{s}^j$.

B. Watermark Computation

Each hash vector $\mathbf{h}$ is permuted using N_p permutations $\boldsymbol{\pi}^k(\cdot)$, $\boldsymbol{\pi}^k : \{0,1\}^{N_h} \rightarrow \{0,1\}^{N_h}$, with $k = 1, \cdots, N_p$. Next, the results are joined to define the length-N_p vectors $\mathbf{t}^l \triangleq (\pi_l^1(\mathbf{h}), \pi_l^2(\mathbf{h}), \cdots, \pi_l^{N_p}(\mathbf{h}))$, $l = 1, \cdots, N_h$. These $\mathbf{t}^l$, jointly with the secret key of the system, and the index of the current image block, are used as seed of a Pseudo-Random Number Generator (PRNG) that generates a length-M sequence with components uniformly distributed on $[-1, +1]$, and that we will denote by $\mathbf{v}^l$. Finally, the watermark $\mathbf{w}$ corresponding to a block of the original host signal $\mathbf{x}$, is constructed as $\mathbf{w} = \sqrt{\frac{3}{N_h}} \sum_{l=1}^{N_h} \mathbf{v}^l$. Assuming a good behavior of the PRNG, it is reasonable to think of the $\mathbf{v}^l$ as being (almost) independent of $\mathbf{x}$ and $\mathbf{s}^j$ (both through $\mathbf{t}^l$), so $\mathbf{w}$ can be considered to be independent of $\mathbf{x}$ and $\mathbf{s}^j$ as well. Furthermore, given that different seeds are used for generating the vectors $\mathbf{v}^l$, it follows that $\mathbf{v}^l$ will be (almost) independent from each other; hence, for values of N_h large enough, one can use the Central Limit Theorem (CLT), and approximate the distribution of $\mathbf{W}$ by $\mathcal{N}(\mathbf{0}, \mathbf{I}_{M \times M})$, with $\mathbf{I}_{M \times M}$ the identity matrix of size M.

This watermark $\mathbf{w}$ is embedded in the host signal using Additive Spread Spectrum [14] in the $\sqrt{M} \times \sqrt{M}$-block pixel domain, so a block of the watermarked image is obtained as $\mathbf{y} = \mathbf{x} + \gamma\mathbf{w}$, where γ is an embedding strength parameter.

C. Detection

On the detector side, the steps described above are followed to obtain an estimate $\hat{\mathbf{w}}$ of the watermark $\mathbf{w}$ from a block of the received signal $\mathbf{z}$. Be aware that even in the absence of attacks, the considered signal $\mathbf{z}$ will be different from the host image (due to the presence of the watermark), so the quantization threshold at the detector will be computed from $\frac{1}{M}\mathbf{Z}^T \cdot \mathbf{S}^j$, and thus

[1] Following the original description by Fridrich and Goljan [11], these blocks correspond to non-overlapping $\sqrt{M} \times \sqrt{M}$-pixel blocks.

it will not necessarily coincide with that obtained at the embedder. In order to make this difference explicit, we will denote the threshold calculated at the detector as T_d. On the other hand, the same considerations made regarding the computation of T_e in Sect. II-A can be raised here.

The main objective of the detector is to decide on the presence of the watermark estimate in the corresponding block of the received signal; if the estimate of the watermark is decided to be present, that block will be declared authentic (i.e., non-modified); otherwise, it will be said to be manipulated.

Since in [11] detection is not addressed, we introduce next the detection process assumed in our analysis. First, the decision on the presence or absence of the estimate of the watermark can be formulated as a binary hypothesis test, namely,

$$\begin{aligned} \mathcal{H}_0 \quad &: \quad \mathbf{z} = \eta\left(\mathbf{x} + \gamma\hat{\mathbf{w}}\right) + \mathbf{n} \\ \mathcal{H}_1 \quad &: \quad \mathbf{z} = \mathbf{x} + \gamma\hat{\mathbf{w}}, \end{aligned} \tag{1}$$

where $\mathcal{H}_0$ represents the hypothesis of the received signal being the sum of a watermarked signal scaled by a given factor $\eta \in [0, 1]$ and some complementary signal $\mathbf{n}$ independent of $\mathbf{x}$, $\mathbf{s}$ and $\mathbf{w}$, with zero mean and variance σ_N^2, whereas $\mathcal{H}_1$ denotes the hypothesis of the received signal being the output of the embedder. This hypothesis test comprises several interesting detection scenarios. For example, by setting $\eta = 0$, the proposed hypothesis test can model the case of deciding whether a block was watermarked with a valid key or it was not watermarked (or watermarked with an invalid key). On the other hand, with $\eta = 1$, $\mathcal{H}_0$ corresponds to the case where the image under test is the result of applying some unacceptable noise/processing (modeled by the addition of $\mathbf{n}$) to an otherwise valid watermarked image. Moreover, other values of η in $(0, 1)$, may model other scenarios, e.g., scaling attacks, image fusion, etc.

Additionally, it is worth pointing out that the watermark detection problem (a.k.a. one-bit and zero-bit watermarking), although out of the scope of this paper, can be also studied in the framework defined by the hypothesis test in 1, just by setting $\eta = 0$ and interpreting $\mathbf{n}$ as a non-watermarked content.

When detection mistake costs are not set, or a priori probabilities for the two hypotheses are not available (as is the case in most practical scenarios), the Neyman-Pearson criterion is customarily used, as it minimizes the probability of false negative for a given probability of false positive [15]. This criterion implies the use of the likelihood ratio test, taking the form

$$L(\mathbf{z}) = \frac{f(\mathbf{z}|\mathcal{H}_1)}{f(\mathbf{z}|\mathcal{H}_0)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \lambda,$$

where λ is the detection threshold. When both the host signal and the noise are independent and Gaussian distributed, the correlation between the received block and the corresponding watermark, i.e., $\rho \triangleq \frac{1}{M}\mathbf{z}^T \cdot \hat{\mathbf{w}}$, is a sufficient statistic for this problem. In most practical detection methods, and due to its simplicity, this statistic is still used, although the mentioned condition on the Gaussianity and independence of the signals is not verified. Given a false positive probability P_{fp} (the probability of deciding that the received signal was not modified, when indeed it was), λ is selected so that the following equation holds

$$P_{fp} = \int_{R(\lambda)} f_{\rho|\mathcal{H}_0}(\tau) d\tau, \tag{2}$$

and the false negative (the probability of deciding that the received signal was modified, when it was not) is then calculated as

