Statistical Analysis of Watermarking Schemes for Copyright Protection of Images

JUAN R. HERNÁNDEZ, STUDENT MEMBER, IEEE,
AND FERNANDO PÉREZ-GONZÁLEZ, MEMBER, IEEE

In this paper, we address the problem of the performance analysis of image watermarking systems that do not require the availability of the original image during ownership verification. We focus on a statistical approach to obtain models that can serve as a basis for the application of the decision theory to the design of efficient detector structures. Special attention is paid to the possible nonexistence of a statistical description of the original image. Different modeling approaches are proposed for the cases when such a statistical characterization is known and when it is not. Watermarks may encode a message, and the performance of the watermarking system is evaluated using as a measure the probability of false alarm, the probability of detection when the presence of the watermark is tested, and the probability of error when the information that it carries is extracted. Finally, the modeling techniques studied are applied to the analysis of two watermarking schemes, one of them defined in the spatial domain, and the other in the direct cosine transform (DCT) domain. The theoretical results are contrasted with empirical data obtained through experimentation covering several cases of interest. We show how choosing an appropriate statistical model for the original image can lead to considerable improvements in performance.

***Keywords*—** *Codes, copyright protection, cryptography, decision-making, image communication, image processing, information theory.*

I. INTRODUCTION

Recent years have witnessed a stunning proliferation of techniques for representation, storage, and distribution of digital multimedia information. Nowadays, network design is oriented to digital data delivery, while content providers are rapidly transforming their archives to a digital format. Unfortunately, all these developments have also a serious drawback: digital copies can be made identical to the original; moreover, it is fairly simple to manipulate or reuse the information in an unauthorized way. Final quality and impracticality of massive copying were in the past key factors that put limits to any widespread distribution of illegal copies; on the contrary, emerging networks allow a fast and bulky dissemination with no loss of quality. This creates new threats to copyright protection and puts the whole creative process in danger.

Cryptography is an effective solution to the digital distribution problem, but it has to be coupled with costly and specialized hardware in order to preclude direct access to data in digital format. However, most cryptographic protocols are concerned with secure communications instead of ulterior copyright infringements. A good example are the cryptographic devices (set-top boxes) used for access control in digital television broadcasting [1]. There, the major goal lies in avoiding unauthorized customers to access programs that are being broadcast in scrambled form [2], but once data have been (even legally) unscrambled it is quite simple to save them for further manipulation or dissemination. In other scenarios, such as the Internet, the "network value" heavily depends on the existence of relatively cheap and general-purpose hardware that eliminates a significant capital cost both to the user and the service provider, as explained by Metcalfe's Law [3]. Consequently, there is an increasing need for software that allows for protection of ownership rights. It is in this context where watermarking techniques come to our help.

A digital watermark is a distinguishing piece of information that is adhered to the data that it is intended to protect. There are two kinds of watermarks: perceptible and imperceptible. For obvious reasons, the latter are more suitable to become part of a digital copyright system. Imperceptible watermarks obstruct illegal copying by ensuring that ownership information is unnoticeably embedded into the digital data. Considering that watermarking can be applied to data of a very different nature, the imperceptibility constraint must be achieved by carefully taking into account the properties of the human senses. For instance, early work in the image watermarking field did not tackle adequately the perceptibility issue and embedded the watermark in the least significant (the most "insignificant") bits of the original image [4]–[7], thus making it easy to remove or alter this additional information. On the other hand, more recent methods consider the characteristics of the human

Manuscript received September 15, 1998; revised January 3, 1999.

The authors are with Departamento de Tecnoloxías das Comunicacións, University of Vigo, Vigo 36200 Spain.

Publisher Item Identifier S 0018-9219(99)04949-X.

visual systems (e.g., low sensitivity to edge changes) to enhance the robustness of the watermark.

The previous example leads us to the robustness issue [8]. In addition to imperceptibility, there are some desirable features that a watermark should have. First, it should be resilient to standard manipulations (e.g., MUSICAM compression in the case of audio signals [9]) as well as intentional manipulations (e.g., watermark-removal programs discussed below). Second, it should be statistically unremovable (or better yet, undetectable), which means that a statistical analysis of different pieces of data watermarked by the same provider should not lead to any gain from the attacker point of view. Finally, the watermark should withstand multiple watermarking to facilitate the tracking of the subsequent transactions to which an image is subject. Note that for different applications, the watermark should resist quite different types of manipulations. An example is photocopying, which is a possible attack for document images, but almost useless when thinking of color images. Furthermore, the computational power required by an attacker will strongly depend on the application; watermark removal in a digital video sequence would be much more expensive than for a still image. Ideally, watermark destruction attacks should also affect the original data in a similar way; however, to what extent ownership should be preserved after data is severely distorted is a difficult question.

Watermarking, like cryptography, needs secret keys that map rights to owners. However, in most applications, embedment of additional information is required. This hidden information may consist, among others, in owner, distributor, or recipient identifiers, transaction dates, serial numbers, etc., that may become vital when tracking some illegal distribution. In some cases, a trusted authority that issues certificates is needed [10]. For instance, signed time stamps could be imperceptibly hidden in the original data to prevent counterfeiting based on successive watermarks [11], [12]. In fact, the issue of data hiding (also called steganography) leads to the problem of correctly extracting (decoding) this information once in possession of the secret key. In most cases, there will be a certain probability of error for the extracted information. This probability of error can be used as a measure of the performance of a watermarking system. Note that this probability will increase with the number of bits in the hidden message, thus imposing a limit on the length of the secret message one wants to convey.

In addition to the data-hiding problem there is the detection problem, in which it is tested whether data were watermarked with a certain key, therefore serving for ownership determination purposes. The detection problem produces a binary answer: data were (or were not) watermarked with a given key. Consequently, the problem can be formulated as a statistical hypothesis test, for which a probability of false alarm (i.e., of deciding that a given key was used while it was not) and a probability of detection (i.e., of correctly deciding that a given key was used) can be defined as quality measures. More precisely, one would fix a certain value of the probability of false alarm and then evaluate if the system gives an acceptable probability of detection. Note that the probability of false alarm should be kept to an extremely low value if the watermarking system is to be used for commercial purposes, since the existence of "false positives" would undermine its credibility.

In this paper we will concentrate on watermarking of still images, which is the case that has generated the largest amount of research in the field. However, it is interesting to point out that watermarking has been also applied to other types of data, such as document images [13]–[20], audio signals [21], [22], video signals [23]–[31], three-dimensional (3-D) objects [32], and even software or hardware [33], [34]. Many image watermarking methods have mushroomed over the past years, even with commercial products available or in preparation (e.g., NEC, Sony, Hitachi, or Kodak). Although some of them need knowledge of the original image [35], [36], we will assume throughout this paper that this is not available during the watermark extraction and detection processes. While such knowledge would greatly simplify the former tasks, especially if the watermarked image has suffered common geometric distortions, any Internet-oriented long-term product should take it as unacceptable, provided that resorting to the originals would be unmanageable when huge quantities of images need be compared by intelligent agents searching the net for unauthorized copies. Then, the detection and extraction tasks will have to be performed without access to the source image.

While cryptographic protocols have achieved the desired security level that expedites their use in electronic commerce applications [37], this cannot be said yet for watermarking systems. Recent advances in watermark-removal programs, such as unZign and StirMark [38] and others [39], [40] that have succeeded in washing the watermark away with little impact on the perceptibility constraint, even for commercial systems, are quite discouraging but will foster new research in the field. Actually, most methods that do not require knowledge of the original image are not robust to simple geometric transformations or cropping. This problem, also known as the synchronization problem, is one of the current salient points in image watermarking that will deserve much attention in the near future.

First stages in the development of watermarking techniques have produced an impressive amount of algorithms, even though in most cases no theoretical limits to their performance were given. We believe that such a theoretical approach is the only way to turn digital copyright protection into a mature discipline, at a level comparable to other branches of communications and cryptography. In this paper we will show how a careful modeling of the problem can help not only to assess the performance of the various methods but also to considerably improve them. We will focus on the study of watermarking systems through the application of statistical analysis techniques, and we will use statistical decision theory to derive the detection structures involved when ownership of an image must be verified. The reasons why a statistical approach is convenient are

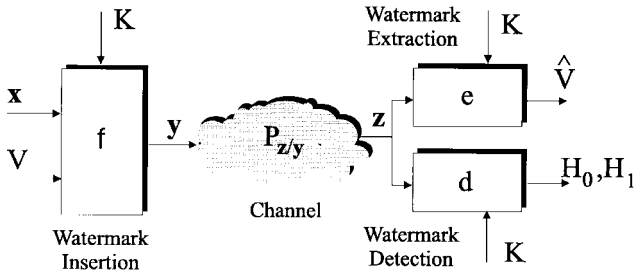


Fig. 1. General model of a watermarking system.

manyfold. First, through a statistical formulation of the problem performance measures can be defined to study rigorously to which extent a watermarking technique can be applied as a copyright protection mechanism. Dependence of the achievable level of protection on the characteristics of images can also be analyzed with such an approach. Furthermore, with this kind of analysis it is possible to obtain efficient detectors that optimize the performance and to assign appropriate values to system parameters so that a certain level of performance is guaranteed.

In Section II, we present a general model of a watermarking system and some definitions of concepts that will be revisited throughout the paper. In Section III, we particularize the general model to watermarking techniques based on the addition of a spread spectrum signal carrying some information. The rest of the paper refers to this kind of watermarks. In Section IV, we formulate statistically the problems of watermark detection and extraction of the information carried by the watermark, when a statistical description of the original image is possible. In Section V, we formulate the same problems when such a statistical characterization of the original image is unknown. Finally, in Sections VIII and IX, we apply the analytical techniques developed in Sections IV and V to the study of two watermarking techniques, one performed in the spatial domain and the other performed in the direct cosine transform (DCT) domain. The theoretical results derived in both sections are contrasted with empirical data obtained through experiments performed with several test images.

II. General Model of a Watermarking System

In Fig. 1 we have represented in block diagram form the general model of a watermarking system. A similar model was proposed in [41] in an information theoretical context. From now on, variables in bold letters will represent vectors whose elements will be referenced using the notation $\boldsymbol{x} = (x_1, \ldots, x_L)$. An image $\boldsymbol{x}$ is transformed into a watermarked version $\boldsymbol{y}$ applying a watermarking function f that also takes as inputs a secret key K only known to the copyright owner and a message V taken from a finite discrete alphabet with M elements. If the source output $\boldsymbol{x}$ is not watermarked, then it is left unaltered.

The watermarked version $\boldsymbol{y}$ is delivered in place of $\boldsymbol{x}$ to the intended recipient. Then it can suffer unintentional distortions or attacks aimed at destroying the watermark information. Note that even intentional attacks are performed without any knowledge about K or $\boldsymbol{x}$, since they are not publicly available. The alterations suffered by $\boldsymbol{y}$ can be thus modeled as a noisy channel whose input $\boldsymbol{y}$ and output $\boldsymbol{z}$ are linked by the conditional distribution $p_{z|y}$.

Two tests are involved in the ownership verification process. First, a watermark detector d decides whether the image $\boldsymbol{z}$ under test contains a watermark generated with a certain key K. Hence, this detector takes as inputs the image $\boldsymbol{z}$ and the secret key K and yields a Boolean output which indicates the decision. If the watermark detector decides that a watermark is present, then authorship by the person who possesses the secret key K is proved and extraction of the hidden message can be performed afterwards. This task is accomplished by a watermark decoder e, whose inputs are $\boldsymbol{z}$ and K. As a result, it outputs an estimate $\hat{V}$ of the hidden message. Note that we have assumed that both the watermark detector and the watermark decoder have no access to the original image.

The watermark detector is characterized by two performance measures: the probability of false alarm P_F and the probability of detection P_D. The former indicates the probability of yielding a positive result in the watermark detection test when $\boldsymbol{z}$ does not actually contain a watermark generated from K. The latter is the probability of getting a positive result when the image does contain such a watermark. These two probabilities have been proposed for quality assessment in [42]. To express these probabilities in mathematical notation, let us denote the event "the image is watermarked" by H_1 and the event "the image is not watermarked" by H_0. Then

$$P_F \triangleq \Pr\{d(\boldsymbol{z}, K) = H_1 | H_0\} \tag{1}$$

$$P_D \triangleq \Pr\{d(\boldsymbol{z}, K) = H_1 | H_1\}. \tag{2}$$

The performance of the watermark decoder is given by the probability of error P_e, defined as the probability of getting a wrong estimate $\hat{V}$

$$P_e \triangleq \Pr\left\{\hat{V} \neq V\right\}. \tag{3}$$

This is a measure of the overall performance of the watermarking system that is almost useless. Two conditional error probabilities are specially interesting for their applicability to the design of detectors and decoders. One of them is the probability of error conditioned to a given original image $\boldsymbol{x}$. The other is the probability of error conditioned to a certain key K

$$P_e(\boldsymbol{x}) \triangleq \Pr\left\{\hat{V} \neq V | \boldsymbol{x}\right\} \tag{4}$$

$$P_e(K) \triangleq \Pr\{\hat{V} \neq V | K\}. \tag{5}$$

The former indicates how appropriate an image is for data hiding and the latter expresses the average performance seen by a copyright holder.

In the following sections we study a special kind of watermarking function based on the addition of a spread spectrum signal. No explicit assumption will be made about the domain where $\boldsymbol{x}$ is defined. It could be just a spatial

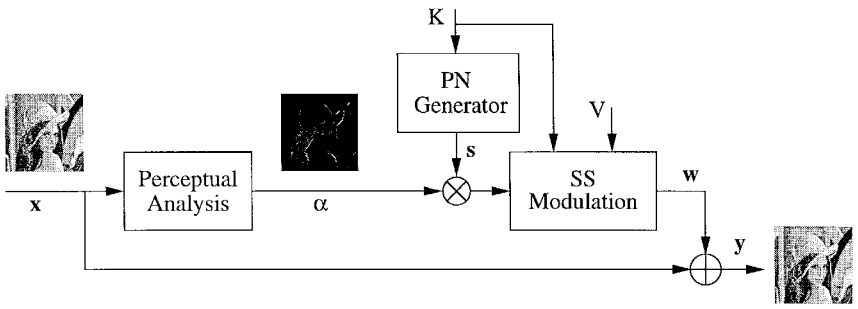


Fig. 2. Generation of a spread spectrum watermark.

domain representation of the image luminance or any kind of representations obtained after applying transforms such as the fast Fourier transform (FFT), DCT, Karhunen–Loeve transform (KLT), wavelet transform, etc.

III. Additive Spread Spectrum Watermarks

Additive spread spectrum watermarking systems, on which most of the proposed watermarking techniques are based, constitute a special case of the general model depicted in Section II. They are inspired by the spread spectrum modulation techniques employed for digital communications in jamming environments [43], [44], since data hiding can be seen as a communication problem in which the original image plays the role of channel noise and attackers may try to disrupt the transfer of information. In this kind of scheme a spread spectrum two-dimensional signal $\boldsymbol{w}$ (the watermark) carrying some hidden information is added to the original image $\boldsymbol{x}$.

The procedure followed to generate a watermark can be summarized as follows (Fig. 2). Suppose that we want to hide N bits of information and $\boldsymbol{x}$ has L elements. First, a pseudorandom sequence $\boldsymbol{s}$ is generated using a pseudonoise generator initialized to a state which depends on the value of K. To guarantee invisibility, this sequence is multiplied element by element by a perceptual mask $\boldsymbol{\alpha}$ obtained after analyzing the original image employing a psychovisual model. This mask takes care of the fact that alterations performed to different elements of $\boldsymbol{x}$ have different influences on the overall perceptual distortion. Next, the set of indexes $\{1, \ldots, L\}$ is partitioned into N subsets that we will denote by $\{\mathcal{S}_i\}_{i=1}^{N}$, satisfying $\mathcal{S}_i \cap \mathcal{S}_j = \emptyset$, $\forall i \neq j$. Each of these sets represents the elements of $\boldsymbol{x}$ in which a certain information bit is going to be hidden. The partition can in general be key dependent to provide an additional level of resilience to attacks directed against specific hidden information bits. Then, the final expression of the watermark $\boldsymbol{w}$ is

$$w_i = b_j \alpha_i s_i, \qquad \forall i \in \mathcal{S}_j \tag{6}$$

where j is any index in $\{1, \ldots, N\}$, and b_j is a coefficient used to encode the jth bit of the hidden message. Finally, the watermarked image $\boldsymbol{y}$ is obtained by adding the watermark to the original image. Using vector notation, the watermark $\boldsymbol{w}$ can be expressed as

$$\boldsymbol{w} = \boldsymbol{P}(K, \boldsymbol{x})\boldsymbol{b} \tag{7}$$

where $\boldsymbol{b} \triangleq (b_1, \ldots, b_N)^T$ encodes the hidden message and $\boldsymbol{P}(K, \boldsymbol{x})$ is an $L \times N$ matrix whose elements p_{ij} satisfy

$$p_{ij} = \begin{cases} \alpha_i s_i, & \text{if } i \in \mathcal{S}_j \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

The columns of this matrix, $\boldsymbol{P} = (\boldsymbol{p}_1 \ \ldots \ \boldsymbol{p}_N)$, will hereafter be called modulation pulses, since the watermark can be expressed as a linear combination of them and can thus be compared to multipulse amplitude modulation schemes used in communications [45], [46]. The resulting watermarked image is $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{w}$. Equation (6) indicates that in the watermark generation process the message vector $(b_1, \ldots, b_N)$, defined in a space with N dimensions, is mapped in a pseudorandom fashion onto a space with many more dimensions ($L \gg N$). This is what in communications theory is called "spreading the spectrum." The high degree of redundancy introduced in this transformation and the dependence of the mapping on the value of a secret key, only known to the copyright owner, are the facts that provide the robustness necessary to resist both unintentional alterations and malicious attacks.

