PERFOMANCE ANALYSIS OF THE FRIDRICH-GOLJAN SELF-EMBEDDING AUTHENTICATION METHOD

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

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

ABSTRACT

A performance analysis of the authentication method proposed by J. Fridrich and M. Goljan is carried out. This method has the particular feature that both the embedder and the detector generate the watermark from a perceptual digest of the image. Hence, in order to accurately analyze the performance, the digest errors caused by the watermark embedding, the addition of a complementary signal and the scaling attacks are also taken into account.

Index Terms— Content-based authentication, robust hash, performance analysis.

1. INTRODUCTION

Due to the ease of modifying digital objects, ensuring the authenticity and/or integrity of digital contents has recently become an unavoidable requirement in medium-high security profile scenarios (e.g., recordings of people entering restricted areas, banks videosurveillance systems, etc.). Several watermarking techniques have been proposed to tackle this issue; these algorithms embed a low-power signal (a.k.a. digital watermark) in the digital content to be protected. Analyzing the inserted watermark, one can check whether the received digital content is authentic or is a forgery. Some of these watermarking techniques generate the watermark as function of a secret key and a perceptual digest of the digital object (a.k.a. robust hash, perceptual hash or soft hash). Ideally, the difference between the robust hash vectors of two digital objects is proportional to the dissimilarity of their meaning.

From the set of authentication techniques which embed a watermark that depends on the image robust hash (also known as watermarking self-embedding authentication), the algorithm proposed by Fridrich and Goljan [1] is one of the most prominent. However, a theoretical performance analysis of this widely-referenced self-embedding authentication algorithm is still lacking. Thus, for performance comparison Monte Carlo techniques need to be used, with the obvious drawback of requiring long time simulations whenever small probabilities are to be estimated. The analysis proposed here is based on computing the hash bit error probability due to the watermark embedding and noise, and showing its impact on the Receiver Operating Characteristic (ROC) of the overall scheme.

In the next section we give a brief introduction to the robust authentication method proposed by Fridrich and Goljan, including the embedding and detection processes. Performance is addressed in Sect. 3, whereas in Sect. 4 simulations are carried out to check the accuracy of our analysis. Finally, Sect. 5 presents the main conclusions.

1.1. 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.

2. DESCRIPTION OF THE FRIDRICH-GOLJAN METHOD

In this section a description of the method by Fridrich and Goljan [1], that will be analyzed in Sect. 3, is given; furthermore, a correlation based detector for that method is proposed.

2.1. Hash and Watermark 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. For each $\mathbf{x}$, a set of N_h length-M pseudo-random sequences $\mathbf{s}^j$, $1 \leq j \leq N_h$ are produced from a $\sqrt{M} \times \sqrt{M}$ pseudo-random matrix (obtained depending on the secret key θ of the system) which are rearranged

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 SPROACTIVE (ref. TEC2007-68094-C02-01/TCM).

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

as the length-M vector $\mathbf{s}^j$. Thus, from each $\mathbf{x}$, and depending on the set of $\mathbf{s}$, 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 derived 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. $\frac{1}{M}\mathbf{X}^T \cdot \mathbf{S}^j$.

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 θ, 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 $\mathbf{w} = \sqrt{\frac{3}{N_h}} \sum_{l=1}^{N_h} \mathbf{v}^l$.

The watermark $\mathbf{w}$ is embedded in the host signal using Additive Spread Spectrum [2] in the pixel domain, so the corresponding block of the watermarked image is obtained as $\mathbf{y} = \mathbf{x} + \gamma\mathbf{w}$, where γ is an embedding strength parameter.

2.2. 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 $\mathbf{z}$. Given that $\mathbf{z}$ and the host image differ, the quantization threshold at the detector T_d is calculated again from $\frac{1}{M}\mathbf{Z}^T \cdot \mathbf{S}^j$ as described in Sect. 2.1. The main objective of the detector is to decide whether the estimate of the watermark is present (authentic block) or not (manipulated block).

In the current work the decision on the presence or absence of the estimate of the watermark is formulated as a binary hypothesis test, namely,

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

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 $\mathbf{n}$, 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.

In order to solve this binary test problem, the well-known likelihood ratio test is used. In this way, 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. This statistic is widely used among the research community due to its simplicity, even when the mentioned conditions on the Gaussianity and independence of the signals are known not to be verified; due to this extensive use, this statistic was chosen for detection in the current work. Under these considerations, the thresholds T_0 and T_1, which define the decision regions, can be obtained from the respective expectations and variances of the distribution of ρ when $\mathcal{H}_0$ (E$\{\rho_{\mathcal{H}_0}\}$,Var$\{\rho_{\mathcal{H}_0}\}$) and $\mathcal{H}_1$ ((E$\{\rho_{\mathcal{H}_1}\}$,Var$\{\rho_{\mathcal{H}_1}\}$) hold [3].