$$P_{fn} = \int_{\bar{R}(\lambda)} f_{\rho|\mathcal{H}_1}(\tau) d\tau, \tag{3}$$

where in the above expressions (2) and (3), $R(\lambda) \triangleq \{\mathbf{x}|L(\mathbf{x}) > \lambda\}$, $\bar{R}(\lambda)$ is the complement of $R(\lambda)$, and $f_{\rho_{\mathcal{H}_0}}$ and $f_{\rho_{\mathcal{H}_1}}$ denote the pdf's of ρ when respectively $\mathcal{H}_0$ and $\mathcal{H}_1$ are true (namely, $\rho_{\mathcal{H}_0}$ and $\rho_{\mathcal{H}_1}$). For the computation of $f_{\rho_{\mathcal{H}_0}}$ and $f_{\rho_{\mathcal{H}_1}}$ we will take into account that ρ is the result of adding $N_h \cdot M$ random variables with finite variance, so for large values of both N_h and M, and whenever $\sigma_X^2 >> \sigma_N^2$ and $\sigma_X^2 >> \gamma^2$, we can model both $f_{\rho_{\mathcal{H}_0}}$ and $f_{\rho_{\mathcal{H}_1}}$ as Gaussian distributions (see App. B for further details); therefore, the detection thresholds on ρ will be given by the solutions to the following equation (see Chapter 2.6 [15])

$$\frac{(\rho - \mathrm{E}\{\rho_{\mathcal{H}_0}\})^2}{2\mathrm{Var}\{\rho_{\mathcal{H}_0}\}} - \frac{(\rho - \mathrm{E}\{\rho_{\mathcal{H}_1}\})^2}{2\mathrm{Var}\{\rho_{\mathcal{H}_1}\}} + \log\left(\sqrt{\frac{\mathrm{Var}\{\rho_{\mathcal{H}_0}\}}{\mathrm{Var}\{\rho_{\mathcal{H}_1}\}}}\right) = \log\lambda, \tag{4}$$

where $\mathrm{E}\{\cdot\}$ denotes expectation, and $\mathrm{Var}\{\cdot\}$ variance. Given that the last equation has in general two solutions in ρ, the detection region is defined by two thresholds, T_0 and T_1 (with $T_0 < T_1$), depending the chosen hypothesis on the ratio $\mathrm{Var}\{\rho_{\mathcal{H}_0}\}/\mathrm{Var}\{\rho_{\mathcal{H}_1}\}$; if that ratio is larger than 1, $\mathcal{H}_1$ holds for $\rho \in [T_0, T_1]$, and $\mathcal{H}_0$ elsewhere (conversely, if the ratio is smaller than 1, the hypothesis choice is the reverse). In the special case where the variances of both Gaussians are the same, (4) becomes a linear equation, and a single solution (one detection threshold) exists. Although in our performance analysis both T_0 and T_1 are used, the analytical expressions will be still valid when $\mathrm{Var}\{\rho_{\mathcal{H}_0}\} = \mathrm{Var}\{\rho_{\mathcal{H}_1}\}$ by setting $T_1 = \infty$.

In order to quantify the performance of the method proposed by Fridrich and Goljan, we will derive its Receiver Operating Characteristic (ROC). This is equivalent to evaluating both the probability of false positive P_{fp} and the corresponding probability of false negative P_{fn} for a range of values of λ; this derivation is the objective of Sect. III.

III. Performance Analysis Considering the Watermark Embedding

$$\begin{aligned} P_e \approx \frac{2}{\eta}\int_{-\infty}^{\infty} A_N e^{-|\beta_N t|^{c_N}} \Bigg[& \frac{\int_{\eta T_e}^{\infty} A_X e^{-|\frac{\beta_X}{\eta}\tau|^{c_X}} \left(1 - \mathcal{Q}\left(\frac{\sqrt{M}(T_d-\tau-t)}{\sqrt{\gamma^2\eta^2\sigma_S^2}}\right) - \mathcal{Q}\left(\frac{\sqrt{M}(T_d+\tau+t)}{\sqrt{\gamma^2\eta^2\sigma_S^2}}\right)\right) d\tau}{\frac{4}{\eta}\int_{\eta T_e}^{\infty} A_X e^{-|\frac{\beta_X}{\eta}\tau|^{c_X}} d\tau} \\ & + \frac{\int_{0}^{\eta T_e} A_X e^{-|\frac{\beta_X}{\eta}\tau|^{c_X}} \left(\mathcal{Q}\left(\frac{\sqrt{M}(T_d-\tau-t)}{\sqrt{\gamma^2\eta^2\sigma_S^2}}\right) + \mathcal{Q}\left(\frac{\sqrt{M}(T_d+\tau+t)}{\sqrt{\gamma^2\eta^2\sigma_S^2}}\right)\right) d\tau}{\frac{4}{\eta}\int_{0}^{\eta T_e} A_X e^{-|\frac{\beta_X}{\eta}\tau|^{c_X}} d\tau} \Bigg] dt. \end{aligned} \tag{5}$$

$$\begin{aligned} P_e \approx \frac{2}{\eta}\Bigg[& \int_{\eta T_e}^{\infty} A_X e^{-|\frac{\beta_X}{\eta}\tau|^{c_X}} \left(1 - \mathcal{Q}\left(\frac{\sqrt{M}(T_d-\tau)}{\sqrt{\gamma^2\eta^2\sigma_S^2+\sigma_S^2\sigma_N^2}}\right) - \mathcal{Q}\left(\frac{\sqrt{M}(T_d+\tau)}{\sqrt{\gamma^2\eta^2\sigma_S^2+\sigma_S^2\sigma_N^2}}\right)\right) d\tau \\ & + \int_{0}^{\eta T_e} A_X e^{-|\frac{\beta_X}{\eta}\tau|^{c_X}} \left(\mathcal{Q}\left(\frac{\sqrt{M}(T_d-\tau)}{\sqrt{\gamma^2\eta^2\sigma_S^2+\sigma_S^2\sigma_N^2}}\right) + \mathcal{Q}\left(\frac{\sqrt{M}(T_d+\tau)}{\sqrt{\gamma^2\eta^2\sigma_S^2+\sigma_S^2\sigma_N^2}}\right)\right) d\tau \Bigg]. \end{aligned} \tag{6}$$

$$P_{fp} \approx \sum_{n_s=0}^{N_h} Pr(N_s = n_s|\mathcal{H}_0) \left(\mathcal{Q}\left(\frac{\sqrt{M}\xi(T_0 - (N_h - n_s)\frac{\gamma\eta}{N_h})}{\sqrt{\eta^2\sigma_X^2 + \gamma^2\eta^2\sigma_{n_s}^2 + \sigma_N^2}}\right) - \xi\mathcal{Q}\left(\frac{\sqrt{M}(T_1 - (N_h - n_s)\frac{\gamma\eta}{N_h})}{\sqrt{\eta^2\sigma_X^2 + \gamma^2\eta^2\sigma_{n_s}^2 + \sigma_N^2}}\right)\right). \tag{7}$$

$$P_{fn} \approx \sum_{n_s=0}^{N_h} Pr(N_s = n_s|\mathcal{H}_1) \left(\mathcal{Q}\left(\frac{\sqrt{M}\xi((N_h - n_s)\frac{\gamma}{N_h} - T_0)}{\sqrt{\sigma_X^2 + \gamma^2\sigma_{n_s}^2}}\right) + \xi\mathcal{Q}\left(\frac{\sqrt{M}(T_1 - (N_h - n_s)\frac{\gamma}{N_h})}{\sqrt{\sigma_X^2 + \gamma^2\sigma_{n_s}^2}}\right)\right). \tag{8}$$