IV. Watermark Detection and Decoding with Known Image Statistics

Given an image $\boldsymbol{z}$ under test and a key K, and assuming that the watermarked image has not suffered neither attacks nor unintentional distortions, the watermark detection test can be formulated as the binary hypothesis test

$$\begin{aligned} H_1&: \quad \boldsymbol{z} = \boldsymbol{x}_1 + \boldsymbol{P}(K, \boldsymbol{x}_1)\boldsymbol{b} \\ H_0&: \quad \boldsymbol{z} = \boldsymbol{x}_0 \end{aligned} \tag{9}$$

where $\boldsymbol{x}_1$ and $\boldsymbol{x}_0$ are images. The goal of the watermark detection test is to decide whether the image $\boldsymbol{z}$ contains a watermark generated by the copyright holder who possesses the key K. It is not necessary to decode the hidden message. Therefore, $\boldsymbol{b}$ must be regarded as a random vector with a probability mass function (pmf) equal to the probability distribution of the messages V. Even though K is known (it is the key under test), the matrix $\boldsymbol{P}$ is a function of the original image, which is supposed to be unknown. For each possible value of the message vector $\boldsymbol{b}$ there is a unique value of $\boldsymbol{x}$ satisfying the equation $\boldsymbol{z} = \boldsymbol{x} + \boldsymbol{P}(K, \boldsymbol{x})\boldsymbol{b}$. The functional relationship between $\boldsymbol{P}$ and $\boldsymbol{x}$ is usually quite complex, so it is in practice very difficult to know the set of possible original images that could have been mapped onto

$\boldsymbol{z}$ during the watermarking process. We can alternatively assume that $\boldsymbol{P}(K, \boldsymbol{x})$ is approximately the same for all the possible values of $\boldsymbol{b}$ and substitute $\boldsymbol{P}(K, \boldsymbol{x})$ by $\boldsymbol{P}(K, \boldsymbol{z})$. This is a reasonable approximation since the distortions introduced by the watermark are small, so they are expected to produce negligible alterations in the perceptual mask. Then, the test in (9) is approximately equivalent to the binary hypothesis test

$$\begin{aligned} H_1&: \quad \boldsymbol{z} = \boldsymbol{x}_1 + \boldsymbol{P}(K, \boldsymbol{z})\boldsymbol{b} \\ H_0&: \quad \boldsymbol{z} = \boldsymbol{x}_0 \end{aligned} \tag{10}$$

and the dependence of $\boldsymbol{P}$ on $\boldsymbol{x}$ has disappeared. Let $S \in \{H_1, H_0\}$ be the decision made in the watermark detection test (function d in Fig. 1). The probability of false alarm, defined as $P_F = \Pr\{S = H_1 | H_0\}$, will be required to lie below a certain maximum value to guarantee that the watermarking system is reliable. The design of the detector structure will be aimed at maximizing the probability of detection $P_D = \Pr\{S = H_1 | H_1\}$ which corresponds to the maximum allowable P_F.

In the watermark decoding test it is assumed that $\boldsymbol{z}$ does contain a watermark belonging to the copyright holder who possesses the secret key under test. For this reason, the decoding process is performed only if H_1 is decided in the watermark detection test. The goal of the decoder is to obtain an estimate $\hat{\boldsymbol{b}}$ of the message vector $\boldsymbol{b}$ in such a way that the probability of error is minimized. Following similar arguments as in the discussion on the watermark detection test, the matrix $\boldsymbol{P}(K, \boldsymbol{x})$ can be approximated by $\boldsymbol{P}(K, \boldsymbol{z})$.

Assuming that K is fixed, there are two random vectors in the statistical decision tests we have just formulated: $\boldsymbol{x}$ and $\boldsymbol{b}$. If the distribution $f_x(\boldsymbol{x})$ is known, the optimum decision tests are given by the Neyman–Pearson rule [47]. In the watermark detection test, for example, the optimum detector which maximizes the P_D conditioned to a given K for any value of P_F is given by the test

$$\ln \frac{f_z(\boldsymbol{z}|H_1, K)}{f_z(\boldsymbol{z}|H_0)} = \ln \sum_{\boldsymbol{b}} \frac{p(\boldsymbol{b}) f_x(\boldsymbol{z} - \boldsymbol{P}\boldsymbol{b})}{f_x(\boldsymbol{z})} \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{11}$$

where η is a threshold. If the messages are assumed to be equiprobable, then $p(\boldsymbol{b}) = 1/M, \forall \boldsymbol{b}$. Then, the optimum watermark decoder which minimizes the probability of error conditioned to K is given by the maximum likelihood structure. The output $\hat{\boldsymbol{b}}$ of the decoder is therefore

$$\hat{\boldsymbol{b}} = \arg \max_{\boldsymbol{b}} \ln f_z(\boldsymbol{z}|\boldsymbol{b}) = \arg \max_{\boldsymbol{b}} \ln f_x(\boldsymbol{z} - \boldsymbol{P}\boldsymbol{b}). \tag{12}$$

Expressions (11) and (12) define the detector and decoder functions $d(\boldsymbol{z}, K)$ and $e(\boldsymbol{z}, K)$, respectively, represented in Fig. 1.

Once these functions have been designed, it is interesting to study the influence of image characteristics in the performance. A measure of the goodness of an image $\boldsymbol{x}$ for watermarking purposes can be obtained by conditioning the probabilities P_e, P_F, and P_D to $\boldsymbol{x}$ and treating the secret key K and the message vector $\boldsymbol{b}$ as the only random variables in the system. In other words, we can regard a given original image $\boldsymbol{x}$ as a fixed deterministic vector and model $\boldsymbol{s}$ and the partition $\{\mathcal{S}_i\}_{i=1}^N$ statistically. This leads to a statistical model of the matrix $\boldsymbol{P}(\boldsymbol{x}, K)$, which can be used to compute the conditional probabilities of false alarm and detection

$$P_F(\boldsymbol{x}) = \Pr\{d(\boldsymbol{x}, K) = H_1\} \tag{13}$$

$$P_D(\boldsymbol{x}) = \Pr\{d(\boldsymbol{x} + \boldsymbol{P}(\boldsymbol{x}, K)\boldsymbol{b}, K) = H_1\} \tag{14}$$

and the conditional probability of error

$$P_e(\boldsymbol{x}) = \Pr\{e(\boldsymbol{x} + \boldsymbol{P}(\boldsymbol{x}, K)\boldsymbol{b}, K) \neq \boldsymbol{b}\}. \tag{15}$$

Given $\boldsymbol{x}$, $P_F(\boldsymbol{x})$ indicates the proportion of keys that, after being applied in the detection test performed to the original image $\boldsymbol{x}$, yield a positive result. The probability $P_D(\boldsymbol{x})$ gives the proportion of keys that, after being applied in the watermarking and detection processes, yield a positive result. The probability $P_e(\boldsymbol{x})$ indicates the proportion of keys for which a decoding error occurs when applied in the watermarking and decoding processes.

Usually the log-likelihood functions involved in the watermark detector and decoder can be expressed as a function of a vector $\boldsymbol{r} = (r_1, \ldots, r_{N'})$ of sufficient statistics. In this case, the conditional probabilities $P_e(\boldsymbol{x})$, $P_F(\boldsymbol{x})$, and $P_D(\boldsymbol{x})$ can be evaluated by studying the conditional distributions $f_r(\boldsymbol{r}|H_0, \boldsymbol{x})$ and $f_r(\boldsymbol{r}|H_1, \boldsymbol{b}, \boldsymbol{x})$.

V. Watermark Detection and Decoding with Unknown Image Statistics

In some cases the distribution $f_x(\boldsymbol{x})$ may be unknown or difficult to approximate. For example, there are no satisfactory statistical models for images in the spatial domain. The approach discussed above for the design of the watermark detector and decoder is not applicable in such a situation.

In Section III, we saw how in the watermark generation process the message vector $\boldsymbol{b}$ was mapped onto a vector with a much higher number of dimensions. Therefore, it is reasonable to apply a transformation $\boldsymbol{r} = h(K, \boldsymbol{z})$ to reduce the number of dimensions of the image under test $\boldsymbol{z}$ and then analyze the statistical decision problem in the transformed space. One such a transformation is, for instance

$$r_i = \sum_{j \in \mathcal{S}_i} \alpha_j s_j z_j, \qquad i \in \{1, \ldots, N\} \tag{16}$$

which is nothing but the correlation receiver applied in spread spectrum communications [43], [44]. Expressed in vector notation

$$\boldsymbol{r} = \boldsymbol{P}^T(K, \boldsymbol{z})\boldsymbol{z}. \tag{17}$$

In practice we will make some assumptions about the distribution of $\boldsymbol{x}$ to design $h(K, \boldsymbol{z})$. In some situations it is possible to exploit statistical properties such as ergodicity and quasi-stationarity to obtain estimates of the first- and second-order moments of $\boldsymbol{x}$. In those cases, even though the exact shape of the distribution of the original image is not known, at least means, variances, and cross covariances can be approximated and this information can help to

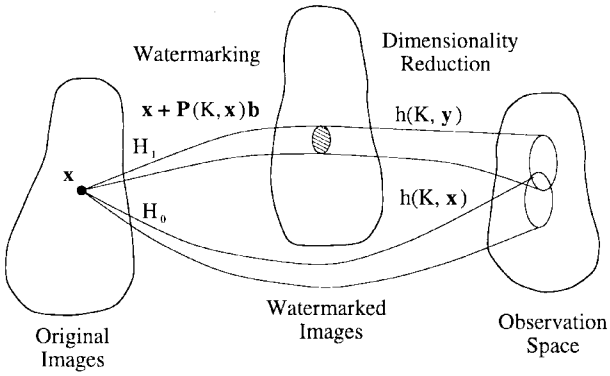


Fig. 3. Dimensionality reduction and watermark detection.

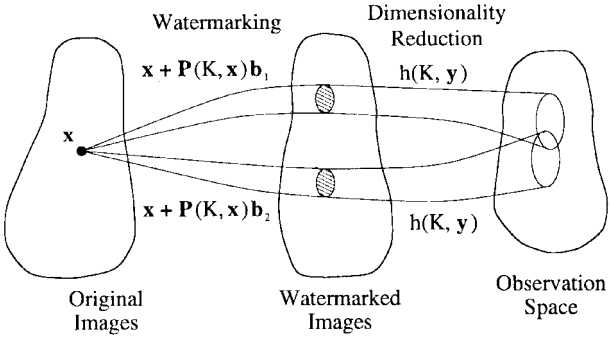


Fig. 4. Dimensionality reduction and watermark decoding.

improve substantially the performance associated with the dimensionality reduction transformation.

For instance, a key-independent linear minimum mean square error (MMSE) estimate of $\boldsymbol{w}$ can be computed before entering the correlation receiver. This filtering operation can considerably reduce the noise contribution due to the original image. Let $\boldsymbol{m}_x \triangleq E[\boldsymbol{x}]$ be the mean of $\boldsymbol{x}$ and let $\boldsymbol{R}_x$ be the covariance matrix of $\boldsymbol{x}$, defined as $\boldsymbol{R}_x \triangleq E[(\boldsymbol{x}-\boldsymbol{m}_x)(\boldsymbol{x}-\boldsymbol{m}_x)^T]$. If we assume that the watermarked image does not suffer any alteration, then $\boldsymbol{z}=\boldsymbol{x}+\boldsymbol{w}$. As we said before, $\boldsymbol{w}$ is actually a function of $\boldsymbol{x}$ since the perceptual mask is computed from it. We can, however, assume that the perceptual mask is approximately the same for all the possible values of $\boldsymbol{x}$ that can be mapped onto $\boldsymbol{z}$ after being watermarked with a certain pair K, $\boldsymbol{b}$. Thus, we can assume that $\boldsymbol{w}$ and $\boldsymbol{x}$ are statistically independent. Under this approximation, the optimum linear MMSE estimator of $\boldsymbol{w}$ is

$$\hat{\boldsymbol{w}} = \boldsymbol{R}_w(\boldsymbol{R}_x + \boldsymbol{R}_w)^{-1}(\boldsymbol{z}-\boldsymbol{m}_z) \tag{18}$$

where $\boldsymbol{m}_z$ is the mean value of $\boldsymbol{z}$. Then, the number of dimensions can be reduced applying the correlator receiver as we did before, obtaining as a result

$$\boldsymbol{r} = \boldsymbol{P}^T\hat{\boldsymbol{w}} = \boldsymbol{P}^T\boldsymbol{R}_w(\boldsymbol{R}_x + \boldsymbol{R}_w)^{-1}(\boldsymbol{z}-\boldsymbol{m}_z). \tag{19}$$

Note that in practice $\boldsymbol{R}_x$ will be estimated from $\boldsymbol{z}$ since $\boldsymbol{x}$ is not available in the watermark verification process. In fact, a reasonable approximation is $\boldsymbol{R}_x+\boldsymbol{R}_w \simeq \hat{\boldsymbol{R}}_z$. We can also estimate $\boldsymbol{R}_w$ from the image under test, since a good approximation to the perceptual mask $\boldsymbol{\alpha}$ can be obtained from $\boldsymbol{z}$.

Once the dimensionality reduction function $h(K, \boldsymbol{z})$ is defined, it is possible to study the statistical properties of its output $\boldsymbol{r}$ for a fixed original image $\boldsymbol{x}$, assuming that K is taken at random. Then we can make use of this statistical characterization to design optimum detector and decoder structures. In Fig. 3 we represent graphically the scenario corresponding to the watermark detection process. Suppose we choose a certain original image $\boldsymbol{x}$. If it is watermarked (hypothesis H_1) taking some pair K, $\boldsymbol{b}$ at random, the resulting vector $\boldsymbol{y}$ has a distribution that corresponds to the transformation of the distribution of K and $\boldsymbol{b}$ when the function $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{P}(K, \boldsymbol{x})\boldsymbol{b}$ is applied. After passing $\boldsymbol{y}$ through the dimensionality reduction function $h(K, \boldsymbol{y})$, the resulting observation vector $\boldsymbol{r}$ will have some conditional distribution $f_r(\boldsymbol{r}|H_1, \boldsymbol{x})$.

If $\boldsymbol{x}$ is not watermarked (hypothesis H_0), the application of the dimensionality reduction function will result in an observation vector $\boldsymbol{r} = h(K, \boldsymbol{x})$ with a conditional distribution $f_r(\boldsymbol{r}|H_0, \boldsymbol{x})$, which ideally can be evaluated applying the transformation $h(K, \boldsymbol{x})$ to the distribution of K [note that the function $h(K, \boldsymbol{x})$ is independent of $\boldsymbol{b}$]. Given an observed vector $\boldsymbol{r}$, the detector which maximizes the conditional probability of detection $P_D(\boldsymbol{x})$ for every value of the conditional probability of false alarm $P_F(\boldsymbol{x})$ is given by

$$\ln \frac{f_r(\boldsymbol{r}|H_1, \boldsymbol{x})}{f_r(\boldsymbol{r}|H_0, \boldsymbol{x})} \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{20}$$

where η is a threshold. Even though the detector depends on $\boldsymbol{x}$, which is not available during the watermark verification process, in practice only some "macroscopic" properties of $\boldsymbol{x}$ will be required, and it will be possible to estimate them from $\boldsymbol{z}$. Furthermore, certain families of functions $h(K, \boldsymbol{y})$ as, for example, those of the form $r_i = \sum_j h_{i,j}(K, y_j)$, allow the application of the central limit theorem to approximate $\boldsymbol{r}$ by a Gaussian distribution, since the elements w_i of the watermark are statistically independent if $\boldsymbol{s}$ is modeled as an independently identically distributed (i.i.d.) sequence. This kind of function appears, for instance, when the modulation pulses are sparsely spread and the image samples are assumed to be statistically independent if they are not too close.

A similar approach can be undertaken for the design of the watermark decoder. In this case the scenario is represented in Fig. 4. After watermarking $\boldsymbol{x}$ with a given vector message $\boldsymbol{b}$ and a random secret key K, the observation vector $\boldsymbol{r}$ has a distribution $f_r(\boldsymbol{r}|\boldsymbol{b}, \boldsymbol{x})$ that can be evaluated by applying the transformation $h(K, \boldsymbol{x} + \boldsymbol{P}(K, \boldsymbol{x})\boldsymbol{b})$ to the distribution of K. Therefore, given an observed vector $\boldsymbol{r} = h(K, \boldsymbol{z})$, the optimum decoder which minimizes the conditional probability of error $P_e(\boldsymbol{x})$ assuming that all codewords $\boldsymbol{b}$ have the same *a priori* probability is given by the maximum likelihood (ML) decoder

$$\hat{\boldsymbol{b}} = \arg\max_{\boldsymbol{b}} f_r(\boldsymbol{r}|\boldsymbol{b}, \boldsymbol{x}). \tag{21}$$

In general, different decoders are associated with different images $\boldsymbol{x}$, since the probability density functions (pdf's) are conditioned to $\boldsymbol{x}$. Even though the original image is assumed to be unknown during the decoding process, the decoder will actually depend on a few image-dependent parameters that can be estimated from the image $\boldsymbol{z}$ under test. The same kind of approximations resulting from the application of the central limit theorem can be made for certain functions $h(\boldsymbol{z}, K)$ as when we talked about the watermark detector. It is even possible that exploiting this kind of approximations a decoder independent of $\boldsymbol{x}$ results (see Section VIII).