3. PERFORMANCE ANALYSIS

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) [4], 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 GGD 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 hypothesis has been validated using the luminance component of a set of 14 images constituted by those images with size 256×256 pixels from volume "miscellaneous" of the USC-SIPI database [5]. Specifically, the Kullback-Leibler divergence (KLD) between the histogram of the projection of the blocks onto the pseudorandom patterns and the corresponding GGD is an order of magnitude lower than that resulting from a Gaussian distribution.

The first step of our analysis is the calculation of the probability of flipping one bit of the hash obtained at the detector with respect to the hash computed at the embedder. For large values of M, it is shown in [6] that the projected watermark $\frac{\gamma\eta}{M}\mathbf{W}^T \cdot \mathbf{S}^j$ can be modeled by $\mathcal{N}(0, \frac{\gamma^2\eta^2\sigma_S^2}{M})$, where σ_S is the standard deviation of the pseudo-random sequences $\mathbf{s}$. On the other hand, the projection of the complementary signal $\mathbf{N}$ onto $\mathbf{S}^j$, i.e. $\frac{1}{M}\mathbf{N}^T \cdot \mathbf{S}^j$, is modeled by a GGD with parameters A_N, β_N and c_N, being its pdf denoted by $f_N(x)$.[2]

Taking this into account, in [6] it is shown that the probability of a hash bit error under hypothesis $\mathcal{H}_0$ can be expressed as (1), where $\mathcal{Q}(x) \triangleq \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{\tau^2}{2}} d\tau$. In (1) the first inner integral gives the probability that the hash bit at the detector is 1 when the counterpart at the embedder is 0, and the second inner integral calculates the probability of changing a bit 1 at the embedder to 0 at the detector. Both of them consider the effect of the host and watermark distribution, whereas the outer integral takes into account the noise effect. It is worth pointing out that the previous expression is valid whenever

[2] Note that this characterization is valid for both the cases where $\mathbf{n}$ is an image or additive Gaussian noise (for which $c_N = 2$).

$$P_e \approx \frac{2}{\eta}\int_{-\infty}^{\infty} A_N e^{-|\beta_N t|^{c_N}} \left[\frac{\int_{\eta T_e}^{\infty} A_X e^{-|\frac{\beta_X}{\eta}\tau|^{c_X}} \left(1 - \mathcal{Q}\left(\frac{\sqrt{M}(T_d-\tau-t)}{\sqrt{\gamma^2\eta^2\sigma_S^2}}\right) - \mathcal{Q}\left(\frac{\sqrt{M}(T_d+\tau+t)}{\sqrt{\gamma^2\eta^2\sigma_S^2}}\right)\right) d\tau}{\frac{4}{\eta}\int_{\eta T_e}^{\infty} A_X e^{-|\frac{\beta_X}{\eta}\tau|^{c_X}} d\tau} \right.$$
$$\left. + \frac{\int_0^{\eta T_e} A_X e^{-|\frac{\beta_X}{\eta}\tau|^{c_X}} \left(\mathcal{Q}\left(\frac{\sqrt{M}(T_d-\tau-t)}{\sqrt{\gamma^2\eta^2\sigma_S^2}}\right) + \mathcal{Q}\left(\frac{\sqrt{M}(T_d+\tau+t)}{\sqrt{\gamma^2\eta^2\sigma_S^2}}\right)\right) d\tau}{\frac{4}{\eta}\int_0^{\eta T_e} A_X e^{-|\frac{\beta_X}{\eta}\tau|^{c_X}} d\tau} \right] dt \quad (1)$$

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

$$P_{fn} \approx \sum_{n_s=0}^{N_h} Pr(N_s = n_s|\mathcal{H}_1)\left[\mathcal{Q}\left(\frac{\sqrt{M}\xi((N_h - n_s)\frac{\gamma}{N_h} - T_0)}{\sqrt{\sigma_X^2 + \gamma^2\sigma_{n_s}^2}}\right) + \xi\mathcal{Q}\left(\frac{\sqrt{M}(T_1 - (N_h - n_s)\frac{\gamma}{N_h})}{\sqrt{\sigma_X^2 + \gamma^2\sigma_{n_s}^2}}\right)\right] \quad (3)$$

$\eta > 0$; 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, (1) can be easily adapted to hypothesis $\mathcal{H}_1$ by setting $\eta = 1$ and $\mathbf{N}^T\mathbf{S}^j = 0$ ($\beta_N = \infty$).