In this section we will analyze the effect of the watermark embedding and the attack on the estimate of the hash on the detector side, and how a non-perfect estimate of the watermark will deteriorate the overall performance. Our first step will be the characterization of the random variable $D_j \triangleq \frac{1}{M}\mathbf{X}^T \cdot \mathbf{S}^j$. Reasoning that projecting onto $\mathbf{s}^j$ resembles computing an almost orthogonal transform somewhat similar to the DCT, whose coefficients have been previously characterized in the literature by a Generalized Gaussian Distribution (GGD) [16], we propose to model D_j by a GGD, i.e., $f_{D_j}(x) \approx A_X e^{-|\beta_X x|^{c_X}}$, where in the above expression A_X, β_X and the shaping parameter c_X are fitted for each block of the image to the experimental data using Maximum Likelihood Estimation (MLE). This crucial hypothesis has been validated using the luminance component of a set of 100 images. This set was built by randomly selecting 100 images from the Uncompressed Colour Image Database of the Austin University [17], resizing the chosen images to 256×256 pixels. For each image, the Kullback-Leibler divergence (KLD) between the histogram obtained by projecting its blocks of size 64×64 over 10000 random vectors and the corresponding GGD with parameters optimized for those projections has been computed. The mean value of the KLD over the 100 images is as small as $2.5 \cdot 10^{-3}$. For the sake of comparison we have also obtained the KLD between the histogram of the projected coefficients for each image and a Gaussian distribution with zero mean and variance empirically estimated from those coefficients; in this case the mean value of KLD is $3.1 \cdot 10^{-2}$, i.e., an order of magnitude difference between the two KLD's which supports the use of a GGD.

In order to derive the probability of the errors produced by watermark embedding, the probability of flipping one bit of the hash obtained at the detector with respect to the robust hash computed at the embedder, has to be calculated. As stated above, for large N_h the components of the watermark can be well-approximated by a $\mathcal{N}(0, \mathbf{I}_{M\times M})$; furthermore, $\mathbf{W}$ and $\mathbf{S}$ can be considered as independent, as stated above. Therefore, for the characterization of $\frac{\gamma\eta}{M}\mathbf{W}^T \cdot \mathbf{S}^j$ we will take into account that for an arbitrary value of the vector $\mathbf{s}^j$, $\frac{1}{M}\mathbf{W}^T \cdot \mathbf{s}^j$ follows a $\mathcal{N}\left(0, \frac{||\mathbf{S}^j||^2}{M^2}\right)$, so in order to derive the pdf of $\frac{1}{M}\mathbf{W}^T \cdot \mathbf{S}^j$ one should average the obtained Gaussian distribution over the possible values of $\frac{||\mathbf{S}^j||^2}{M^2}$. Thus, given that $\mathbf{S}^j$ is obtained by processing an i.i.d. random vector with a low-pass filter and subtracting the mean, due to the Law of Large Numbers (see Chapter VII, [18]) for large values of M, the distribution of $\frac{||\mathbf{S}^j||^2}{M}$ will converge to $\sigma_Q^2 \cdot ||\mathbf{g}||^2$, where $\mathbf{g}$ denotes the coefficients of the mentioned filter and σ_Q^2 is the variance of the original i.i.d. signal.[2] Let $\sigma_S^2 \triangleq \sigma_Q^2 \cdot ||\mathbf{g}||^2$, then we can approximate the projected watermark $\frac{\gamma\eta}{M}\mathbf{W}^T \cdot \mathbf{S}^j$ by $\mathcal{N}(0, \frac{\gamma^2\eta^2\sigma_S^2}{M})$ when M is large. On the other hand, the projection of the complementary signal $\mathbf{N}$ onto $\mathbf{S}^j$, i.e. $\frac{1}{M}\mathbf{N}^T \cdot \mathbf{S}^j$, will be modeled by a GGD with parameters A_N, β_N and c_N. This characterization is valid for both the cases where $\mathbf{n}$ is an image (since, as we have discussed at the beginning of this section, its projection will be well-modeled by a GGD) or Gaussian additive noise (for which $c_N = 2$).

Thus, as shown in App. A, the probability of a hash bit error under hypothesis $\mathcal{H}_0$ can be expressed as (5), where $\mathcal{Q}(x) \triangleq \frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-\frac{\tau^2}{2}}d\tau$. It is worth pointing out that (5) is valid whenever $\eta > 0$; in the particular case where $\eta = 0$, due to the assumption of independence between $\mathbf{X}$ and $\mathbf{N}$, it is clear that $P_e \approx 1/2$. Additionally, (5) can be easily adapted to hypothesis $\mathcal{H}_1$ by setting $\eta = 1$ and $\mathbf{N}^T\mathbf{S}^j = 0$ ($\beta_N = \infty$).

Expression (5) admits further simplifications under certain circumstances. First, recalling that the threshold T_e is set so that it is exceeded half of the time, the denominators of the two summands in (5) will be approximately 1 when the statistics of the different projected blocks are similar. Moreover, when $\mathbf{N}$ corresponds to Gaussian noise, the outer integral in (5) can be explicitly solved. To this end, we will focus on the first summand in (5), since a similar derivation applies to the second term.

[2]Border effects were not considered in this derivation due to their marginal consequences.

Recalling that $\mathbf{N}^T\mathbf{S}^j \sim \mathcal{N}(0, \frac{\sigma_S^2\sigma_N^2}{M})$, we can write the numerator of the chosen term as

$$\begin{aligned}
&\frac{2}{\eta}\int_{\eta T_e}^{\infty} A_X e^{-|\frac{\beta_X}{\eta}\tau|^{c_X}} \int_{-\infty}^{\infty}\int_{-T_d+\tau+t}^{T_d+\tau+t} \\
&\frac{\sqrt{M}}{\sqrt{2\pi\sigma_S^2\sigma_N^2}} e^{\frac{-Mt^2}{2\sigma_S^2\sigma_N^2}} \frac{\sqrt{M}}{\sqrt{2\pi\gamma^2\eta^2\sigma_S^2}} e^{\frac{-Mx_1^2}{2\gamma^2\eta^2\sigma_S^2}} dx_1 dt d\tau \\
&= \frac{2}{\eta}\int_{\eta T_e}^{\infty} A_X e^{-|\frac{\beta_X}{\eta}\tau|^{c_X}} \int_{-T_d+\tau}^{T_d+\tau}\int_{-\infty}^{\infty} \\
&\frac{\sqrt{M}}{\sqrt{2\pi\sigma_S^2\sigma_N^2}} e^{\frac{-Mt^2}{2\sigma_S^2\sigma_N^2}} \frac{\sqrt{M}}{\sqrt{2\pi\gamma^2\eta^2\sigma_S^2}} e^{\frac{-M(x_2-t)^2}{2\gamma^2\eta^2\sigma_S^2}} dx_2 dt d\tau, \qquad (9)
\end{aligned}$$