VI. The Impact of Distortions and Attacks

In Sections IV and V, we have assumed that the watermarked image $\boldsymbol{y}$ did not suffer any alteration during distribution. However, in practice the watermarked image may be altered either on purpose or accidentally by linear filtering distortions, cropping, scaling, rotations, etc., and the watermarking system should still be able to detect and extract the watermark. The distortions are limited to those not producing excessive degradations, since otherwise the image would become unusable. Distortions and attacks introduce an additional transformation between watermarking and verification that changes the statistical distributions of $\boldsymbol{r}$ involved in the watermark detection and decoding tests. As a consequence, the performance of these tests can be degraded. Given a watermarking system with a certain structure, the goal of an attacker is to alter the image in such a way that it is not severely distorted and the distribution of $\boldsymbol{r}$ is transformed so that the probability of detection is decreased. The main obstacle the attacker must deal with is the uncertainty about the value of the secret key used by the copyright owner.

The ideal solution against attacks is the application of robust statistical decision theory [48], [49]. Instead of deriving the optimum watermark detector and decoder for certain distributions of the observation vector $\boldsymbol{r}$, robust detection, and decoding devices are designed to maximize the worst-case performance, associated with the worst-case attack. Robustness criteria can also be applied to the watermarking system as a whole, searching simultaneously the watermarking function f and the watermark decoder c and detector d (Fig. 1) such that the worst-case performance is optimized. However, it is difficult to model all the possible attacks that can appear in a practical situation. For this reason, the application of robust statistical theory to the design of watermarking schemes is an ambitious task that can hardly provide useful results.

The most harmful attack against an image watermarking system based on additive spread spectrum watermarks is that consisting in geometrical transformations such as scalings and rotations in the spatial domain. The sensitivity of the watermark detector to this kind of manipulation is due to the white nature of the pseudorandom sequence. The reason is that with such a sequence, a slight mismatch between the modulation pulses generated during the verification test and the ones actually present in the image produces drastic degradations in the distribution of $\boldsymbol{r}$ conditioned to H_1, so that it can even become indistinguishable from the distribution of $\boldsymbol{r}$ conditioned to H_0. A promising approach to achieve robustness against geometrical distortions is the use of transforms invariant to rotations and scalings [50].

VII. Error-Protection Codes

The performance of the watermark decoder can be improved if channel codes are used to encode the hidden messages carried by the watermark. Let L be the number of elements of $\boldsymbol{x}$ and $\{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_M\}$ the message vectors associated with the M possible messages that can be encoded by the watermark employing the generation mechanism explained in Section III. The watermarks that result after multiplying each of these vectors by the matrix $\boldsymbol{P}$ can be seen as points in the L-dimensional space $\mathbb{R}^L$. A channel encoder basically transforms these points into a different set of points in such a way that the distances between any two of them is increased [51]. Placing the watermarks farther from each other in the space helps to reduce the probability of error in the watermark decoding stage. In fact, the watermark generation procedure exposed in Section III can be seen as an encoding scheme that places the codewords in the subspace spanned by the vectors $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_N$.

One of the problems that image watermarking has to deal with is the extremely low signal-to-noise ratio (SNR) in each dimension of the L-dimensional space. In other words, the power of the original image, which is unknown to the decoder, is much stronger than the power of the watermark. As a consequence, an acceptable probability of error in watermark decoding can be achieved only by adding a large amount of redundancy during the encoding process. For this reason, the number of hidden message bits that can be embedded into an image is limited, depending on the size of the image (L) and its power.

Ideally, channel codes should be designed with as many degrees of freedom as dimensions are available. Unfortunately, this is a difficult task considering the low SNR per dimension that usually appears. A practical approach that can be undertaken is to use a block code or a convolutional code and increase the number of pulses so that the message length is left the same.

Channel codes have been successfully used in image watermarking systems [50]. The use of Bose Chaudhuri Hocquenghem (BCH) and Golay codes [51], [52] in the context of spatial-domain image watermarking and the influence of parameters such as the redundancy and the minimum distance in the performance of the watermark decoder as well as the watermark detection test are studied in [53]. Promising results, not published, have also been obtained for convolutional codes.

VIII. Spatial Domain Watermarking

Let us now apply the ideas exposed in Sections IV and V to the analysis of spread spectrum watermarking of images defined in the spatial domain. We will thus assume in this

section that $\boldsymbol{x}$ represents the luminance component of a digitized image. Although we will concentrate our attention on grey-level images for simplicity, our derivations can be easily extended to color images by including in the image mathematical model three vectors, each associated with one of the three components in a luminance and color differences representation.

Unfortunately, there are no statistical distributions suitable for modeling the luminance component of common images in the spatial domain [54]. Without a satisfactory statistical model for the original image $\boldsymbol{x}$, we cannot apply the decision theory as described in Section IV to the design of the optimal watermark detector and decoder structures that optimize the performance conditioned to each value of the secret key (the performance that each copyright owner sees).

However, we can project the image $\boldsymbol{z}$ under test onto a subspace and apply the design techniques presented in Section V. An interesting candidate among the transformations that can be used to reduce the number of dimensions is the correlation receiver $\boldsymbol{r} = \boldsymbol{P}^T\boldsymbol{z}$ already discussed in Section V. The correlation receiver provides sufficient statistics for both the watermark detection and the watermark decoding problems when the watermark $\boldsymbol{w}$ is immersed in zero-mean white Gaussian noise. For this reason, this is a reasonably good choice.

Images found in practical situations are nonstationary in the spatial domain since the broad range of objects that can be represented in the same image may result in considerable variations in statistical properties of luminance samples along the image. Nevertheless, in most cases the aforementioned statistical properties do not substantially differ in adjacent pixels. In other words, the image $\boldsymbol{x}$ can be approximated as a quasi-stationary random process. If we also assume that it is ergodic, then we can estimate the first and second moments of the original image at each pixel by computing averages in block neighborhoods. As we said before, the original image is not available in the copyright verification process, so these statistics must be calculated from the image intended to be tested.

As discussed in Section V, knowledge about the first- and second-order moments of the original image can be used to improve the performance of detection and decoding when done in the projection subspace, even if this knowledge is an approximation to the actual values. A Wiener filter, for instance, can be used to obtain a linear MMSE estimate of the watermark. This estimate eliminates part of the original image component before the projection process, thus improving the achievable performance in detection and decoding.

We will also assume that the watermarked image $\boldsymbol{y}$ may have been distorted by a linear filter, either as a consequence of alterations occurring during distribution or as a result of attacks aimed at destroying or corrupting the watermark. This filter can be combined with the Wiener filter discussed above to form an equivalent linear system represented by an $L \times L$ matrix that will be called $\boldsymbol{H}$. Therefore, we will find the signal $\boldsymbol{z} = \boldsymbol{H}\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{H}\boldsymbol{P}\boldsymbol{b}$ at the input of the correlation receiver. After being projected, we obtain the vector

$$\boldsymbol{r} = \boldsymbol{P}^T\boldsymbol{z} = \boldsymbol{P}^T\boldsymbol{H}\boldsymbol{x} + \boldsymbol{P}^T\boldsymbol{H}\boldsymbol{P}\boldsymbol{b}. \tag{22}$$

Recall that K is the only random element in our model. Hence, $\boldsymbol{x}$ is deterministic and the matrix $\boldsymbol{P}$ is random. The coefficients of the Wiener filter are computed from the image under test and depend, therefore, on the value of K. This fact implies that the filter $\boldsymbol{H}$ is actually random and statistically dependent on $\boldsymbol{P}$. However, considering the small alterations that watermarks produce, the variability experienced by the filter coefficients can be expected to be small. Thus, it is legitimate to assume that $\boldsymbol{H}$ is deterministic and independent of $\boldsymbol{P}$. After this approximation, only matrix $\boldsymbol{P}$ is random in (22).

To obtain a statistical characterization of $\boldsymbol{r}$, we need to define a model for the matrix $\boldsymbol{P}$. As indicated in Section III, the columns of this matrix are the pulses $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_N$ that compose the watermark. Since these pulses are nonoverlapping, each row of $\boldsymbol{P}$ has only one nonzero element for every value of K. Furthermore, if we take the ith row, the value of the nonzero element is $\alpha_i s_i$. The pseudorandom sequence $\boldsymbol{s}$ is key dependent and thus must be treated as a random vector. In order to simplify the discussion, we will model it as L outcomes of an i.i.d. random process with a discrete marginal distribution with two equiprobable levels $\{-1, 1\}$. The results given below can be straightforwardly extended to any kind of marginal distribution.

We will assume that the pulses $\{\boldsymbol{p}_i\}_{i=1}^N$ are sparsely scattered over the whole image in a key-dependent pseudorandom fashion to provide diversity that strengthens the robustness to attacks directed against particular bits. We also assume that the watermark covers all the pixels of the image. Then, the sets $\{\mathcal{S}_i\}_{i=1}^N$ constitute a partition of the set $\{1, \ldots, L\}$, so each image pixel is assigned to only one of those sets. This pixel assignment mechanism must be modeled as a random procedure since it is key dependent. We will assume that every index $j \in \{1, \ldots, L\}$ is included in any of the sets $\{\mathcal{S}_i\}_{i=1}^N$ with the same probability $1/N$ and that the assignment is performed independently for each index. As a consequence, for every value of K, any row of $\boldsymbol{P}$ has one and only one nonzero element, which belongs to any of the columns $1, \ldots, N$ with probability $1/N$. Under these assumptions, and after some algebraic manipulations, the first- and second-order moments of the elements of $\boldsymbol{r}$ conditioned to an original image $\boldsymbol{x}$ and a message vector $\boldsymbol{b}$ can be shown to be [45], [55]

$$E[r_i|H_1, \boldsymbol{b}, \boldsymbol{x}] = b_i \frac{1}{N} \sum_{k=1}^{L} h_{k,k}\alpha_k^2 \tag{23}$$

$$\begin{aligned}\mathrm{Var}(r_i|H_1, \boldsymbol{b}, \boldsymbol{x}) &= \frac{1}{N}\sum_{k=1}^{L}\alpha_k^2 x_{f_k}^2 + \frac{b_i^2}{N}\left(E[s^4]-1\right)\sum_{k=1}^{L} h_{k,k}^2\alpha_k^4 \\ &\quad + \frac{\sum_j b_j^2}{N^2}\sum_{k=1}^{L}\sum_{l\neq k} h_{k,l}^2\alpha_k^2\alpha_l^2 + b_i^2\frac{N-1}{N^2}\sum_{k=1}^{L} h_{k,k}^2\alpha_k^4\end{aligned} \tag{24}$$

$$\mathrm{Cov}(r_i, r_j|H_1, \boldsymbol{b}, \boldsymbol{x}) = -b_i b_j \frac{1}{N^2} \sum_{k=1}^{L} h_{k,k}^2 \alpha_k^4, \qquad i \neq j \quad (25)$$

where $h_{i,j}$ are the elements of $\boldsymbol{H}$ and $\boldsymbol{x}_f \triangleq \boldsymbol{H}\boldsymbol{x}$. In practical situations the cross-covariance terms are negligible compared to the terms in the diagonal of the covariance matrix. Hence, we can assume that the elements of $\boldsymbol{r}$ are approximately uncorrelated.

Since every modulation pulse is sparsely spread out over the whole image, if the kernel of the filter $\boldsymbol{H}$ for every element x_i is small compared to the image size, i.e., if every row of $\boldsymbol{H}$ has only a few nonzero elements, then the elements of $\boldsymbol{r}$ can be expressed as a sum of statistically independent terms. The number of terms summed up, L/N on average, is in practice large since a high level of redundancy is necessary if an acceptable performance is desired. Hence, we can apply the central limit theorem and assume that $\boldsymbol{r}$ is approximately Gaussian.

Thus we have come up with a Gaussian model for the observed vector $\boldsymbol{r}$ that can be exploited to obtain detector and decoder structures.

A. Watermark Decoder

The optimum ML watermark decoder, derived in Section V, is given by (21), where $f_r(\boldsymbol{r}|\boldsymbol{x}, \boldsymbol{b})$ is the distribution at which we have just arrived. If we observe (23) and (24), assuming that $b_i \in \{-1, 1\}$, $\forall i \in \{1, \ldots, N\}$ and that the covariance matrix is approximately diagonal, we can infer that the observation vector $\boldsymbol{r}$ can be modeled as the output of an additive white Gaussian noise (AWGN) channel, $r_i = ab_i + n_i$, $i \in \{1, \ldots, N\}$, where

$$a = \frac{1}{N} \sum_{k=1}^{L} h_{k,k} \alpha_k^2 \quad (26)$$

and $n_1, \ldots, n_N$ are samples of an i.i.d. zero-mean Gaussian random process with variance

$$\sigma^2 = \frac{1}{N} \sum_{k=1}^{L} \alpha_k^2 x_{f_k}^2 + \frac{1}{N} (E[s^4] - 1) \sum_{k=1}^{L} h_{k,k}^2 \alpha_k^4 + \frac{1}{N} \sum_{k=1}^{L} \sum_{l \neq k} h_{k,l}^2 \alpha_k^2 \alpha_l^2 + \frac{N-1}{N^2} \sum_{k=1}^{L} h_{k,k}^2 \alpha_k^4. \quad (27)$$

We know from communication theory that the optimum ML decoder for an AWGN channel seeks the message vector $\boldsymbol{b}$ closest to the observation vector $\boldsymbol{r}$ in the Euclidean distance sense. Therefore, this decoder structure minimizes the probability of error conditioned to the original image $\boldsymbol{x}$. In other words, given some original image $\boldsymbol{x}$, this detector minimizes the chances that the key under test yields an error while extracting the hidden message.

When a binary antipodal constellation is used to encode $M = 2^N$ possible messages, i.e., when all possible combinations of N elements taken from $\{-1, 1\}$ are valid message vectors, the minimum Euclidean distance decoder is equivalent to a bit-by-bit hard decisor with the decision threshold located at the origin. Then, the output of the decoder is

$$\hat{b}_i = \mathrm{sign}(r_i), \qquad i \in \{1, \ldots, N\} \quad (28)$$

and the probability of making an error when decoding a bit [also known as bit error rate (BER)] is[1]

$$P_b = Q\left(\frac{a}{\sigma}\right) \quad (29)$$

which can be easily computed from the channel parameters.

B. Watermark Detector

The optimum watermark detector, whose structure has been already derived in Section V, is given by (20). An equivalent expression for this Neyman–Pearson test is

$$\ln \sum_{\boldsymbol{b}} p(\boldsymbol{b}) \frac{f_r(\boldsymbol{r}|H_1, \boldsymbol{x}, \boldsymbol{b})}{f_r(\boldsymbol{r}|H_0, \boldsymbol{x})} \underset{H_0}{\overset{H_1}{\gtrless}} \eta \quad (30)$$

where $f_r(\boldsymbol{r}|H_1, \boldsymbol{x}, \boldsymbol{b})$ is the Gaussian pdf that has just been derived in the analysis of the ML watermark decoder. In order to obtain the pdf of $\boldsymbol{r}$ under hypothesis H_0, i.e., when $\boldsymbol{x}$ is not watermarked, let us suppose that $\boldsymbol{x}$ instead of $\boldsymbol{y}$ is tested. Then, the observation vector is

$$\boldsymbol{r} = \boldsymbol{P}^T \boldsymbol{H} \boldsymbol{x}. \quad (31)$$

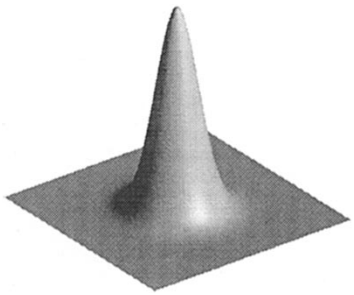
In this case, it can be easily shown that $\boldsymbol{r}$ is zero mean, white, and the variance of its elements is

$$\sigma_0^2 = \frac{1}{N} \sum_{k=1}^{L} \alpha_k^2 x_{f_k}^2. \quad (32)$$

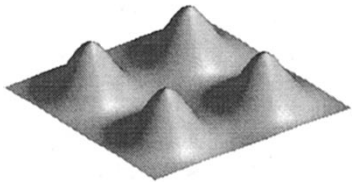
Every element of $\boldsymbol{r}$ can still be expressed as a sum of independent random variables. Therefore, under hypothesis H_0, the observation vector can be accurately approximated by a zero-mean white Gaussian vector with variance given by (32).

In Fig. 5 we have represented graphically an example of the Gaussian distributions conditioned to the hypothesis H_0 and H_1 when $N = 2$ and there are four equiprobable messages. Given an observation $\boldsymbol{r}$, the watermark detector must decide to which of these distributions $\boldsymbol{r}$ belongs. Suppose that N_s pulses, e.g., $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_{N_s}$, are modulated by message-independent known coefficients. These reserved pulses can be used to improve the performance of the watermark detection process since the uncertainty about the possible message vectors that the watermark may carry is thus reduced. Assume also that the remaining pulses are modulated by coefficients in a binary antipodal constellation with $N - N_s$ dimensions (hence, $M = 2^{N-N_s}$). Then, the logarithmic function in (30), which will be denoted by $l(\boldsymbol{r})$,

[1] $Q(x) \triangleq (1/\sqrt{2\pi}) \int_x^\infty e^{-t^2/2}\, dt$.