As it is described in Sect. 2.1, 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 $\mathbf{t}^l$, denoted by N_s. Hence, in order to quantify the watermark estimation error it is necessary to know the probability of the number N_S of wrong vectors $\mathbf{t}^l$. In this way, the probability of the number of wrong $\mathbf{t}^l$ vectors after k permutations (denoted by $N_{s,k}$) being m_k given that $N_e = n_e$ is calculated by the recursive formula

$$Pr(N_{s,k} = m_k|n_e) = \begin{cases} 1, & \text{if } k = 1 \text{ and } m_k = n_e, \\ 0, & \text{if } k = 1 \text{ and } m_k \neq n_e, \\ \displaystyle\sum_{m_{k-1}=0}^{N_h} Pr\left(N_{s,k} = m_k|m_{k-1}, n_e\right) & \\ \times Pr\left(N_{s,k-1} = m_{k-1}|n_e\right), & \text{otherwise,} \end{cases} \quad (4)$$

which depends on both the probability of $N_{s,k-1} = m_{k-1}$ given $N_e = n_e$ (i.e., $Pr(N_{s,k-1} = m_{k-1}|n_e)$), and the probability of $N_{s,k}$ being equal to m_k, given that $N_{s,k-1} = m_{k-1}$ and $N_e = n_e$ (i.e., $Pr(N_{s,k} = m_k|m_{k-1}, n_e)$). This last probability is calculated as $Pr\left(N_{s,k} = m_k|m_{k-1}, n_e\right) = \binom{n_e}{m_k - m_{k-1}} \frac{\left(\prod_{l=m_k-n_e+1}^{m_{k-1}} l\right)\left(\prod_{l=m_{k-1}}^{m_k-1}(N_h - l)\right)}{\prod_{l=0}^{n_e-1}(N_h - l)}$ if $m_k \leq N_h$ and $0 \leq m_k - m_{k-1} \leq n_e$, and 0 elsewhere, and where $2 \leq k \leq N_p$, $n_e \leq N_h$. By setting $k = N_p$ and $m_k = n_S$ in (4), the probability of N_s after N_p permutations is obtained as the expectation of $Pr(N_{s,N_p} = n_s|N_e = n_e)$ with respect to the distribution of N_e, i.e.,

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

where $Pr(N_e = n_e)$ can be calculated by combinatorial analysis as

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

In [6] it is shown that the values of the probability of false positive and false negative are respectively given by (2) and (3), where $Pr(N_s = n_s|\mathcal{H}_0)$ and $Pr(N_s = n_s|\mathcal{H}_1)$ denote the probability of $N_s = n_s$ under hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively, and $\xi = \text{sign}\left(\text{Var}\{\rho_{\mathcal{H}_0}\} - \text{Var}\{\rho_{\mathcal{H}_1}\}\right)$ when $\text{Var}\{\rho_{\mathcal{H}_0}\} \neq \text{Var}\{\rho_{\mathcal{H}_1}\}$. When $\text{Var}\{\rho_{\mathcal{H}_0}\} = \text{Var}\{\rho_{\mathcal{H}_1}\}$ the obtained expressions are still valid by doing $T_1 = \infty$ and $\xi = 1$. Furthermore, σ_X denotes the standard deviation of the image block $\mathbf{x}$ and 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 $N_s = n_s$, which can be calculated as

$$\sigma_{n_s}^2 \triangleq \frac{1}{N_h^2 \cdot \sigma_V^4}\left[(N_h - n_s)\cdot\sigma_{V^2}^2 + (N_h - 1)\cdot N_h \cdot \sigma_V^4 \right.$$
$$\left. + n_s \cdot \sigma_V^4 + (N_h - n_s)\cdot(N_h - n_s - 1)\cdot\sigma_V^4\right],$$

where $\sigma_V^4 = 1/9$ and $\sigma_{V^2}^2 = 4/45$, as $\sigma_{V^2}^2 \triangleq \text{Var}\{V_i^2\}$ and $V_i \sim \mathcal{U}(-1, 1)$. Further details on the presented analysis and proofs can be found in [6].

4. EXPERIMENTAL RESULTS

In this section we experimentally check the accuracy of our model. In order to do so, 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 14 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, in reasonable work scenarios, better performance, in terms of the ROC, is achieved with larger values of γ, although one should also notice that a larger γ implies a larger distortion. Hence, a trade-off between distortion and performance should be achieved.