Given that the inner integral in (9) is the convolution of $\mathcal{N}(0, \frac{\sigma_S^2\sigma_N^2}{M})$ and $\mathcal{N}(0, \frac{\gamma^2\eta^2\sigma_S^2}{M})$, we can rewrite (9) as

$$\begin{aligned}
&\frac{2}{\eta}\int_{\eta T_e}^{\infty} A_X e^{-|\frac{\beta_X}{\eta}\tau|^{c_X}} \left(1 - \mathcal{Q}\left(\frac{\sqrt{M}(T_d-\tau)}{\sqrt{\gamma^2\eta^2\sigma_S^2+\sigma_S^2\sigma_N^2}}\right)\right. \\
&\left. -\mathcal{Q}\left(\frac{\sqrt{M}(T_d+\tau)}{\sqrt{\gamma^2\eta^2\sigma_S^2+\sigma_S^2\sigma_N^2}}\right)\right) d\tau.
\end{aligned}$$

Summarizing, when $\mathbf{N}$ is Gaussian-distributed, and the projected host blocks have similar statistics, P_e becomes (6).

Once we have obtained the probability of error of each hash bit, we want to relate this quantity to the errors made in the estimate of the watermark. As it was described in Sect. II-B, the estimated watermark $\hat{\mathbf{w}}$ is generated from N_p permutations of the reconstructed hash vector $\hat{\mathbf{h}}$; one bit of each of these permutations is picked to form the vector $\mathbf{t}^l$, $1 \leq l \leq N_h$. Thus, N_e errors in the estimate of the hash vector, with $N_e \leq N_h$, will be spread to at most $\min\{N_e \cdot N_p, N_h\}$ different vectors $\mathbf{t}^l$. This implies that the correlation between $\mathbf{w}$ and $\hat{\mathbf{w}}$ for a given block will depend, through the generation of $\mathbf{v}^l$, on the number of wrong vectors $\mathbf{t}^l$, that we will denote by N_s. Hence, it is necessary to know the probability that the number of wrong vectors $\mathbf{t}^l$ is n_s, when there are n_e bit errors in the estimate of the hash. In App. B it is shown that the values of the probability of false positive and false negative are given by (7) and (8), respectively. In both expressions, $Pr(N_s = n_s|\mathcal{H}_0)$ and $Pr(N_s = n_s|\mathcal{H}_1)$ (which are derived in App. B) denote the probability that the number of wrong vectors $\mathbf{t}^l$ is n_s under hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively, and when $\mathrm{Var}\{\rho_{\mathcal{H}_0}\} \neq \mathrm{Var}\{\rho_{\mathcal{H}_1}\}$, $\xi = \mathrm{sign}\left(\mathrm{Var}\{\rho_{\mathcal{H}_0}\} - \mathrm{Var}\{\rho_{\mathcal{H}_1}\}\right)$; for the case where $\mathrm{Var}\{\rho_{\mathcal{H}_0}\} = \mathrm{Var}\{\rho_{\mathcal{H}_1}\}$ the obtained expressions are still valid by making $T_1 = \infty$ and $\xi = 1$. In addition, σ_{n_s} represents the standard deviation of the projection of the original watermark $\mathbf{w}$ onto $\hat{\mathbf{w}}$ computed at the detector when the number of wrong vectors $\mathbf{t}^l$ is n_s.

IV. Experimental Results

Next we present the results of several experiments conducted on the set of images described above. First, we experimentally verify the validity of our model. In order to do so, in our first experiment we study the scenario where the detector must decide whether a given image bears the right watermark. In this setup, the null hypothesis $\mathcal{H}_0$ is particularized to $\eta = 0$ and $\mathbf{n}$ is the block of a non-watermarked image. The aforementioned set of 100 images was used, with block size of 64×64 pixels, $N_h = 16$ and $N_p = 5$. The results are plotted in Fig. 1, where the empirical and analytical ROC curves almost perfectly match. Furthermore, the curves for different values of γ ($\gamma \in \{2, 4, 8, 10\}$), show that a better performance, in terms of the ROC, is achieved with larger values of γ, although one should also consider that a larger γ also implies a larger distortion. Hence, in this case a trade-off between distortion and performance should be achieved. However, the reasoning "larger distortion implies better performance", is not always verified. For example, if one considers the case $\eta = 1$, and a fixed distribution of $\mathbf{N}$, an

increase of γ reduces the detection probability for a given probability of false positive, as the effect of $\mathbf{N}$, which in this scenario is the part of the received signal that helps us to decide what hypothesis holds, is masked by the watermark.

In the second considered scenario we analyze the performance of the Fridrich-Goljan authentication method when trying to detect luminance scaling attacks. Fig. 2 compares the analytical and empirical ROC curves for $\eta \in \{0.3, 0.5, 0.7\}$ and $\gamma = 10$ with the same parameters of the previous experiment (i.e., blocks of 64×64 pixels, $N_h = 16$ and $N_p = 5$) for Lena without additive noise, i.e., $\mathbf{n} = \mathbf{0}$. The close resemblance between the analytical and empirical results shows again the goodness of the proposed model and the subsequent analysis.

It is worth discussing the role of the number N_p of permutations of $\mathbf{h}$ in the generation of $\mathbf{w}$. This parameter is used to control the performance degradation due to image modification, as it was outlined at the end of the Sect. III. On one hand, it is important to avoid that a convex combination of a block from an authentic signal and another from a non-authentic signal be used to produce a forgery. In that case, we would be interested in having a large value of N_p, producing a sharp degradation of the correlation statistic, and therefore implying a reduction in the false positive probability. Nevertheless, we are also interested

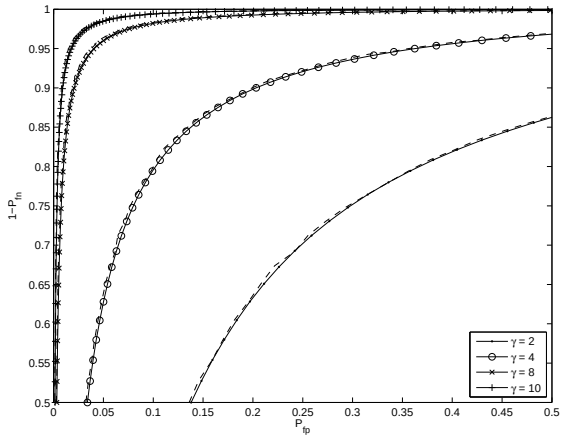

Figure 1. Analytical and empirical ROC curves for the set of 100 images. $M = 4096$, $N_h = 16$, $N_p = 5$, $\eta = 0$, $\gamma \in \{2, 4, 8, 10\}$, and $\mathbf{n} = \mathbf{x}$. The solid and the dashed lines correspond with the analytical and experimental curves, respectively.