(a)



(b)

Fig. 5. PDF's involved in the watermark detection problem.

can be expressed as [45]

$$l(\boldsymbol{r}) = N \ln \frac{\sigma_0}{\sigma} - \frac{a^2 N}{2\sigma^2} - \frac{1}{2}\left(\frac{1}{\sigma^2} - \frac{1}{\sigma_0^2}\right)\sum_{i=1}^{N} r_i^2 + \frac{a}{\sigma^2}\sum_{i=1}^{N_s} r_i + \sum_{i=N_s+1}^{N} \ln\left(\cosh\left(\frac{a r_i}{\sigma^2}\right)\right) \underset{H_0}{\overset{H_1}{\gtrless}} \eta. \tag{33}$$

The probability of false alarm $P_F(\boldsymbol{x})$ indicates the chance that, given a certain nonwatermarked image $\boldsymbol{x}$, the key under test yields a positive result in the watermark detection test when applied to $\boldsymbol{x}$. On the other hand, the probability of detection $P_D(\boldsymbol{x})$ indicates the chance that, given a certain original image $\boldsymbol{x}$, the secret key under test yields a positive result in the watermark detection test when applied to watermark $\boldsymbol{x}$ and to detect the presence of the watermark. Although there is no close form expression for the probabilities of false alarm and detection conditioned to the original image $\boldsymbol{x}$ as a function of the threshold η, it is possible, however, to obtain Chernoff bounds [47]. These bounds, as well as approximations, were derived in [45].

C. Attacks

Attacks suffered by the watermarked image $\boldsymbol{y}$ affect the channel parameters a and σ, thus altering the achievable performance in both the watermark detection and decoding tests. Linear filtering attacks are already included in the model we have assumed when we began the analysis. Their impact in performance can be studied by assigning values to the coefficients of the matrix $\boldsymbol{H}$. The effect of attacks in which the image is cropped, so that some watermark energy is lost, can be studied by taking out from summations in (26) and (27) those terms whose index k corresponds to pixels that do not survive the attack. Undoubtedly, the most harmful kind of attack is that consisting in geometrical transformations of the image in the spatial domain [56]. If the watermarked image is either scaled or rotated, there will be a mismatch between the modulation pulses generated during the projection process and the pulses that are actually in the image under test. Given the white nature of the pseudorandom sequence from which these pulses are generated, we can expect a rapid degradation of the equivalent channel parameters in terms of the signal to noise ratio (defined as $20 \log a/\sigma$) as we increment or decrement the rotation angle or the scaling factor.

A countermeasure to weaken the effect of attacks based on geometrical transformations is the design of a spatial synchronization algorithm that estimates the transformation suffered by the image. Since the original image is not available, only knowledge about the watermark can be exploited to recover the size and orientation that it had before the attack. Let $\boldsymbol{\xi}$ be a vector of parameters defining the geometrical transformation that was performed by the attacker. An estimate of $\boldsymbol{\xi}$ can be obtained if we search the vector of parameters that, after being applied to transform the modulation pulses locally generated in the projection process, maximizes the log-likelihood function in (30). In mathematical notation [45]

$$\hat{\boldsymbol{\xi}} = \arg\max_{\boldsymbol{\xi}'} \ln \sum_{\boldsymbol{b}} p(\boldsymbol{b}) \frac{f_r(\boldsymbol{r}|H_1, \boldsymbol{x}, \boldsymbol{b}, \boldsymbol{\xi}')}{f_r(\boldsymbol{r}|H_0, \boldsymbol{x}, \boldsymbol{\xi}')}. \tag{34}$$

However, this exhaustive search technique is impractical due to the narrowness of the peak of the log-likelihood function, which is a consequence of the white nature of the watermark. Resilience to scaling and rotation distortions is still in fact a challenging problem in image watermarking.

D. Experimental Results

To verify the validity of the model described in previous sections, analytical results have been contrasted with empirical data obtained through experimentation. We have used five test images, shown in Fig. 6. These images were chosen for their different characteristics in terms of flat areas, noisy textures, etc. In Fig. 7 we show an example of a watermark and a watermarked version for "Lena."

The perceptual model we have chosen for the experimental work in spatial domain watermarking is described in [54], [57], and [58] and exploits the spatial masking properties of the human visual system (HVS). The perceptual mask $\boldsymbol{\alpha}$, obtained after the analysis of the original image, indicates the maximum allowable standard deviation of noisy alterations at each pixel. The pseudorandom sequence $\boldsymbol{s}$ is assumed to have a discrete marginal distribution with two levels, $\{-1, 1\}$. This means that $\boldsymbol{s}$ has unit variance at every pixel, so if it is multiplied element by element by $\boldsymbol{\alpha}$ the invisibility constraint will be satisfied.

In our experiments, a spatially variant Wiener filter is used before the correlation receiver to eliminate part of the noise contribution due to the original image. The coefficients of this filter are computed using (18) under the assumption that the original image is white, i.e., that $\boldsymbol{R}_x$ is a diagonal matrix. Let $\boldsymbol{C}$ denote the matrix associated with the Wiener filtering operation. If we assume that the watermarked image has not suffered any alteration, then

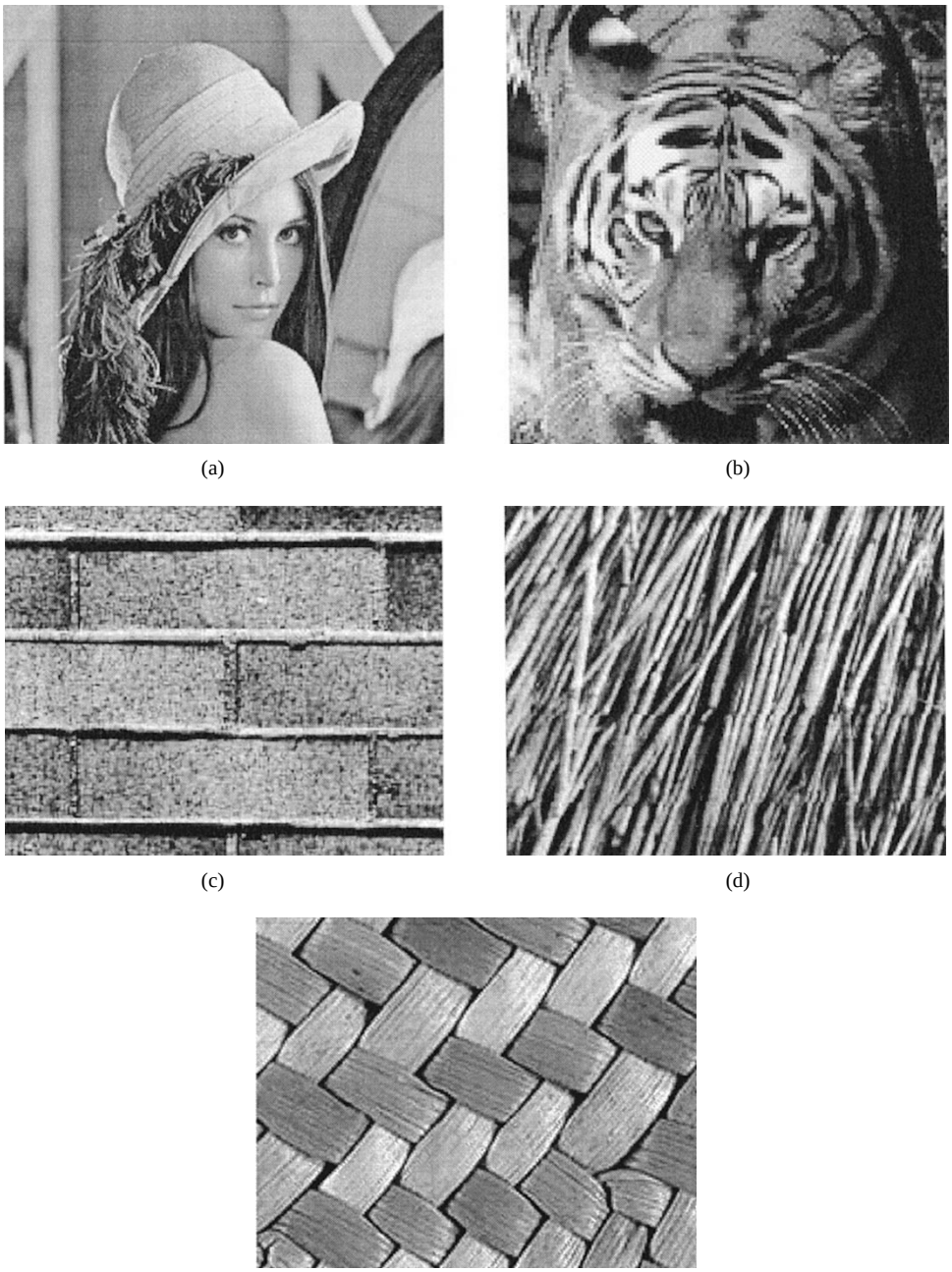


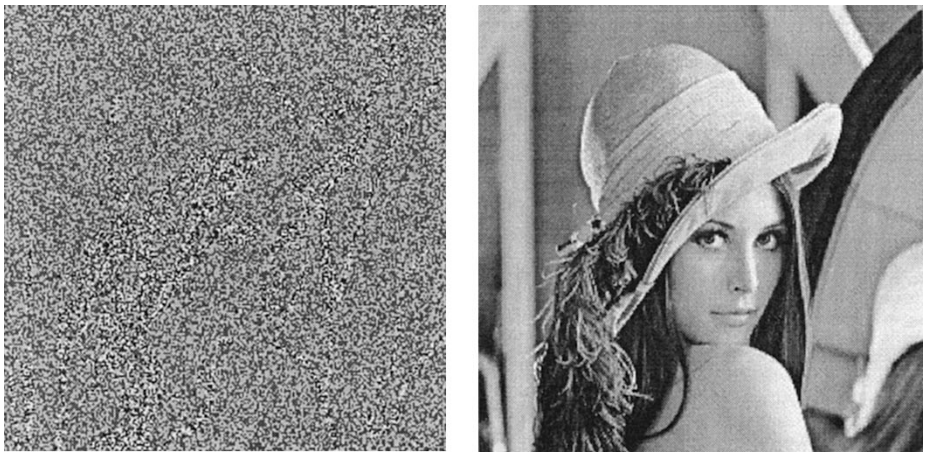
Fig. 6. Original images used in the experiments.

$\boldsymbol{R}_z = \boldsymbol{R}_x + \boldsymbol{R}_w$ since $\boldsymbol{x}$ and $\boldsymbol{w}$ are uncorrelated. Hence, the Wiener filter can be expressed as $\boldsymbol{C} = \boldsymbol{R}_w \boldsymbol{R}_z^{-1}$, whose coefficients are

$$c_{i,j} = \begin{cases} \dfrac{\alpha_i^2}{\sigma_{z_i}^2}, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases} \tag{35}$$

where $\sigma_{z_i}^2$ is the variance of z_i, which can be estimated from the pixels in a block neighborhood around z_i. The mean vector $\boldsymbol{m}_z$ in (18) can be similarly estimated from $\boldsymbol{z}$.

In Figs. 8–10 we show plots of the BER associated to the watermark decoder derived in Section VIII-A as a function of the average size of the modulation pulses. In all cases the empirical measures have been obtained by taking 100 keys at random. The first figure represents the BER when the watermarked image is not altered during distribution. In Fig. 9 we have plotted the BER when an attacker adds to the watermarked image Gaussian noise whose variance at each pixel is shaped by the perceptual mask $\boldsymbol{\alpha}$ so that the perceptual distortion is minimized. Fig. 10 shows the BER that results when the image is distorted by a Wiener filter aimed at obtaining an estimate of the original image so that the watermark is partially destroyed. We can see that the analytical approximations are reasonably close to the empirical results. The dependence of the performance on the image characteristics is also evident from the plots. The shift of theoretical curves with respect to the empirical



(a) (b)

Fig. 7. Example of (a) a watermark and (b) a watermarked version for "Lena" (256 × 256).

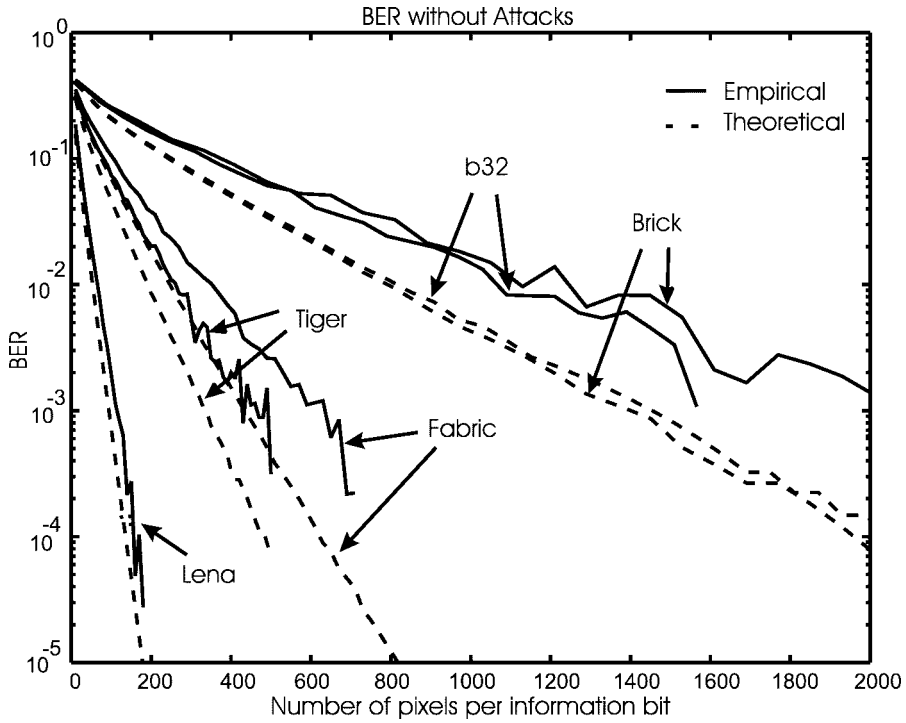


Fig. 8. BER without attacks.

data is due to the watermark-dependent nature of the Wiener filter coefficients, which has not been statistically modeled.

We have also measured through experimentation the performance of the watermark detector derived in Section VIII-B and contrasted the empirical data with the analytical results. Measures have been obtained by taking 400 keys at random. In all cases the watermark carries 240 bits of hidden information. In Figs. 11–15 plots of the receiver operating characteristic (ROC) are shown for all the test images. The empirical curves actually represent the experimentally measured P_D versus the Chernoff bound for the P_F because the values of P_F in the range of thresholds in which P_D begins to fall down are so small that they cannot be estimated through simulations. For a fair comparison of the performance results, all the images have been cropped down to a size of 128 × 128 pixels. The curves show that the Chernoff bound provides a fairly good approximation of the ROC. Comparing the figures we can see that performance of the watermark detection test clearly depends on the characteristics of the image contents. Note for instance the difference between the ROC of "Lena," an image with many flat regions, and "Brick" or "b32," both with more noisy textures.

IX. DCT Domain Watermarking

In this section we will assume that the vector $\boldsymbol{x}$ is the DCT of the luminance component of the original image, applied in blocks of 8 × 8 pixels, as in the JPEG algorithm. Then, $\boldsymbol{x}$ can be splitted into 64 vectors, each gathering

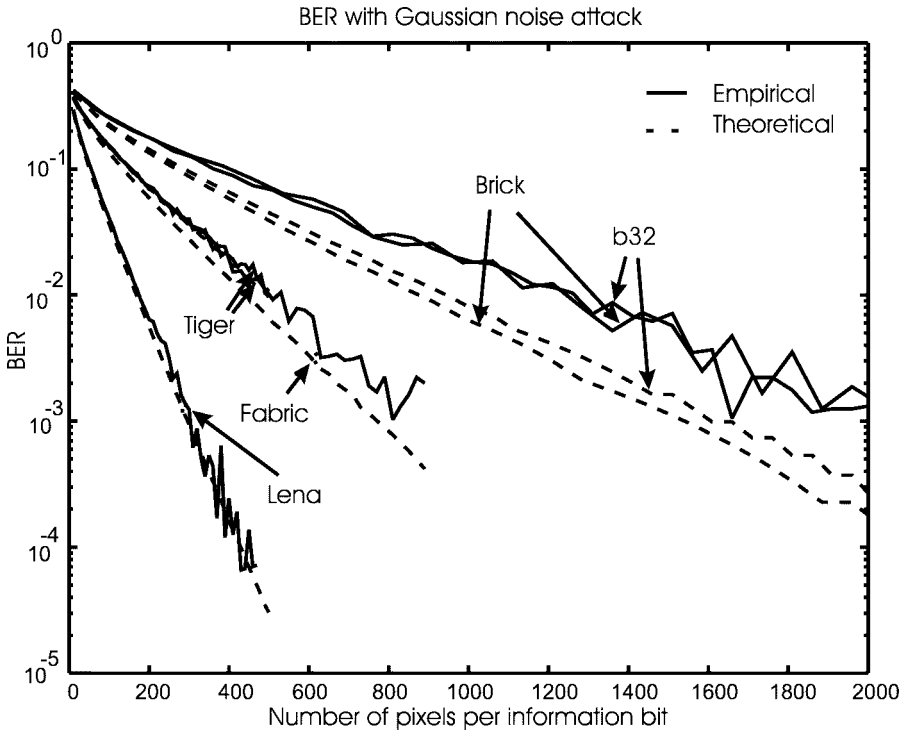


Fig. 9. BER with additive Gaussian noise attack.