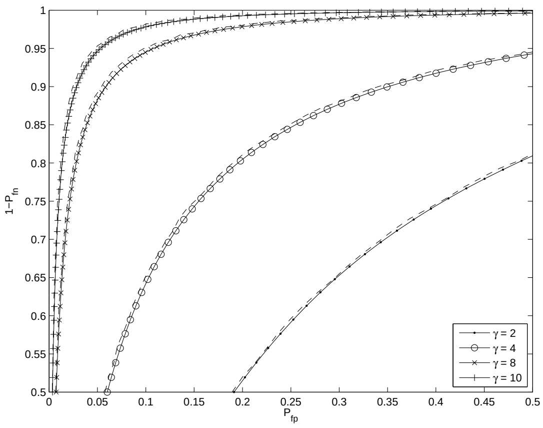

Fig. 1. Analytical (dashed lines) and empirical (solid lines) ROCs for a set of 14 images [5]. $M = 4096$, $N_h = 16$, $N_p = 5$, $\eta = 0$, and $\gamma \in \{2, 4, 8, 10\}$.

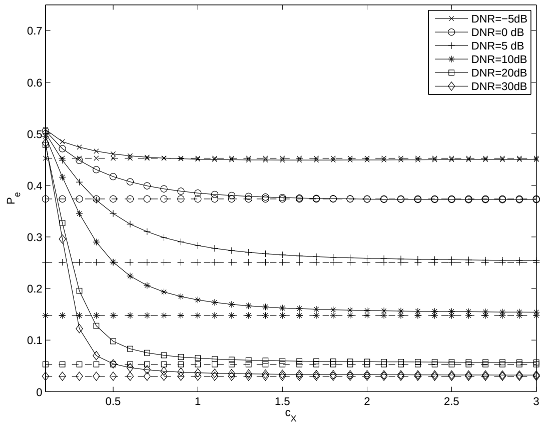

Fig. 2. P_e vs shaping factor c_X (solid lines) and P_e for uniform $f_D(x)$ (dashed lines) for different DNRs when $\mathcal{H}_0$ holds. N Gaussian distributed, $M = 4096$, $\eta = 1$, $\sigma_X = \frac{0.1\sqrt{M}}{\sigma_S}$ and $\gamma = 10$.

The dependence of P_e on the MLE estimated parameters of the GGD used for modeling $\mathbf{X}^T \cdot \mathbf{S}^j$ when $\mathcal{H}_0$ holds is illustrated in Fig. 2, where P_e is plotted as a function of the shaping factor c_X for several values of Data to Noise Ratio (DNR=σ_X^2/σ_N^2), and $\mathbf{n}$ is Gaussian distributed. The simulations were obtained with the following parameters: $\sigma_X = \frac{0.1\sqrt{M}}{\sigma_S}$ (this is a typical value for real images), $M = 4096$, $\eta = 1$ and $\gamma = 10$. On one hand, and according to intuition, it can be seen that the larger the values of DNR, the smaller the hash bit error probabilities due to watermark embedding and noise; indeed for large values of DNR, the hash vector errors are mainly produced by the watermark embedding distortion (which is the same for every curve), explaining the resemblance of the results for DNR=20 dB and DNR=30 dB. On the other hand, for low values of DNR the added Gaussian noise dominates over the watermark embedding. Concerning the dependency of P_e with c_X it is worth pointing out that, although P_e seems to be monotonically decreasing with c_X, a closer examination of Fig. 2 reveals that this is not in fact the case. This behavior is especially noticeable for the DNR=-5 dB plot, where P_e is monotonically increasing with c_X in an interval of this parameter. In this sense, Fig. 2 also shows P_e when c_X goes to ∞, i.e. f_D is uniformly distributed; although the obtained value is not really a bound on P_e, it seems to be a reasonable approximation for large values of c_X.

5. CONCLUSIONS

A performance analysis of the self-embedding authentication method proposed by Fridrich and Goljan was carried out. An important characteristic of any 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, from a performance perspective, a larger embedding distortion is usually preferred, as in realistic scenarios it typically increases the correlation between the received signal and the watermark estimated at the detector; therefore, a trade-off between distortion and performance must be considered.

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. These results have been compared with an approximation to P_e for large c_X values, showing to be reasonably close to the real probability for $c_X \geq 1.5$.

6. REFERENCES

[1] 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.

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

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

[4] K.A. Birney and T.R. 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.

[5] "The University of Southern California-SIPI Image Database," .

[6] G. Domínguez-Conde, P. Comesaña, and F. Pérez-González, "Performance analysis of Fridrich-Goljan self-embedding authentication method," *IEEE Trans. Information Forensics and Security*, October 2008, Submitted.