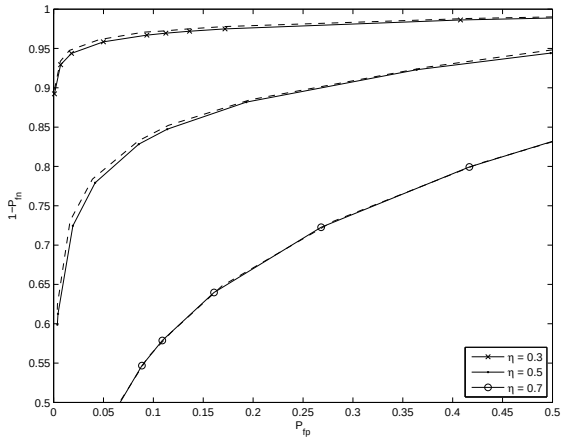

Figure 2. ROC curves with $\eta = \{0.3, 0.5, 0.7\}$ without additive noise for Lena. Here, $M = 4096$, $N_h = 16$, $N_p = 5$ and $\gamma = 10$. The solid lines correspond with the analytical curves and the dashed lines with the empirical curves.

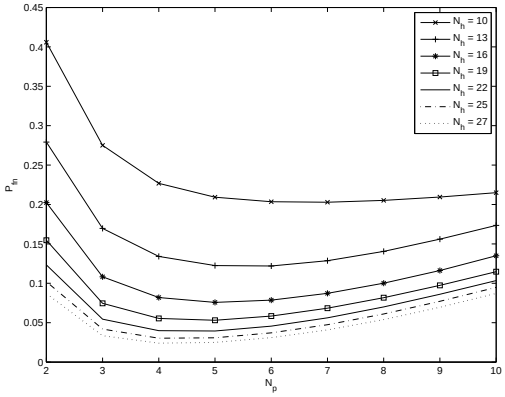

Figure 3. P_{fn} vs N_p for Lena for different values of N_h with $P_{fp} = 0.005$, $M = 4096$, $\eta = 0$ and $\gamma = 15$.

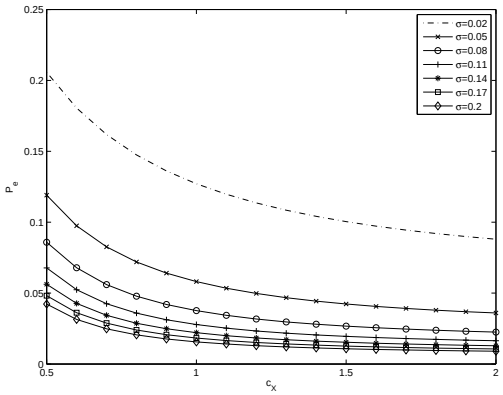

Figure 4. P_e vs shaping factor c_X curves for different standard deviations σ of the GGD which models the projection of image block $\mathbf{x}$ onto the pseudorandom patterns (i.e. $\sigma = \frac{1}{\beta_X}\sqrt{\frac{\Gamma(3/c_X)}{\Gamma(1/c_X)}}$) when $\mathcal{H}_1$ holds using the following parameters: $M = 4096$ and $\gamma = 10$.

in having a smooth degradation of the correlation statistic with the number of mistakes in the hash, as the authentication scheme should be robust to slight modifications, such as transcoding or the watermark embedding itself. In that sense, the smaller the value of N_p, the smoother the performance degradation, and therefore the more robust the authentication scheme. As a conclusion, we can establish that a trade-off exists between robustness and probability of false positive.

In order to illustrate the dependence of the probability of false negative on N_p, in Fig. 3 the behavior of P_{fn} is plotted as a function of N_p for different values of N_h for Lena (with $M = 4096$, $\eta = 0$ and $\gamma = 15$), after fixing the probability of false positive to $P_{fp} = 0.005$. In the particular case studied in Fig. 3, the value of N_p which minimizes P_{fn} just slightly depends on N_h. Notice that if $\mathbf{z}$ is not watermarked or the detector and the embedder use different keys then $P_e \approx 1/2$. Therefore, if we increase N_p, the error probability in the seed used to generate $\mathbf{v}^l$ will rise and consequently P_{fn} will also be larger.

The dependence of P_e on the parameters of the GGD used for modeling $\mathbf{X}^T \cdot \mathbf{S}^j$ when $\mathcal{H}_1$ holds is illustrated in Fig. 4, where P_e is plotted as a function of the shaping factor c_X for different standard deviations of the projected blocks σ_X, $M = 4096$ and $\gamma = 10$. The considered range of σ_X corresponds to typical values of the standard deviation obtained for the projected coefficients of real images. On one hand, and according to intuition, it can be seen that for larger values of the Document to Watermark Ratio (proportional to $\frac{\sigma_X^2}{\gamma^2}$) the probability that the watermark embedding flips bits of the original robust hash will be smaller. On the other hand, the results on Fig. 4 show that P_e decreases with c_X.

Finally, in order to identify the limitations of the analysis presented above, we will show how it can be adapted to deal with typical image processing attacks such as JPEG compression, intensity transformations and linear filtering.

The JPEG standard compresses an image by quantizing its 8×8 block-DCT coefficients so that the step-size of the used uniform scalar quantizer depends on both the desired JPEG quality factor (QF) and the frequency of each particular coefficient (quantization step-sizes for low-frequency coefficients are usually smaller than high-frequency ones). Our analytical framework can be adapted to deal with large QF JPEG compression by using a fine quantization approximation, i.e., by modeling the effect of the quantization error in the 8×8 block-DCT with a noise random variable independent of the to-be-quantized coefficient and uniformly distributed in $[-\Delta_i/2, \Delta_i/2]$, where Δ_i is the step-size of the i$-th$ coefficient scalar quantizer. Considering the

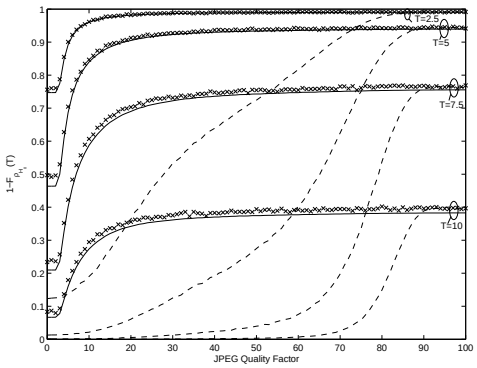

Figure 5. $1 - F_{\rho_{\mathcal{H}_0}}(T)$ vs JPEG quality factor curves using Lena for different values of T, with $N = 4096$, $N_h = 16$, $N_p = 5$, $\eta = 1$, and $\gamma = 10$. Solid lines correspond to the analytical curves, dashed lines were obtained by compressing the watermarked image with JPEG, and $\times$ symbols correspond to curves experimentally obtained by adding uniform noise in the 8×8 block-DCT domain to the watermarked signal.