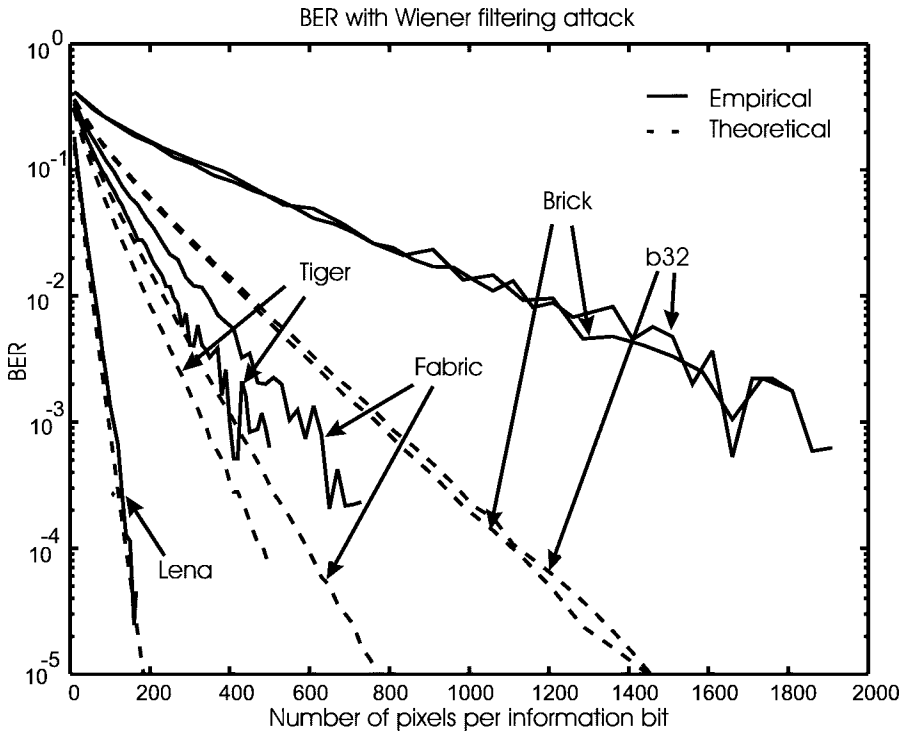


Fig. 10. BER with Wiener filtering attack.

the elements that correspond to one of the 8×8 DCT coefficients.

A. Statistical Model for the DCT Coefficients

The DCT coefficients of common images have interesting properties that can be exploited to obtain good watermark detectors and decoders. One of the most interesting characteristics of the DCT is energy compaction. In fact, it has been proved that the DCT converges to the Karhunen–Loéve transform (KLT) for images that can be statistically modeled as first-order Markov processes with a correlation factor close to one [59]. As a consequence, we can assume that the DCT coefficients are uncorrelated.

Another interesting property is that the histograms of the ac coefficients can be better approximated than luminance samples in the spatial domain by analytical expressions of

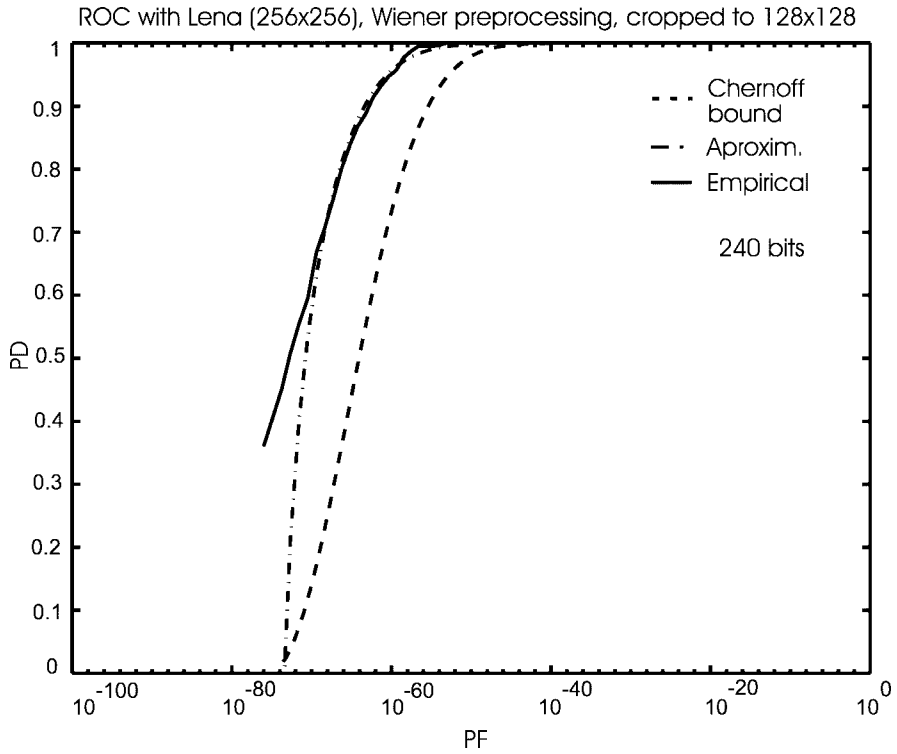


Fig. 11. ROC for "Lena" (256 × 256), cropped down to 128 × 128 pixels, with $N = 240$, $N_s = 0$.

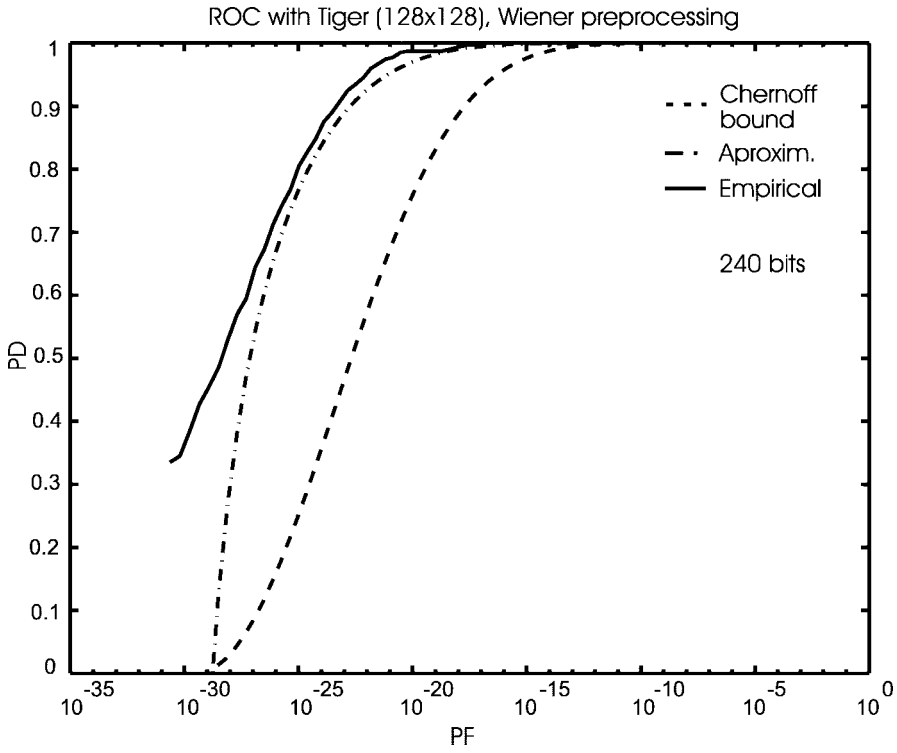


Fig. 12. ROC for "Tiger" (128 × 128), with $N = 240$, $N_s = 0$.

known pdf's. An early proposal as a statistical description of the ac coefficients is the Gaussian model. However, more recent studies have shown that a more accurate approximation is the generalized Gaussian pdf, given by the expression [60]

$$f_x(x) = Ae^{-|\beta x|^c} \tag{36}$$

where both A and β can be expressed as a function of c and the standard deviation σ

$$\beta = \frac{1}{\sigma}\left(\frac{\Gamma(3/c)}{\Gamma(1/c)}\right)^{1/2} \tag{37}$$

$$A = \frac{\beta c}{2\Gamma(1/c)}. \tag{38}$$

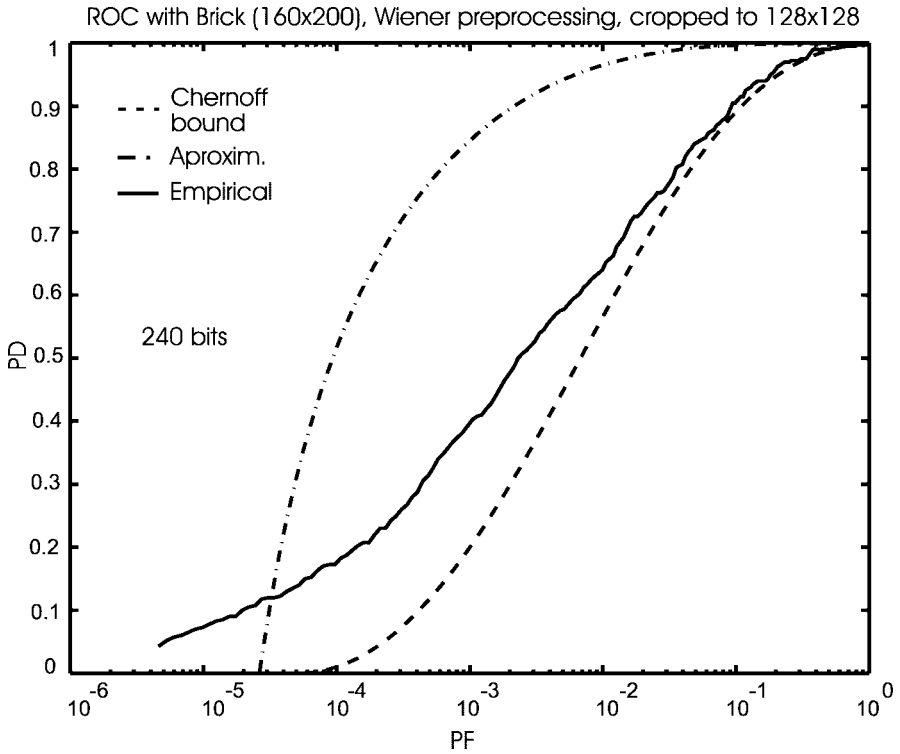


Fig. 13. ROC for "Brick" (160 × 200), cropped down to 128 × 128 pixels, with $N = 240$, $N_s = 0$.

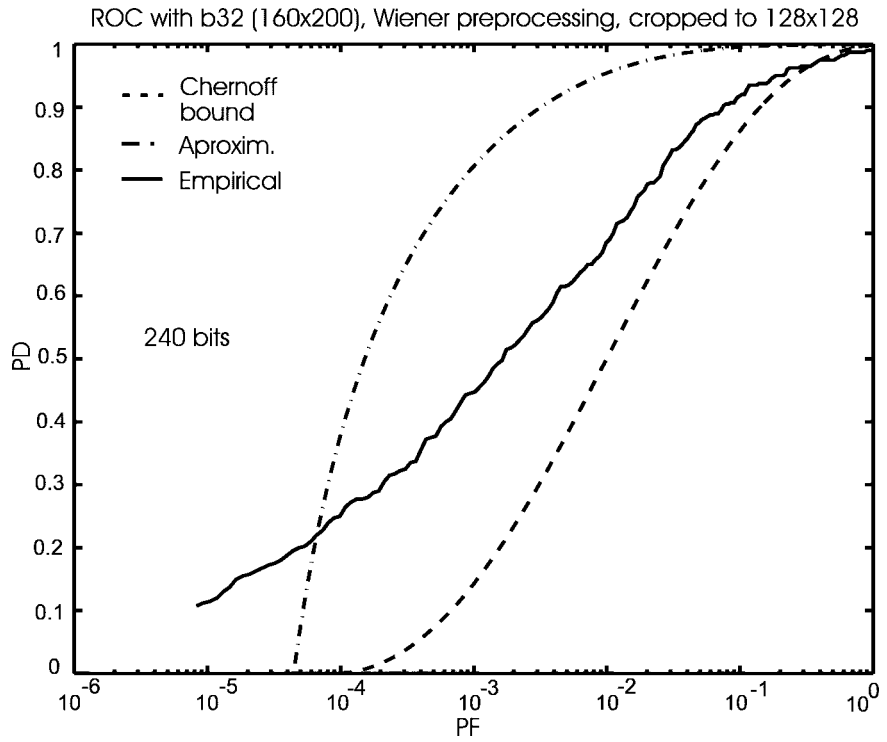


Fig. 14. ROC for "b32" (160 × 200), cropped down to 128 × 128 pixels, with $N = 240$, $N_s = 0$.

Note that the Gaussian as well as the Laplace distributions are just special cases of this pdf, given by $c = 2$ and $c = 1$, respectively. It turns out that coefficients in the low-frequency range are reasonably well modeled by a generalized Gaussian distribution with $c = 1/2$ and sometimes by a Laplace distribution ($c = 1$). Coefficients at high frequencies, however, are in many cases better modeled by a Laplace distribution and sometimes even by a Gaussian distribution [60]. What seems to be clear is that the Gaussian model is not a good model for DCT coefficients in most cases, especially at low and medium frequencies.

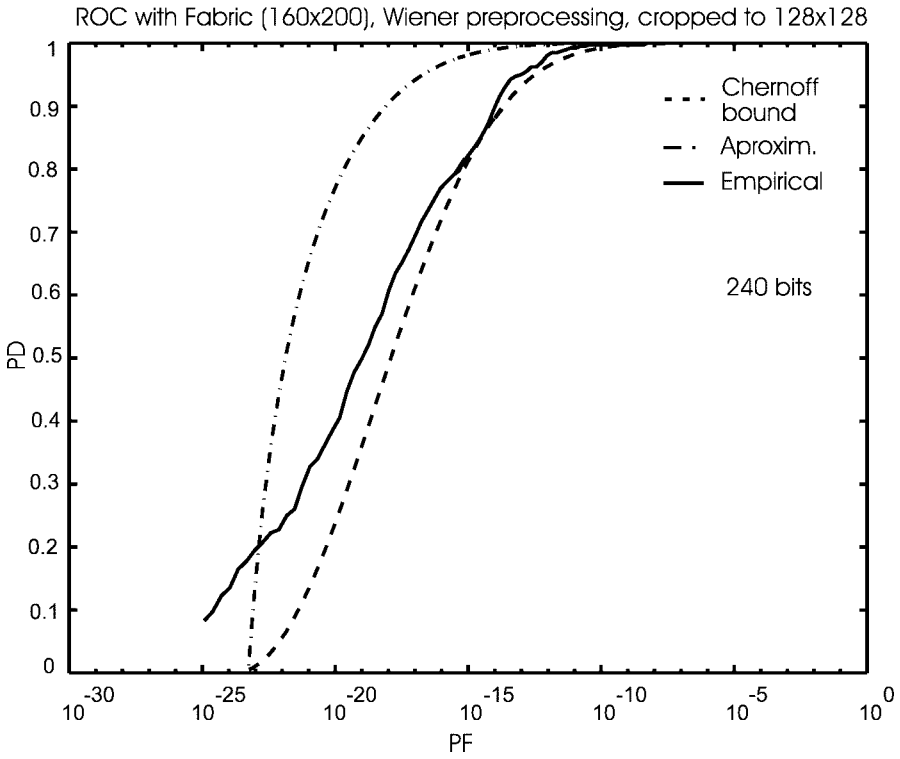


Fig. 15. ROC for "Fabric" (160 × 200), cropped down to 128 × 128 pixels, with $N = 240$, $N_s = 0$.

We will assume that the DCT coefficients are statistically independent, even though this property is not necessarily implied from the fact that the DCT coefficients are approximately uncorrelated, considering that a non-Gaussian model is more accurate as an approximation to their distribution. It will also be assumed that samples of DCT coefficients at the same frequency, in different blocks, are statistically independent. Thus, we will model the elements of $\boldsymbol{x}$ associated with the same DCT coefficient as outcomes from an i.i.d. random process with a generalized Gaussian marginal pdf. The parameters c and σ, which completely specify the distribution, can be different for each DCT coefficient. We will denote by c_i and σ_i the values of such parameters for the DCT coefficient to which the element x_i belongs.

The statistical characterization of the DCT coefficients of the original image is an invaluable help for the design of satisfactory watermark detectors and decoders in terms of performance. Since the original image is unknown, the parameters defining the distribution of its DCT coefficients must be estimated from the image under test. Given the small alterations produced during the watermarking process due to the limitations imposed by the invisibility constraint, and considering that the watermark can also be modeled statistically, fairly good estimates of the distribution parameters can be obtained in practice.