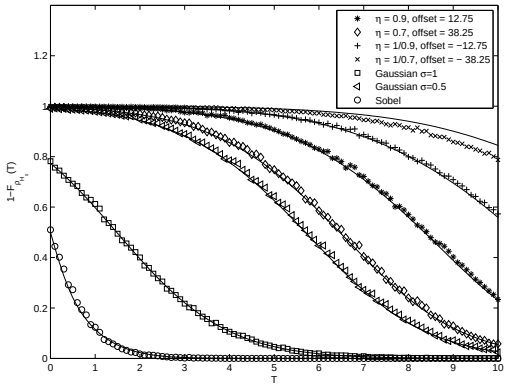

Figure 6. $1 - F_{\rho_{\mathcal{H}_0}}(T)$ corresponding to image Lena, for intensity correction and linear filtering attacks. $N = 4096$, $N_h = 16$, $N_p = 5$, and $\gamma = 10$. Solid and symbols correspond to the analytical and empirical curves, respectively.

8×8 IDCT transform, one can obtain the distribution of $\mathbf{N}$ in (1). Obviously, when the QF is decreased, the fine quantization approximation fails and our analysis can be no longer applied. This is illustrated in Fig. 5 where we compare the values of the complementary cumulative density function of $\rho_{\mathcal{H}_0}$ (i.e. $1 - F_{\rho_{\mathcal{H}_0}}(T)$) as function of the QF for the actual JPEG compression, the analytical results obtained by adapting our analysis considering fine-quantization, and the empirical results obtained when noise uniformly distributed in $[-\Delta_i/2, \Delta_i/2]$ is added to the corresponding 8×8 DCT coefficients. As expected, the plots corresponding to the two latter are very similar for the full range of QF, whereas both of them are close to the actual JPEG compression just for QF> 90, i.e. while the fine-quantization approximation holds.

Affine point-wise intensity transformations consist in scaling the luminance values of an image and centering the histogram of the resulting image by adding an offset. This processing can be modeled by changing η and $\mathbf{n}$ in (1). Analytical and empirical curves for different intensity change affine correction functions can be found in Fig. 6, showing the good match between both curves; the only plot with a significant difference between both results corresponds to $\eta = 1/0.7$, where the luminance transformation becomes non-linear due to clipping. Notice that non-linear transformations are not encompassed by by our model.

In addition, the proposed analysis can be also adapted to deal with Linear Space Invariant (LSI) image filtering, by modeling both the image block projection distribution and the distribution of the correlation between the watermark and the filtered watermarked signal. To this end, it is reasonable to neglect border effects and approximate image filtering by a 64×64-block circular filtering, so the pdf of the mentioned correlation ρ can be calculated by using the circular convolution theorem. Due to space limitations, we have chosen a low-pass (Gaussian) and a high-pass (Sobel) filters as representatives of the class of LSI filters. Concerning the Gaussian filtering, it is worth noting that the obtained P_e will be very similar to the P_e without filtering, as the Gaussian filters just slightly modify the low frequency components used to generate the robust hash. Knowing that host signal variance is concentrated in the low frequencies, and we are using a white watermark, for large document-to-watermark ratios the correlation between the Gaussian filtered watermarked signal and the reconstructed watermark will have a variance similar to the non-filtered case, whereas its mean will be reduced. On the other hand, when the image is Sobel-filtered P_e can

be approximated by 0.5, as most frequencies used to produce the robust hash are erased; as a consequence, the mean of the correlation between the Sobel filtered watermarked signal and the reconstructed watermark will be approximately null, whereas its variance will be dramatically reduced due to the low frequencies host power allocation. The behavior of both filters is shown in Fig. 6, in accordance to our previous discussion.

Finally, we would like to emphasize that, as stated above, the computation time needed to assess the performance of the analyzed self-embedding authentication algorithm using our analytical methodology is dramatically smaller than the time required by Monte Carlo techniques; as an example, a common PC with an Intel Core 2 Quad CPU at 2.4 Ghz and 4 Gb of RAM with Matlab 7.4 spends more than six days carrying out the Monte Carlo simulations needed to obtain the curves of Fig. 1. In contrast, less than five minutes are needed to generate the corresponding analytical plots.

V. Conclusions

$$Pr\left(N_{s,k}=m_k|N_{s,k-1}=m_{k-1},N_e=n_e\right)=\begin{cases}\binom{n_e}{m_k-m_{k-1}}\dfrac{\left(\prod_{l=m_k-n_e+1}^{m_{k-1}}l\right)\left(\prod_{l=m_{k-1}}^{m_k-1}(N_h-l)\right)}{\prod_{l=0}^{n_e-1}(N_h-l)}, & \text{if} \quad m_k\leq N_h \text{ and} \\ & 0\leq m_k-m_{k-1}\leq n_e \\ 0, & \text{otherwise}\end{cases} . \quad (10)$$

$$Pr(N_{s,k} \quad =m_k|N_e=n_e)=\begin{cases}1, & \text{if } k=1 \text{ and } m_k=n_e \\ 0, & \text{if } k=1 \text{ and } m_k\neq n_e \\ \sum_{m_{k-1}=0}^{N_h}Pr\left(N_{s,k}=m_k|N_{s,k-1}=m_{k-1},N_e=n_e\right) & \\ \quad \cdot Pr\left(N_{s,k-1}=m_{k-1}|N_e=n_e\right), & \text{otherwise}\end{cases} . \quad (11)$$

A thorough performance analysis of the self-embedding authentication method proposed by Fridrich and Goljan was given in this paper. In this analysis the hash bit error is firstly computed; then, the pdf of the projection of the received image blocks onto the reconstructed watermark is approximated depending on that robust hash bit error probability. This allows us to obtain closed formulas for the false positive and false negative probabilities.

An important characteristic of the self-embedding authentication method is that the embedding process itself can modify the robust hash of the image and consequently corrupt the reconstructed watermark; however in practical cases, a larger embedding distortion could increase the correlation between the received signal and the watermark estimated at the detector. Our analysis shows that, although the hash bit error probabilities increase with the embedding distortion, the overall performance of the authentication system improves. We have also observed, using (7) and (8), that given P_{fp}, P_{fn} is a convex function of N_p, with the optimal N_p almost invariant with N_h. Furthermore, we have seen how P_e depends on the standard deviation and the shaping parameter of the projection (modeled by a GGD) of the image blocks onto the pseudorandom patterns.

Future research will focus on the analysis of self-embedding authentication techniques where the spread-spectrum embedding is replaced by informed embedding strategies, e.g. [19].

Acknowledgments

The authors would like to thank the anonymous reviewers for their comments, that helped to improve the present manuscript.

APPENDIX A
DERIVATION OF HASH BIT ERROR PROBABILITY

Defining $G_j \triangleq \frac{1}{M}\mathbf{Z}^T\mathbf{S}^j = \frac{1}{M}\left(\eta\mathbf{X}^T\mathbf{S}^j + \gamma\eta\mathbf{W}^T\mathbf{S}^j + \mathbf{N}^T\mathbf{S}^j\right)$, the hash bit error probability can be written as

$$\begin{aligned}
P_e &= \frac{1}{2}Pr(|G_j| < T_d|H_j = 1) + \frac{1}{2}Pr(|G_j| > T_d|H_j = 0) \\
&= \frac{1}{2}Pr\left(|G_j| < T_d|\frac{1}{M}|\eta\mathbf{X}^T\mathbf{S}^j| > \eta T_e\right) \\
&\quad + \frac{1}{2}Pr\left(|G_j| > T_d|\frac{1}{M}|\eta\mathbf{X}^T\mathbf{S}^j| < \eta T_e\right) \\
&= \frac{Pr\left(|G_j| < T_d, \frac{1}{M}|\eta\mathbf{X}^T\mathbf{S}^j| > \eta T_e\right)}{2Pr\left(\frac{1}{M}|\eta\mathbf{X}^T\mathbf{S}^j| > \eta T_e\right)} \\
&\quad + \frac{Pr\left(|G_j| > T_d, \frac{1}{M}|\eta\mathbf{X}^T\mathbf{S}^j| < \eta T_e\right)}{2Pr\left(\frac{1}{M}|\eta\mathbf{X}^T\mathbf{S}^j| < \eta T_e\right)} \\
&= \int_{-\infty}^{\infty} Pr\left(\frac{1}{M}\mathbf{N}^T\mathbf{S}^j = t\right) \\
&\quad \cdot \left[\frac{\int_{\eta T_e}^{\infty} Pr\left(\frac{1}{M}|F_j + t| < T_d, \frac{1}{M}|\eta\mathbf{X}^T\mathbf{S}^j| = \tau\right)d\tau}{2Pr\left(\frac{1}{M}|\eta\mathbf{X}^T\mathbf{S}^j| > \eta T_e\right)}\right. \\
&\quad \left. + \frac{\int_{0}^{\eta T_e} Pr\left(\frac{1}{M}|F_j + t| > T_d, \frac{1}{M}|\eta\mathbf{X}^T\mathbf{S}^j| = \tau\right)d\tau}{2Pr\left(\frac{1}{M}|\eta\mathbf{X}^T\mathbf{S}^j| < \eta T_e\right)}\right] dt,
\end{aligned} \tag{12}$$

which yields (5) and where in (12) $F_j \triangleq \eta\mathbf{X}^T\mathbf{S}^j + \gamma\eta\mathbf{W}^T\mathbf{S}^j$.

APPENDIX B
DERIVATION OF THE DETECTION ERROR PROBABILITIES

First, we will derive the probability that given a number of bit errors n_e in the estimate of the hash, the number of wrong vectors $\mathbf{t}^l$ is n_s. In order to do so, we will obtain the probabilities $Pr(N_{s,k} = m_k|N_{s,k-1} = m_{k-1}, N_e = n_e)$, where $N_{s,k}$ denotes the number of wrong vectors $\mathbf{t}^l$ after k permutations of the reconstructed hash vector, $1 \leq k \leq N_p$. These probabilities can be shown to be given by (10), where $2 \leq k \leq N_p$, $n_e \leq N_h$. From there it is possible to write $Pr(N_{s,k} = m_k|N_e = n_e)$ as (11). Finally, given that the probability of having n_e mistakes in the hash estimate is

$$Pr(N_e = n_e) = \binom{N_h}{n_e} P_e^{n_e}\left(1 - P_e\right)^{(N_h - n_e)},$$

the probability of having n_s wrong vectors $\mathbf{t}^l$ after N_p permutations can be written as

$$Pr(N_s = n_s) = \sum_{n_e=0}^{N_h} Pr(N_{s,N_p} = m|N_e = n_e)Pr(N_e = n_e). \tag{13}$$

On the other hand, the correlation statistic under the hypothesis $\mathcal{H}_0$ is given by

$$\begin{aligned}
\rho_{\mathcal{H}_0} &= \frac{1}{M}\left(\eta\mathbf{x}^T\hat{\mathbf{w}} + \gamma\eta\mathbf{w}^T\hat{\mathbf{w}} + \mathbf{n}^T\hat{\mathbf{w}}\right) \\
&= \frac{1}{M}\left(\eta\mathbf{x}^T\hat{\mathbf{w}} + \frac{3}{N_h}\gamma\eta\sum_{l=1}^{N_h}\sum_{j=1}^{N_h}\mathbf{v}^{lT}\hat{\mathbf{v}}^j + \mathbf{n}^T\hat{\mathbf{w}}\right).
\end{aligned} \tag{14}$$

When $\mathbf{x}$ and $\hat{\mathbf{w}}$ are uncorrelated, so are $\mathbf{v}^l$ and $\hat{\mathbf{v}}^j$ when $l \neq j$, while if $N_s = n_s$, the vectors $\mathbf{v}^l$ and $\hat{\mathbf{v}}^l$ coincide in precisely $N_h - n_s$ values; then, using (14) it is possible to conclude that the mean of $\rho_{\mathcal{H}_0}$ is

$$\mathrm{E}\{\rho_{\mathcal{H}_0}|N_s = n_s\} = (N_h - n_s)\cdot\frac{\gamma\eta}{N_h}. \tag{15}$$

Similarly, the variance of $\rho_{\mathcal{H}_0}$ becomes

$$\text{Var}\left\{\rho_{\mathcal{H}_0}|N_s=n_s\right\}=\frac{1}{M}\left(\eta^2\cdot\sigma_X^2+\gamma^2\cdot\eta^2\cdot\sigma_{n_s}^2+\sigma_N^2\right), \tag{16}$$

where

$$\begin{aligned}\sigma_{n_s}^2 \quad \triangleq \quad & \frac{1}{N_h^2\sigma_V^4}\left((N_h-n_s)\sigma_{V^2}^2+n_s\sigma_V^4\right.\\ & \left.+(N_h-1)N_h\sigma_V^4+(N_h-n_s)(N_h-n_s-1)\sigma_V^4\right),\end{aligned}$$