B. Watermark Decoder

Let $\mathcal{B} \triangleq \{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_M\}$ be the message vectors that correspond to the M possible hidden messages. Let us also define the watermark obtained from each of these vectors as $\boldsymbol{w_l} = \boldsymbol{P}\boldsymbol{b}_l$, $l \in \{1, \ldots, M\}$, whose elements will be denoted by $(w_{l1}, \ldots, w_{lN})$. The decoder that minimizes the probability of error conditioned to a given value of the secret key, assuming that all the codewords $\{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_M\}$ have the same *a priori* probability (i.e., the ML decoder), is the one that seeks the message vector $\boldsymbol{b_l}$ satisfying

$$\ln \frac{f_z(\boldsymbol{z}|\boldsymbol{b_l})}{f_z(\boldsymbol{z}|\boldsymbol{b}_m)} = \ln \frac{f_x(\boldsymbol{z}-\boldsymbol{w_l})}{f_x(\boldsymbol{z}-\boldsymbol{w}_m)} > 0, \qquad \forall m \neq l. \quad (39)$$

Using (36) and (37), and given that the elements of $\boldsymbol{x}$ are assumed to be independent, this is equivalent to

$$\sum_{i=1}^{L} \frac{|z_i - w_{m,i}|^{c_i} - |z_i - w_{l,i}|^{c_i}}{\sigma_i^{c_i}} > 0, \qquad \forall m \neq l. \quad (40)$$

It can be shown that, assuming that the message vectors verify $b_{l,i} \in \{-1, 1\}$, $\forall l \in \{1, \ldots, M\}$, $i \in \{1, \ldots, N\}$, the expressions

$$r_i \triangleq \sum_{k \in \mathcal{S}_i} \frac{|z_k + \alpha_k s_k|^{c_k} - |z_k - \alpha_k s_k|^{c_k}}{\sigma_k^{c_k}} \quad (41)$$

are sufficient statistics for the hidden information decoding problem, and the ML decoder is

$$\hat{\boldsymbol{b}} = \arg \max_{\boldsymbol{b} \in \mathcal{B}} \boldsymbol{b}^T \boldsymbol{r} \quad (42)$$

where $\boldsymbol{r} = (r_1, \ldots, r_N)^T$. When message vectors form a binary antipodal constellation, in which all possible combinations of N elements in $\{-1, 1\}$ are valid codewords representing $M = 2^N$ different messages, the ML detector is a bit-by-bit hard decisor

$$\hat{b}_i = \text{sign}(r_i), \qquad \forall i \in \{1, \ldots, N\}. \quad (43)$$

So finally we have found a transformation that, after being applied to $\boldsymbol{z}$, drastically reduces the number of dimensions without loosing any useful information that can help in the decoding problem. Now we can analyze the performance of the proposed watermark decoder in terms of the probability of error conditioned to a given original image. For doing so, we have to switch back to a statistical model in which the original image is fixed and deterministic and only the secret key is random.

Assuming that the watermarked image has not suffered any alteration, then $z_k = x_k + b_i\alpha_k s_k$, $\forall k \in \mathcal{S}_i$. After plugging this expression into (41), and treating the terms s_i and the sets $\mathcal{S}_i$ as the only random elements in the system, we can compute first- and second-order moments and draw some conclusions. Let us assume, without loss of generality, that $b_i = 1$. Then

$$r_i = \sum_{k\in\mathcal{S}_i} \frac{|x_k + 2\alpha_k s_k|^{c_k} - |x_k|^{c_k}}{\sigma_k^{c_k}}. \tag{44}$$

Let us define the vector $\boldsymbol{q} = (q_1, \ldots, q_L)^T$, where $q_k = |x_k + 2\alpha_k s_k|^{c_k} - |x_k|^{c_k}$. If the pseudorandom sequence $\boldsymbol{s}$ is modeled as L outcomes of an i.i.d. random process, as we did in Section VIII, then the mean and variance of r_i conditioned to a certain partition $\mathcal{T} = \{\mathcal{S}_j\}_{j=1}^N$ are

$$E[r_i|\mathcal{T}] = \sum_{k\in\mathcal{S}_i} \frac{E[q_k]}{\sigma_k^{c_k}} \tag{45}$$

$$\mathrm{Var}(r_i|\mathcal{T}) = \sum_{k\in\mathcal{S}_i} \frac{\mathrm{Var}(q_k)}{\sigma_k^{2c_k}}. \tag{46}$$

If the sets $\{\mathcal{S}_j\}_{j=1}^N$ are sparsely scattered over the whole image, and the same statistical model as that used in Section VIII is applicable here, then it can be proved after some algebra that

$$E[r_i] = \frac{1}{N}\sum_{k=1}^{L} \frac{E[q_k]}{\sigma_k^{c_k}} \tag{47}$$

$$\mathrm{Var}(r_i) = \frac{1}{N}\sum_{k=1}^{L} \frac{\mathrm{Var}(q_k)}{\sigma_k^{2c_k}} + \frac{N-1}{N^2}\sum_{k=1}^{L} \frac{E^2[q_k]}{\sigma_k^{2c_k}} \tag{48}$$

where we have used the relations $E[r_i] = E_{\mathcal{T}}[E[r_i|\mathcal{T}]]$ and $\mathrm{Var}(r_i) = E_{\mathcal{T}}[\mathrm{Var}(r_i|\mathcal{T})] + \mathrm{Var}_{\mathcal{T}}(E[r_i|\mathcal{T}])$. If we assume that the pseudorandom sequence has a uniform discrete marginal distribution with two levels, $\{-1, 1\}$, as we did in Section VIII, then, by observing the definition of $\boldsymbol{q}$ we can infer that

$$E[q_k] = \tfrac{1}{2}\left[(|x_k| + 2\alpha_k)^{c_k} + ||x_k| - 2\alpha_k|^{c_k}\right] - |x_k|^{c_k} \tag{49}$$

$$\mathrm{Var}(q_k) = \tfrac{1}{4}\left[(|x_k| + 2\alpha_k)^{c_k} - ||x_k| - 2\alpha_k|^{c_k}\right]^2 \tag{50}$$

and we can substitute these expressions in (47) and (48) to obtain the mean and variance of the elements of the vector $\boldsymbol{r}$ conditioned to a given original image $\boldsymbol{x}$. It can also be proved that when $b_i = -1$, the expected value of r_i is negative, with amplitude given by (47). The variance of r_i in this case is exactly the same as that given by (48).

In the definition of the sufficient statistics r_i we can see that they can each be expressed as a sum of statistically independent terms. Therefore, if N is not too large, we can apply the central limit theorem and approximate the distribution of $\boldsymbol{r}$ by a vector Gaussian pdf.

Let us define the SNR

$$\mathrm{SNR} \triangleq \frac{E^2[r_i]}{\mathrm{Var}(r_i)}. \tag{51}$$

Then, under the Gaussian approximation and assuming that message vectors form a binary antipodal constellation with $M = 2^N$ points, so a bit-by-bit hard decisor is used, the probability of bit error is

$$P_b = Q\left(\sqrt{\mathrm{SNR}}\right). \tag{52}$$

C. Watermark Detector

Given a certain key K, the detector that maximizes the probability of detection for any desired probability of false alarm is given by (11), repeated here

$$l(\boldsymbol{z}) = \ln \sum_{\boldsymbol{b}} \frac{p(\boldsymbol{b}) f_x(\boldsymbol{z} - \boldsymbol{P}\boldsymbol{b})}{f_x(\boldsymbol{z})} \underset{H_0}{\overset{H_1}{\gtrless}} \eta. \tag{53}$$

Assuming equiprobable messages and using the expression of the generalized Gaussian pdf given in (36), this is equivalent to

$$l(\boldsymbol{z}) = -\ln M + \sum_{k=1}^{L} \beta_k^{c_k} |z_k|^{c_k} + \ln\left(\sum_{l=1}^{M}\prod_{i=1}^{N} \exp\left\{-\sum_{k\in\mathcal{S}_i} \beta_k^{c_k} |z_k - b_{l,i}\alpha_k s_k|^{c_k}\right\}\right) \tag{54}$$

where $(b_{l,1}, \ldots, b_{l,N})^T$ is the lth message vector. For simplicity, we will concentrate on the case in which a "pure" watermark not carrying any hidden information is employed. Then, under this assumption the log-likelihood function is considerably simpler

$$l(\boldsymbol{z}) = \sum_{k=1}^{L} \beta_k^{c_k} (|z_k|^{c_k} - |z_k - \alpha_k s_k|^{c_k}). \tag{55}$$

Once we have derived the optimum structure based on a statistical characterization of the original image, we can study the probabilities of false alarm and detection conditioned to a given fixed original image, assuming that the secret key is taken at random. The goal is to obtain estimates of the proportion of keys that produce a false positive when the detection test is applied directly to the original image and the proportion of keys that yield a positive result when they are applied both in the watermarking stage and the detection test. Let us first study the distribution of the log-likelihood function when hypothesis H_0 is true. In this case, and assuming that the

image does not suffer any alteration, we have that $z_k = x_k$, $\forall k \in \{1, \ldots, L\}$. Hence $l(\boldsymbol{z})$ has the form

$$l(\boldsymbol{z}) = \sum_{k=1}^{L} \beta_k^{c_k} (|x_k|^{c_k} - |x_k - \alpha_k s_k|^{c_k}) \tag{56}$$

which is a sum of L statistically independent terms ($\boldsymbol{s}$ is i.i.d.) and, applying the central limit theorem, can thus be approximated by a Gaussian random variable. Assuming that all the elements of the pseudorandom sequence have two equiprobable levels, $\{-1, 1\}$, it can be easily shown that the mean and variance of $l(\boldsymbol{z})$ are

$$E[l(\boldsymbol{z})|H_0] = \sum_{k=1}^{L} \beta_k^{c_k} |x_k|^{c_k} - \tfrac{1}{2} \sum_{k=1}^{L} \beta_k^{c_k} \cdot (|x_k + \alpha_k|^{c_k} + |x_k - \alpha_k|^{c_k}) \tag{57}$$

$$\mathrm{Var}(l(\boldsymbol{z})|H_0) = \tfrac{1}{4} \sum_{k=1}^{L} \beta_k^{2c_k} (|x_k + \alpha_k|^{c_k} - |x_k - \alpha_k|^{c_k})^2. \tag{58}$$

When hypothesis H_1 is true, i.e., when $z_k = x_k + \alpha_k s_k$, the log-likelihood function has the form

$$l(\boldsymbol{z}) = \sum_{k=1}^{L} \beta_k^{c_k} (|x_k + \alpha_k s_k|^{c_k} - |x_k|^{c_k}) \tag{59}$$

which is also a sum of statistically independent terms. Hence $l(\boldsymbol{z})$ is also approximately Gaussian under hypothesis H_1. If we compare this expression to (56), considering the fact that $s_k \in \{-1, 1\}$ equiprobably, for any $k \in \{1, \ldots, L\}$, then we can infer that each term in the summation may take the same two values as in (56), with opposite sign. Thus, the distribution of $l(\boldsymbol{z})$ under H_1 is symmetrical to the distribution under H_0 with respect to the origin. Let us define $m_1 \triangleq E[l(\boldsymbol{z})|H_1]$ and $\sigma_1^2 \triangleq \mathrm{Var}(l(\boldsymbol{z})|H_1)$. Then, under the Gaussian approximation, the probabilities of false alarm and detection are

$$P_F = Q\left(\frac{\eta + m_1}{\sigma_1}\right) \tag{60}$$

$$P_D = Q\left(\frac{\eta - m_1}{\sigma_1}\right). \tag{61}$$

If we define the signal-to-noise ratio

$$\mathrm{SNR}_1 \triangleq \frac{m_1^2}{\sigma_1^2} \tag{62}$$

and we call $Q^{-1}(P_F)$ the value $x \in \mathbb{R}$ such that $Q(x) = P_F$, then the ROC is given by the expression

$$P_D = Q\left(Q^{-1}(P_F) - 2\sqrt{\mathrm{SNR}_1}\right) \tag{63}$$

which depends exclusively on the value of SNR_1. Therefore, this SNR can be used to compare the performance of the ML watermark detector for different images.

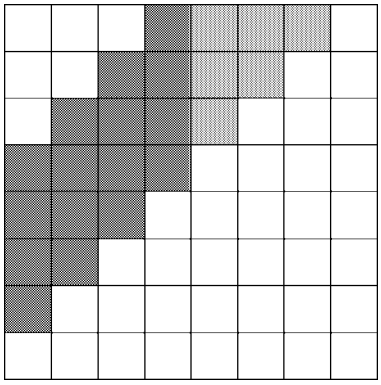


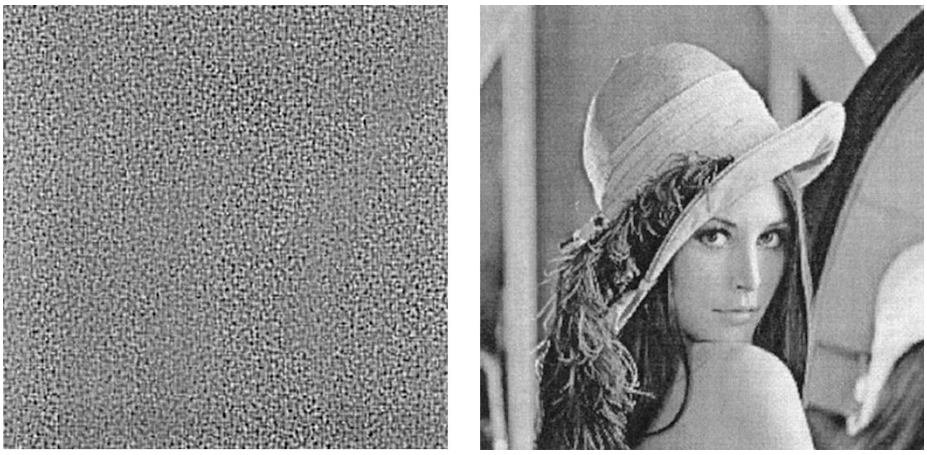
Fig. 16. DCT coefficients where the watermark is embedded.

D. Experimental Results

To contrast the analytical expressions with empirical results, we have performed experiments with two of the images shown in Section VII-D, "Lena" and "Brick." The former is a good representative of images with flat areas and sharp edges, while the latter is an example of images containing noisy textures. In all the experiments the DCT coefficients in middle frequencies shown in Fig. 16 have been altered, following the ideas presented in [61]–[63].

We assume that the perceptual mask determines the maximum amplitude distortion that each coefficient of the original image may suffer while satisfying the invisibility constraint. A good psychovisual model in the DCT domain (with 8×8 blocks) is capital to render the sequence $\boldsymbol{\alpha}$. For the work presented in this section we have followed the model proposed in [64] and [65], similar to those proposed in [66] and [67], that has been also applied to derive adaptive quantization matrices for the JPEG algorithm [68]. This model has been here simplified by disregarding the so-called contrast-masking effect, for which the perceptual mask at a certain coefficient depends on the amplitude of the coefficient itself. This effect has been taken into account by other authors [69], [70]. On the other hand, the background intensity effect, for which the mask depends on the magnitude of the dc coefficient (i.e., the background), has been taken into account. The watermark power obtained from the application of this model has been further reduced by 12 dB to introduce a certain degree of conservativeness in the watermark due to those effects that have been overlooked (e.g., spatial masking in the frequency domain [54]). In Fig. 17 we show an example of a watermark and a watermarked version for "Lena."

In the experiments, we have used the same value of the distribution parameter c for all the DCT coefficients, leaving it as a system parameter. The variance of each original image coefficient is estimated from the watermarked image. In Figs. 18 and 19 we show plots of the BER for the two test images. Both empirical curves and analytical approximations corresponding to four values of c are included in each plot. In all cases, empirical measures have been performed by taking 100 different keys at random. We can see that the values computed using the analytical expressions derived in Section IX-B are good



(a) (b)

Fig. 17. Example of (a) a watermark and (b) a watermarked version for "Lena" (256 × 256).

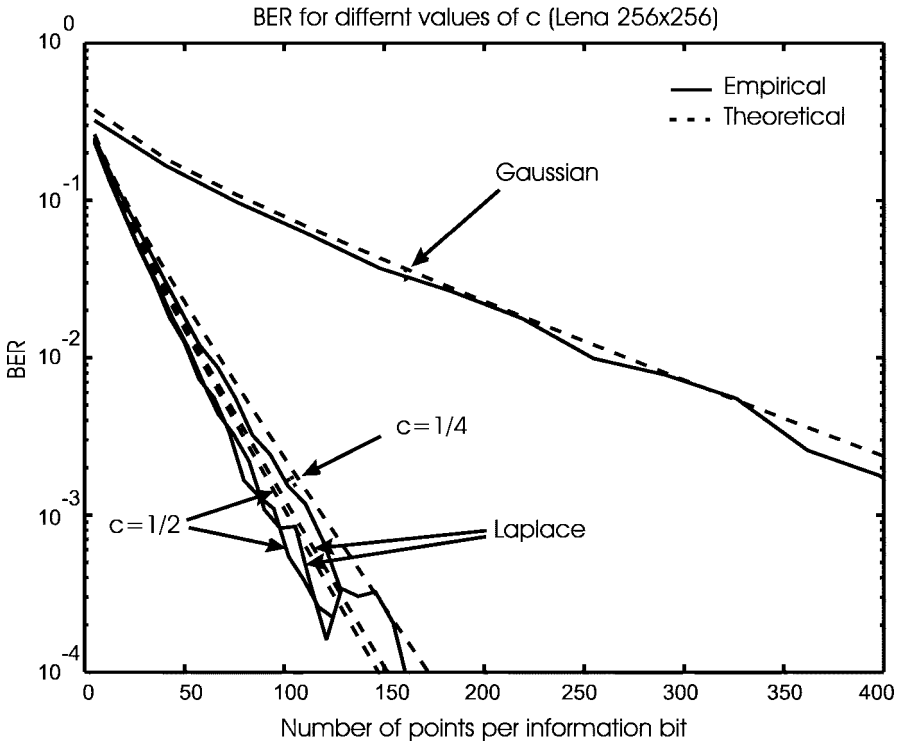


Fig. 18. BER versus pulse size for "Lena" (256 × 256).

approximations of the empirical values. Note that in both cases the best performance is achieved with $c = 1/2$. It is clear from the figures that by choosing an appropriate value of c, performance can be substantially improved. For example, it is clear that in both images the decoder based on the Gaussian model for the DCT coefficients of the original image considerably degrades the BER. This suggests that the correlation receiver used in [69] and [71], which is optimum in the Gaussian case, is not a good candidate for watermark detection purposes.