and $\sigma_{V^2}^2 \triangleq \text{Var}\{V_i^2\}$, $i=1,\cdots,M$. Since $V_i\sim\mathcal{U}(-1,1)$, then $\sigma_V^4=1/9$ and $\sigma_{V^2}^2=4/45$. In order to analyze the distribution of $\rho_{\mathcal{H}_0}$, one can straightforwardly adapt the discussion in Sect. III about the distribution of $\mathbf{W}^T\cdot\mathbf{S}^j$ to obtain the distribution of $\mathbf{X}^T\cdot\mathbf{W}$ and $\mathbf{N}^T\cdot\mathbf{W}$, and apply the CLT for approximating the pdf of $\hat{\mathbf{W}}^T\cdot\mathbf{W}$ by a Gaussian for large values of N_h and M. The resulting approximation is $\rho_{\mathcal{H}_0|N_s=n_s}\sim\mathcal{N}\left((N_h-n_s)\cdot\frac{\gamma\eta}{N_h},\frac{1}{M}\left(\eta^2\cdot\sigma_X^2+\gamma^2\cdot\eta^2\cdot\sigma_{n_s}^2+\sigma_N^2\right)\right)$. Equivalently, it is clear that for $\mathcal{H}_1$ we can approximate $\rho_{\mathcal{H}_1|N_s=n_s}\sim\mathcal{N}\left((N_h-n_s)\cdot\frac{\gamma}{N_h},\frac{1}{M}\left(\sigma_X^2+\gamma^2\cdot\sigma_{n_s}^2\right)\right)$. In this way, when $\text{Var}\{\rho_{\mathcal{H}_0}\}/\text{Var}\{\rho_{\mathcal{H}_1}\}>1$, $P_{fp}=\sum_{n_s=0}^{N_h}Pr(N_s=n_s|\mathcal{H}_0)Pr(T_0\leq\rho_{\mathcal{H}_0}\leq T_1)$ and $P_{fn}=\sum_{n_s=0}^{N_h}Pr(N_s=n_s|\mathcal{H}_1)[Pr(\rho_{\mathcal{H}_1}<T_0)+Pr(\rho_{\mathcal{H}_1}>T_1)]$; on the other hand, when $\text{Var}\{\rho_{\mathcal{H}_0}\}/\text{Var}\{\rho_{\mathcal{H}_1}\}<1$, $P_{fp}=\sum_{n_s=0}^{N_h}Pr(N_s=n_s|\mathcal{H}_0)[Pr(\rho_{\mathcal{H}_0}<T_0)+Pr(\rho_{\mathcal{H}_0}>T_1)]$ and $P_{fn}=\sum_{n_s=0}^{N_h}Pr(N_s=n_s|\mathcal{H}_1)Pr(T_0\leq\rho_{\mathcal{H}_1}\leq T_1)$. Finally, when $\text{Var}\{\rho_{\mathcal{H}_0}\}=\text{Var}\{\rho_{\mathcal{H}_1}\}$, $T_1=\infty$ so $P_{fp}=\sum_{n_s=0}^{N_h}Pr(N_s=n_s|\mathcal{H}_0)Pr(T_0\leq\rho_{\mathcal{H}_0})$ and $P_{fn}=\sum_{n_s=0}^{N_h}Pr(N_s=n_s|\mathcal{H}_1)Pr(T_0>\rho_{\mathcal{H}_1})$. Replacing in the previous expressions the probabilities on $\rho_{\mathcal{H}_i}$ by their integral form, one obtains (7) and (8), respectively. Aside from the obvious differences between the expressions (7) and (8), it is important to note that $Pr(N_s=n_s|\mathcal{H}_0)$ and $Pr(N_s=n_s|\mathcal{H}_1)$ are different, since the expression for P_e depends on the considered hypothesis, as stated in Sect. III.

REFERENCES

[1] M. Schneider and S.-F. Chang, "A robust content based digital signature for image authentication," in *Proc. IEEE Int. Conf. Image Processing*, vol. 3, September 1996, pp. 227–230.

[2] G. L. Friedman, "The trustworthy digital camera: Restoring credibility to the photographic image," *IEEE Trans. Consumer Electronics*, vol. 39, no. 4, pp. 905–910, November 1993.

[3] S. Bhattacharjee and M. Kutter, "Compression tolerant image authentication," in *Proc. IEEE Int. Conf. Image Processing*, vol. 1, October 1998, pp. 435–439.

[4] D. Kundur and D. Hatzinakos, "Digital watermarking for telltale tamper proofing and authentication," *Proc. of the IEEE*, vol. 87, no. 7, pp. 1167–1180, July 1999.

[5] R. Venkatesan, S.-M. Koon, M. Jakubowski, and P. Moulin, "Robust image hashing," in *Proc. IEEE Int. Conf. Image Processing*, vol. 3, September 2000, pp. 664–666.

[6] C.-Y. Lin and S.-F. Chang, "A robust image authentication method distinguishing JPEG compression from malicious manipulation," *IEEE Trans. Circuits and Systems of Video Technology*, vol. 11, no. 2, pp. 153–168, February 2001.

[7] J. Cannons and P. Moulin, "Design and statistical analysis of a hash-aided image watermarking system," *IEEE Trans. Image Processing*, vol. 13, no. 10, pp. 1393–1408, October 2004.

[8] A. Swaminathan, Y. Mao, and M. Wu, "Robust and secure image hashing," *IEEE Trans. Information Forensics and Security*, vol. 1, no. 2, pp. 215–230, June 2006.

[9] V. Monga and K. Mıhçak, "Robust and secure image hashing via non-negative matrix factorizations," *IEEE Trans. Information Forensics and Security*, vol. 2, no. 3, pp. 376–390, September 2007.

[10] J. Fridrich, "Robust bit extraction from images," in *Proc. IEEE Int. Conf. Multimedia Computing and Systems*, vol. 2, July 1999, pp. 536–540.

[11] J. Fridrich and M. Goljan, “Robust hash functions for digital watermarking,” in *Proc. Int. Conf. Information Technology: Coding and Computing*, March 2000, pp. 178–183.

[12] J. Fridrich, “Visual hash for oblivious watermarking,” in *Proc. SPIE Photonic West Electronic Imaging 2000, Security and Watermarking of Multimedia Contents*, January 2000, pp. 286–294.

[13] Özgür Ekici, B. Sankur, B. Coskun, U. Naci, and M. Akcay, “Comparative evaluation of semifragile watermarking algorithms,” *Journal of Electronic Imaging*, vol. 13, no. 1, pp. 209–219, January 2004.

[14] I. Cox, J. Kilian, T. Leighton, and T. Shamoon, “Secure spread spectrum watermarking for images, audio and video,” in *Proc. Int. Conf Image Processing*, vol. 3, September 1996, pp. 243–246.

[15] H. L. van Trees, *Detection, Estimation, and Modulation Theory. Part I*. John Wiley and Sons, 2001.

[16] K. Birney and T. Fischer, “On the modeling of dct and subband image data for compression,” *IEEE Trans. Image Processing*, vol. 4, no. 2, pp. 186–193, February 1995.

[17] “Uncompressed Colour Image Database (UCID) of Austin University (v2).” [Online]. Available: http://vision.cs.aston.ac.uk/datasets/UCID/ucid.html

[18] W. Feller, *An Introduction to Probability Theory and Its Applications. Vol II*. Wiley, 1971.

[19] V. Kitanovski, D. Taskovski, and S. Bogdanova, “Watermark generation using image-dependent key for image authentication,” in *Proc. EUROCON Int. Conf. Computer as a Tool*, vol. 2, November 2005, pp. 947–950.