In Figs. 20 and 21 we show in more detail how the parameter c influences the value of the SNR defined in (51). The curves have been computed using the theoretical expressions derived in Section IX-B and the points of the curve corresponding to $c = 1/2$, $c = 1$ (Laplace), and $c = 2$ (Gaussian) have been circled. Note that the value of c achieving the optimal performance is different in each image. While the maximum SNR lies somewhere between $c = 1/2$ and $c = 1$ for "Lena," it falls down to approximately $c = 1/4$ for "Brick." Again, there is a patent difference between the performance achieved with the correlation receiver (Gaussian case, $c = 2$) and the maximum of the curve.

We have also performed experiments with the two aforementioned images to measure the performance of the watermark detector test derived in Section IX-C. In all cases, empirical measures have been obtained by taking 1000 keys at random. The DCT coefficients in which the watermark

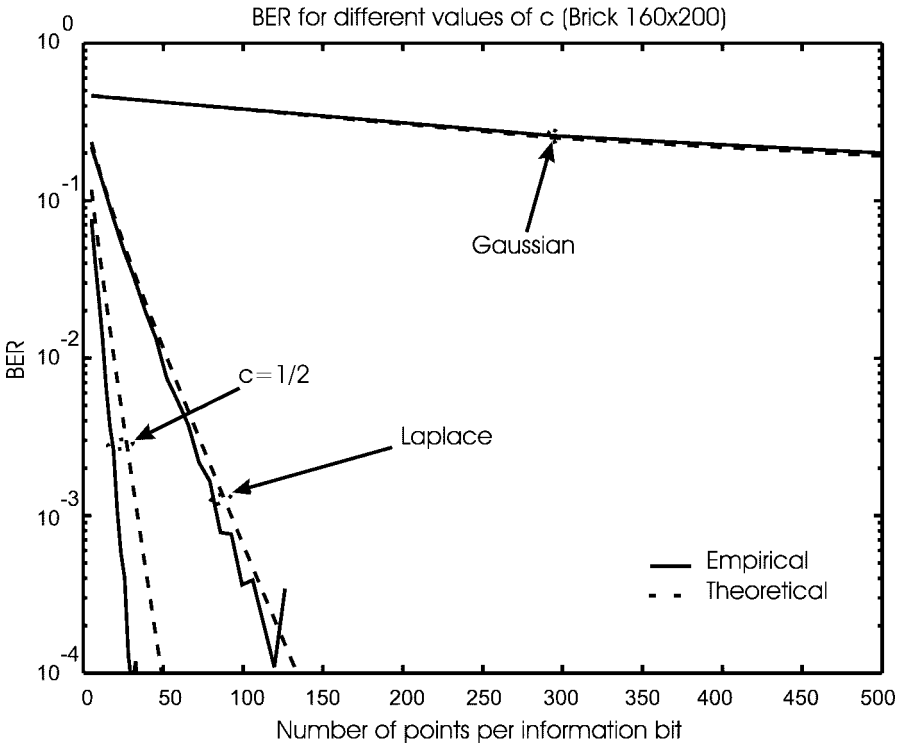


Fig. 19. BER versus pulse size for "Brick" (160 × 200).

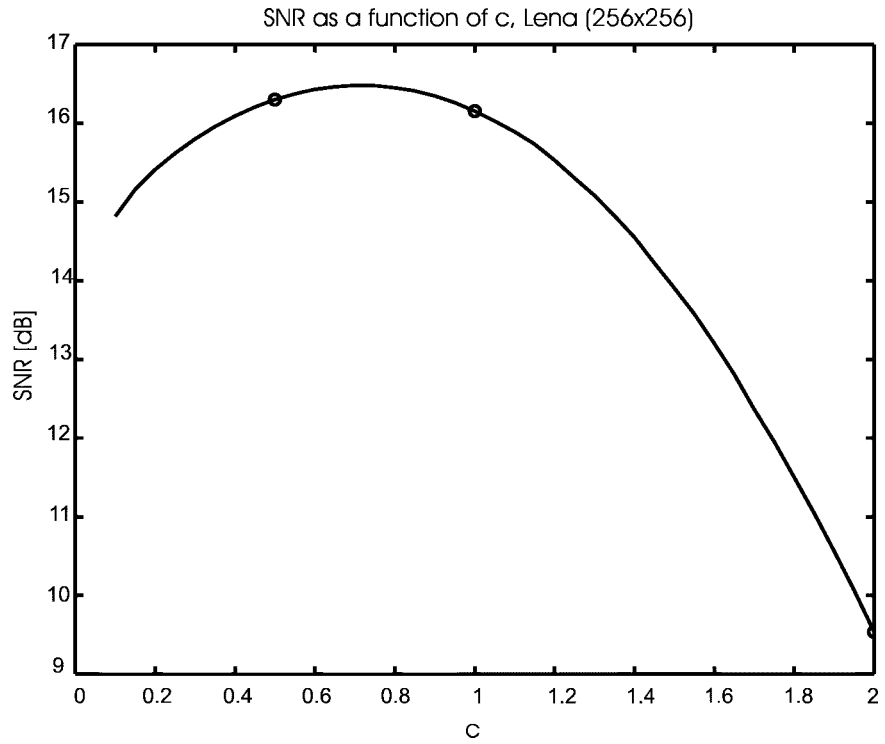


Fig. 20. SNR as a function of c for "Lena" (256 × 256).

has been embedded are also the ones shown in Fig. 16. In Table 1 we have gathered both theoretical values and empirical measures of the signal to noise ratio SNR_1 defined in (62). As we know, this parameter completely determines the shape of the ROC, so it can be used as a performance measure for comparison purposes. We also show in Figs. 22–25 curves of the theoretical and empirical P_F and P_D as a function of the threshold. We can see that the analytical approximations derived in Section IX-C are quite accurate. The different levels of performance achieved

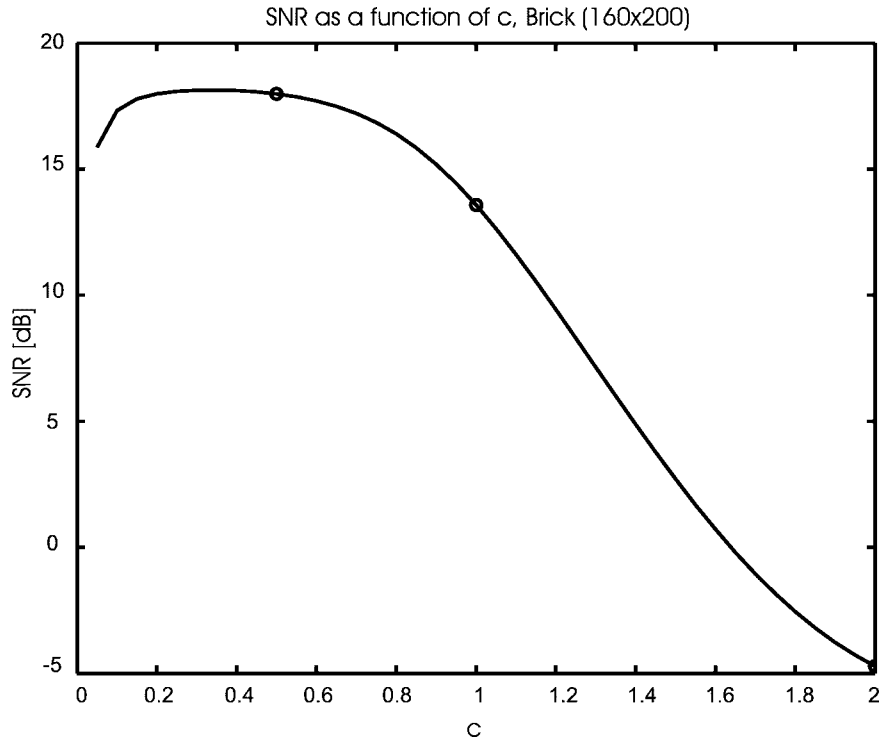


Fig. 21. SNR as a function of c for "Brick" (160 × 200).

Table 1
Empirical and Theoretical Signal-to-Noise Ratio SNR_1 (in dB)

Image	$c = 1/2$		*Laplace*		*Gaussian*	
	Empirical	*Theoretical*	*Empirical*	*Theoretical*	*Empirical*	*Theoretical*
Lena	29.38	29.34	29.07	28.71	21.39	20.78
Brick	42.97	43.23	30.89	30.81	6.00	6.26

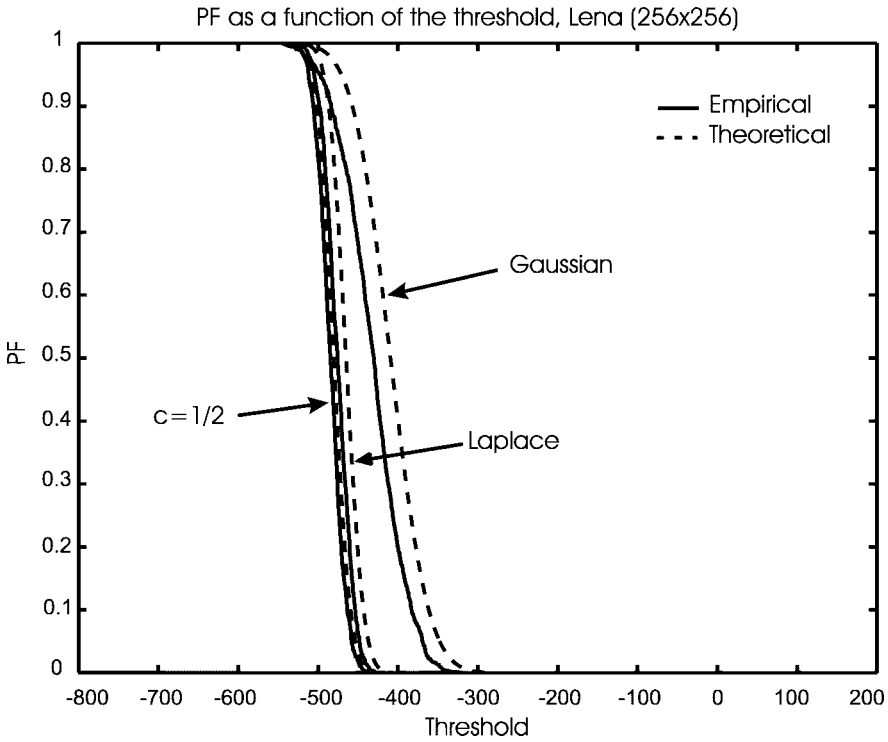


Fig. 22. Probability of false alarm with "Lena" (256 × 256).

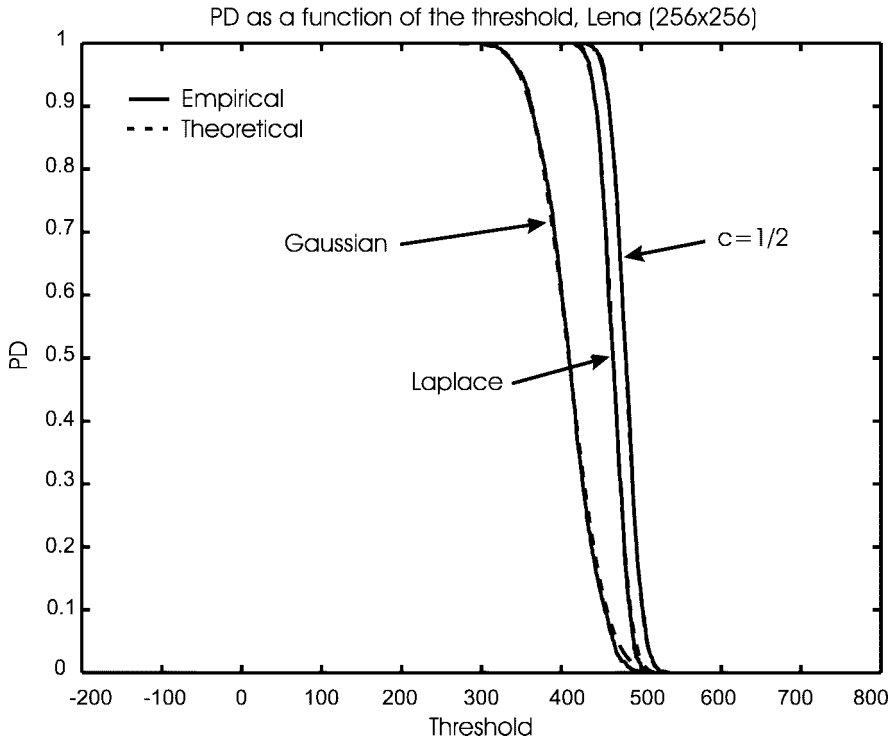


Fig. 23. Probability of detection with "Lena" (256 × 256).

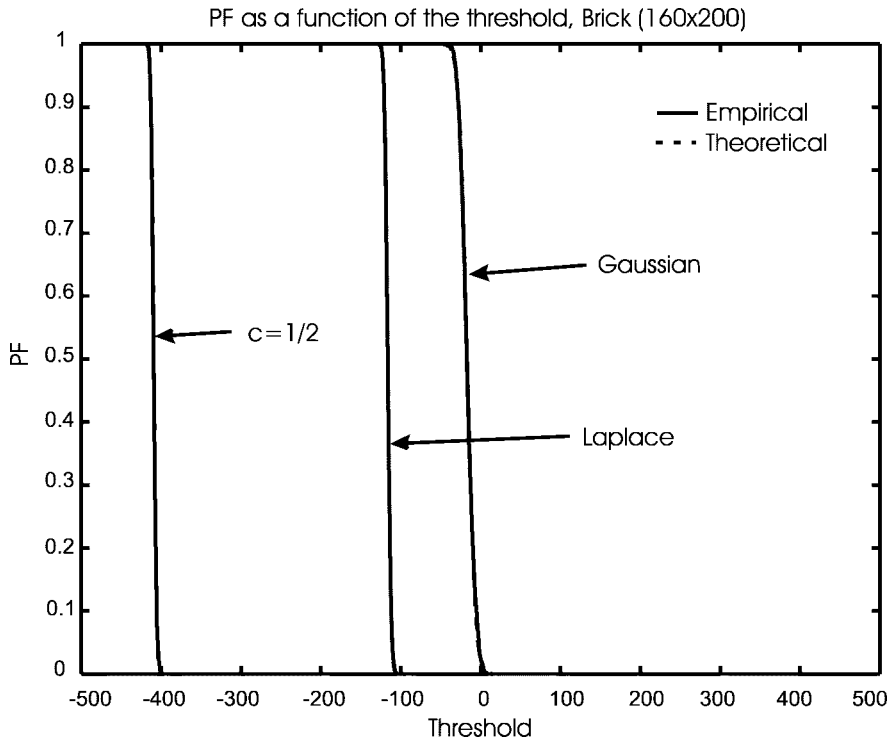


Fig. 24. Probability of false alarm with "Brick" (160 × 200).

with different values of c are also evident. We can see that for both images the Gaussian assumption leads to the worst performance results.

X. Concluding Remarks

In this paper we have discussed the statistical analysis of image watermarking algorithms in which the original image is not needed during the watermark detection and extraction processes. In this context, watermarking can be seen as a communication problem in which a signal carrying some information is transmitted through a noisy channel where the noise is the original image itself, unknown to the receiver. Watermark verification can be seen, hence, as a statistical decision problem involving two tests: first,

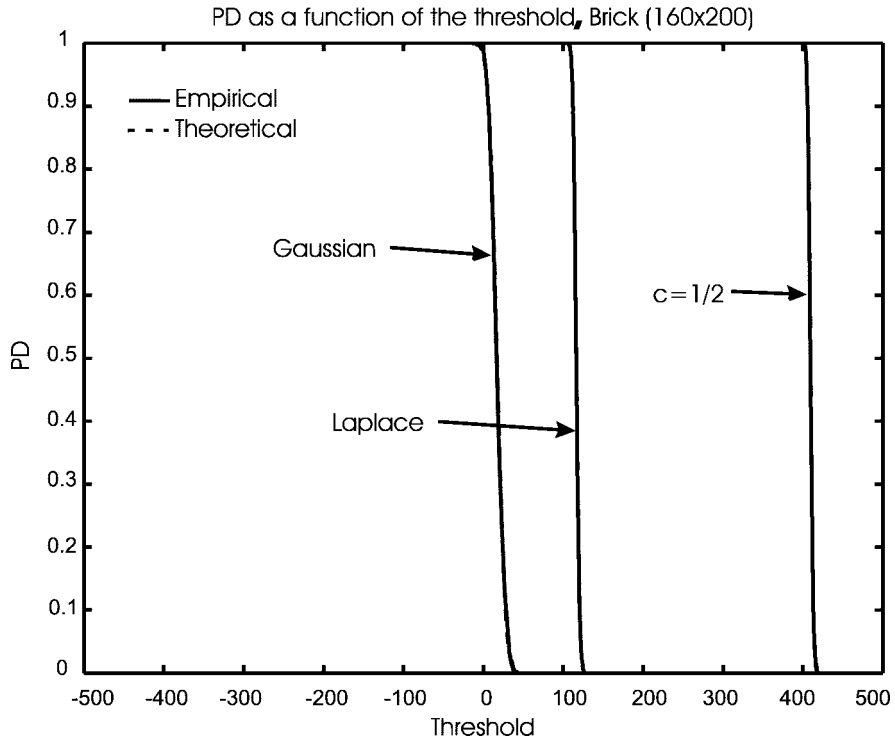


Fig. 25. Probability of detection with "Brick" (160 × 200).

detect the very presence of the watermark, then estimate the information it can optionally carry.

Special attention has been paid to the possible nonavailability of adequate statistical models for the original image (i.e., the channel noise). When such a characterization is impossible, a design approach based on a statistical analysis conditioned to a given original image has been proposed that allows us to derive efficient structures for watermark detection and extraction. If a reasonably accurate model exists, we have shown that it can be exploited to derive detection structures that can considerably improve the performance of the watermarking system. In both cases, a careful theoretical analysis of watermarking techniques using a statistical approach constitutes a rigorous basis for the development of adequate embedding and detection algorithms.

In addition, a theoretical analysis is, in our opinion, vital in order to get a better understanding of the different problems that arise in watermarking and to assess rigorously the suitability of different algorithms, considering the performance requirements of copyright protection applications. With these ideas in mind, we have focused on spread spectrum techniques, and we have analyzed how image characteristics, different kinds of attacks, and system parameters such as the length of the bit string carried by the watermark influence the overall performance of the system.

There are still many open research problems in the field of watermarking for copyright protection. A theoretical approach to the study of watermarking techniques will produce immediate benefits, as we have shown in this paper.

REFERENCES

[1] W. Ciciora, "Inside the set-top box," *IEEE Spectrum,* vol. 32, pp. 70–75, Apr. 1995.

[2] B. Macq and J. Quisquater, "Cryptology for digital TV broadcasting," *Proc. IEEE,* vol. 83, pp. 944–957, Feb. 1995.

[3] A. J. Viterbi, "Four laws of nature and society: The governing principles of digital wireless communications networks," in *Wireless Communications. Signal Processing Perspectives,* H. Poor and G. Wornell, Eds. Englewood Cliffs, NJ: Prentice-Hall, 1998.

[4] R. G. van Schyndel, A. Z. Tirkel, and C. F. Osborne, "A digital watermark," in *Proc. IEEE Int. Conf. Image Processing,* Austin, TX, 1994, pp. 86–89.

[5] R. B. Wolfgang and E. J. Delp, "A watermarking technique for digital imagery: Further studies," in *Proc. Int. Conf. Imaging Science,* Las Vegas, NV, June/July 1997, pp. 279–287.

[6] ——, "A watermark for digital images," in *Proc. 1996 Int. Conf. Image Processing,* Lausanne, Switzerland, Sept. 1996, vol. 3, pp. 219–222.

[7] K. Matsui and K. Tanaka, "Video-steganography: How to embed a signature in a picture," in *Proc. IMA Intellectual Property,* Jan. 1994, vol. 1, pp. 187–206.

[8] I. J. Cox, J. Kilian, F. T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for multimedia," *IEEE Trans. Image Processing,* vol. 6, pp. 1673–1687, Dec. 1997.

[9] P. Noll, "Digital audio coding for visual communications," *Proc. IEEE,* vol. 83, pp. 925–943, June 1995.

[10] J. Nechvatal, "Public key cryptography," in *Contemporary Cryptology,* G. Simmons, Ed. New York: IEEE Press, 1992.

[11] D. Bayer, S. Haber, and W. Stornetta, "Improving the efficiency and reliability of digital time-stamping," in *Sequences II: Methods in Communications, Security and Computer Science.* New York: Springer-Verlag, 1993, pp. 329–334.

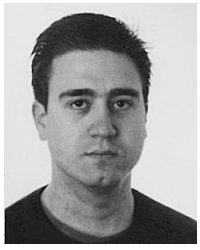
[12] S. Haber and W. Stornetta, "How to time-stamp a digital document," *J. Cryptology,* vol. 3, no. 2, pp. 99–112, 1991.

[13] J. Brassil, S. Low, N. Maxemchuk, and L. O'Gorman, "Electronic marking and identification techniques to discourage document copying," in *Proc. Infocom'94,* pp. 1278–1287.

[14] ——, "Electronic marking and identification techniques to discourage document copying," *IEEE J. Selec. Areas Commun.,* vol. 13, pp. 1495–1504, Oct. 1995.

[15] S. H. Low, N. F. Maxemchuk, J. T. Brassil, and L. O'Gorman, "Document marking and identification using both line and word shifting," in *Proc. Infocom'95*, Apr. 1995.
[16] J. Brassil, S. Low, N. F. Maxemchuk, and L. O'Gorman, "Hiding information in document images," in *Proc. Conf. Information Sciences and Systems (CISS-95)*, Mar. 1995, Johns Hopkins Univ., Baltimore, MD, pp. 482–489.
[17] A. K. Choudhury, N. F. Maxemchuk, S. Paul, and H. G. Schulzrinne, "Copyright protection for electronic publishing over computer networks," *IEEE Network Mag.*, vol. 9, pp. 12–21, May/June 1995.
[18] N. F. Maxemchuk, "Electronic document distribution," *AT&T Tech. J.*, vol. 73, no. 5, pp. 73–80, Sept./Oct. 1994.
[19] S. H. Low and N. F. Maxemchuk, "Performance comparison of two text marking methods," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 561–572, May 1998.
[20] S. H. Low, N. F. Maxemchuk, and A. M. Lapone, "Document identification for copyright protection using centroid detection," *IEEE Trans. Commun.*, vol. 46, pp. 372–383, Mar. 1998.
[21] L. Boney, A. H. Tewfik, and K. N. Hamdy, "Digital watermarks for audio signals," in *EUSIPCO'96, VIII European Signal Proc. Conf.*, Trieste, Italy, Sept. 1996, pp. 1697–1700.
[22] C. Neubauer, J. Herre, and K. Brandenburg, "Continuous steganographic data transmission using uncompressed audio," in *Proc. Information Hiding Workshop*, Portland, OR, Apr. 1998, pp. 208–217.
[23] M. D. Swanson, B. Zhu, B. Chau, and A. H. Tewfik, "Object-based transparent video watermarking," in *Electron. Proc. IEEE SPS Workshop Multimedia Signal Processing*, Princeton, NJ, June 1997.
[24] M. D. Swanson, B. Zhu, and A. H. Tewfik, "Multiresolution scene-based video watermarking using perceptual models," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 540–550, May 1998.
[25] F. Hartung and B. Girod, "Digital watermarking of raw and compressed video," in *Digital Compression Technologies and Systems for Video Communications* (SPIE Proceedings Series), vol. 2952, N. Ohta, Ed. San Diego, CA: SPIE, 1996, pp. 205–213.
[26] ——, "Fast public-key watermarking of compressed video," in *Proc. IEEE ICIP'97*, Santa Barbara, CA, Oct. 1997, vol. I, pp. 528–531.
[27] F. Hartung, P. Eisert, and B. Girod, "Digital watermarking of MPEG-4 facial animation parameters," *Comput. & Graphics*, vol. 22, no. 3, pp. 425–435, Aug. 1998.
[28] F. Hartung and B. Girod, "Copyright protection in video delivery networks by watermarking of pre-compressed video," in *Multimedia Applications, Services and Techniques—ECMAST'97* (Springer Lecture Notes in Computer Science), vol. 1242, S. Fdida and M. Morganti, Eds. Heidelberg, Germany: Springer, 1997, pp. 423–436.
[29] ——, "Digital watermarking of MPEG-2 coded video in the bitstream domain," in *Proc. ICASSP'97*, Munich, Germany, Apr. 1997, pp. 2621–2624.
[30] G. C. Langelaar, R. L. Lagendijk, and J. Biemond, "Real-time labeling methods for MPEG compressed video," in *Proc. 18th Symp. Information Theory in the Benelux*, Veldhoven, The Netherlands, May 1996, pp. 25–32.
[31] T.-L. Wu and S. F. Wu, "Selective encryption and watermarking of MPEG video," submitted for publication.
[32] R. Ohbuchi, H. Masuda, and M. Aono, "Watermarking three-dimensional polygonal models through geometric and topological modifications," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 551–560, May 1998.
[33] G. Bleumer, "Biometric yet privacy protecting person authentication," in *Proceedings Information Hiding: Second International Workshop*, D. Aucsmith, Ed. Berlin, Germany: Springer-Verlag, 1998, pp. 101–112.
[34] T. Sander and C. F. Tschudin, "On software protection via function hiding," in *Information Hiding: Second International Workshop*, D. Aucsmith, Ed. Berlin, Germany: Springer-Verlag, 1998, pp. 113–125.
[35] I. J. Cox, J. Kilian, T. Leighton, and T. Shamoon, "A secure, robust watermark for multimedia," in *Information Hiding*, G. Goos, J. Hartmanis, and J. Leeuwen, Eds. Berlin, Germany: Springer-Verlag, pp. 185–206.
[36] C. I. Podilchuk and W. Zeng, "Perceptual watermarking of still images," in *Electron. Proc. IEEE SPS Workshop Multimedia Signal Processing*, Princeton, NJ, June 1997.
[37] D. Lynch and L. Lundquist, *Digital Money: The New Era of Internet Commerce.* New York: Wiley, 1995.
[38] F. A. Petitcolas, R. J. Anderson, and M. G. Kuhn, "Attacks on copyright marking systems," in *Information Hiding: Second International Workshop*, D. Aucsmith, Ed. Berlin, Germany: Springer-Verlag, 1998, pp. 219–239.
[39] J.-P. M. Linnartz and M. van Dijk, "Analysis of the sensitivity attack against electronic watermarks in images," in *Information Hiding: Second International Workshop*, D. Aucsmith, Ed. Berlin, Germany: Springer-Verlag, 1998,pp. 259–273.
[40] M. Maes, "Twin peaks: The histogram attack on fixed depth image watermarks," in *Information Hiding: Second International Workshop*, D. Aucsmith, Ed. Berlin, Germany: Springer-Verlag, 1998, pp. 291–306.
[41] J. R. Hernández and F. Pérez-González, "Shedding more light on image watermarks," in *Information Hiding: Second International Workshop*, D. Aucsmith, Ed. Berlin, Germany: Springer-Verlag, 1998, pp. 192–208.
[42] N. Nikolaidis and I. Pitas, "Robust image watermarking in the spatial domain," *Signal Processing*, vol. 66, pp. 385–404, May 1998.
[43] S. Glisic and B. Vucetic, *Spread Spectrum CDMA for Wireless Communications.* Norwood, MA: Artech House, 1997.
[44] A. Viterbi, *CDMA. Principles of Spread Spectrum Communication.* Reading, MA: Addison-Wesley, 1995.
[45] J. R. Hernández, F. Pérez-González, J. M. Rodríguez, and G. Nieto, "Performance analysis of a 2D-multipulse amplitude modulation scheme for data hiding and watermarking of still images," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 510–524, May 1998.
[46] B. Sklar, *Digital Communications. Fundamentals and Applications.* Englewood Cliffs, NJ: Prentice-Hall, 1988.
[47] H. L. V. Trees, *Detection, Estimation and Modulation Theory*. New York: Wiley, 1968, pt. I.
[48] S. A. Kassam and H. V. Poor, "Robust techniques for signal processing: A survey," *Proc. IEEE*, vol. 73, pp. 433–481, Mar. 1985.
[49] P. J. Huber, *Robust Statistics.* New York: Wiley, 1981.
[50] J. J. K. O. Ruanaidh and T. Pun, "Rotation, scale and translation invariant spread spectrum digital image watermarking," *Signal Processing*, vol. 66, pp. 303–318, May 1998.
[51] S. Wilson, *Digital Modulation and Coding.* Englewood Cliffs, NJ: Prentice-Hall, 1996.
[52] S. Lin and D. Costello, *Error Control Coding: Fundamentals and Applications.* Englewood Cliffs, NJ: Prentice-Hall, 1983.
[53] J. R. Hernández, J. M. Rodríguez, and F. Pérez-González, "Improving the performance of spatial watermarking of images using channel coding," *Signal Processing*, to be published.
[54] A. N. Netravali and B. G. Haskell, *Digital Pictures. Representation, Compression and Standards.* New York: Plenum, 1995.
[55] J. R. Hernández, F. Pérez-González, and J. M. Rodríguez, "The impact of channel coding on the performance of spatial watermarking for copyright protection," in *Proc. ICASSP'98*, Seattle, WA, May 1998, vol. 5, pp. 2973–2976.
[56] ——, "Coding and synchronization: A boost and a bottleneck for the development of image watermarking," in *Proc. COST 254 Workshop Intelligent Communications*, SSGRR, L'Aquila, Italia, June 1998, pp. 77–82.
[57] J. S. Lim, *Two-Dimensional Signal and Image Processing.* Englewood Cliffs, NJ: Prentice-Hall, 1990.
[58] G. L. Anderson and A. N. Netravali, "Image restoration based on a subjective criterion," *IEEE Trans. Syst., Man., Cybern.*, vol. SMC-6, pp. 845–853, Dec. 1976.
[59] R. Clarke, "Relation between the Karhunen–Loéve and cosine transforms," *Proc. Inst. Elect. Eng.*, vol. 128, pt. F, pp. 359–360, 1981.
[60] R. J. Clarke, *Transform Coding of Images.* New York: Academic, 1985.
[61] J. Zhao and E. Koch, "Embedding robust labels into images for copyright protection," in *Proc. Int. Congr. Intellectual Property Rights for Specialized Information, Knowledge and New Technologies*, Vienna, Austria, Aug. 21–25, 1995, pp. 242–251.

[62] E. Koch, J. Rindfrey, and J. Zhao, "Copyright protection for multimedia data," in *Digital Media and Electronc Publishing*. New York: Academic, 1996, pp. 203–213.
[63] J. J. K. O. Ruanaidh, W. J. Dowling, and F. M. Boland, "Watermarking digital images for copyright protection," in *IEE Proc. Vision, Image and Signal Processing*, Aug. 1996, vol. 143, no. 4, pp. 250–256.
[64] A. J. Ahumada and H. A. Peterson, "Luminance-model-based DCT quantization for color image compression," in *Human Vision, Visual Processing, and Digital Display III (Proc. SPIE)*, 1992.
[65] J. A. Solomon, A. B. Watson, and A. J. Ahumada, "Visibility of DCT basis functions: Effects of contrast masking," in *Proc. Data Compression Conf.*,Snowbird, UT, 1994, pp. 361–370.
[66] C. I. Podilchuk and W. Zeng, "Image-adaptive watermarking using visual models," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 525–539, May 1998.
[67] J. Delaigle, C. D. Vleeschouwer, and B. Macq, "Watermarking algorithm based on a human visual system," *Signal Processing*, vol. 66, pp. 319–336, May 1998.
[68] A. B. Watson, "Visual optimization of DCT quantization matrices for individual images," in *Proc., AIAA Computing in Aerospace 9*, San Diego, CA, 1993, pp. 286–291.
[69] M. Barni, F. Bartolini, V. Capellini, and A. Piva, "A DCT-domain system for robust image watermarking," *Signal Processing*, vol. 66, pp. 357–372, May 1998.
[70] A. Piva, M. Barni, F. Bartolini, and V. Cappellini, "DCT-based watermark recovering without resorting to the uncorrupted original image," in *Proc. IEEE ICIP'97*, Santa Barbara, CA, vol. I, Oct. 1997, pp. 520–523.
[71] J.-P. Linnartz, T. Kalker, and G. Depovere, "Modeling the false alarm and missed detection rate for electronic watermarks," in *Information Hiding: Second International Workshop*, D. Aucsmith, Ed. Berlin, Germany: Springer-Verlag, 1998, pp. 330–344.



Juan R. Hernández (Student Member, IEEE) received the Ingeniero de Telecomunicación degree from the University of Vigo, Spain, in 1993, the M.S. degree in electrical engineering from Stanford University, Stanford, CA, in 1996, and the Ph.D. degree in telecommunications engineering from the University of Vigo, Spain, in 1998.

From 1993 to 1995, he was a member of the Department of Communication Technologies, University of Vigo, where he worked on hardware for digital signal processing and access control systems for digital television. Since 1996, he has been a member of this same department, where he is working as a Research Assistant. His research interests include digital communications and copyright protection in multimedia.

Fernando Pérez-González (Member, IEEE) received the Ingeniero de Telecomunicación degree from the University of Santiago, Spain, in 1990 and the Ph.D. degree from the University of Vigo, Spain, in 1993, both in telecommunications engineering.

He joined the faculty of the School of Telecommunications Engineering, University of Vigo, as an Assistant Professor in 1990 and is currently an Associate Professor in the same institution. He has visited the University of New Mexico, Albuquerque, NM, for different periods spanning ten months. His research interests lie in the areas of digital communications, adaptive algorithms, robust control, and copyright protection. He has been the Project Manager of different projects concerned with digital television, both for satellite and terrestrial broadcasting. He is coeditor of the book *Intelligent Methods in Signal Processing and Communications* (Birkhauser, 1997) and has been Guest Editor of a special section of *Signal Processing* magazine devoted to signal processing for communications.