Performance Analysis of a 2-D-Multipulse Amplitude Modulation Scheme for Data Hiding and Watermarking of Still Images

Juan R. Hernández, *Student Member, IEEE*, Fernando Pérez-González, *Member, IEEE*, José Manuel Rodríguez, and Gustavo Nieto

***Abstract*—In this paper a watermarking scheme for copyright protection of still images is modeled and analyzed. In this scheme a signal following a key-dependent two-dimensional multipulse modulation is added to the image for ownership enforcement purposes. The main contribution of this paper is the introduction of an analytical point of view to the estimation of performance measurements. Two topics are covered in the analysis: the ownership verification process, also called watermark detection test, and the data-hiding process. In the first case, bounds and approximations to the receiver operating characteristic are derived. These results can be used to determine the threshold associated to a required probability of false alarm and the corresponding probability of detection. The data-hiding process is modeled as a communications system and approximations for the bit error rate are derived. Finally, analytical expressions are contrasted with experimental results.**

***Index Terms*—Copyright protection, codes, cryptography, decision-making, image processing, image communication, information theory.**

I. Introduction

THE enormous progress that digital technologies have experienced during the last decades has contributed to the popularization of the use of electronic media for transmission and storage of documents, images, audio, video, and other types of information. Information stored in digital format can be copied without quality loss and distributed efficiently at fairly low costs. These developments have also increased the potential for interception, manipulation, misuse, and unauthorized distribution of information. This is, in fact, one of the main impediments to commercial use of communication networks and electronic storage media for distribution of digital information. For this reason, the design of techniques for preserving the ownership of digital information is the cornerstone of the development of future multimedia services.

Previous research on copyright protection of still images has resulted in the appearance of several watermarking methods. In all these techniques, the contents of the original image is altered in a fashion determined by a secret key and, optionally, by a certain amount of information to be hidden into the image.

Manuscript received March 10, 1997; revised July 15, 1997.

The authors are with the Departmento de Tecnologías de las Comunicaciones, E.T.S.I. Telecomunicaciòn, University of Vigo, Vigo 36200, Pontevedra, Spain (e-mail: jhernan@tsc.uvigo.es, fperez@tsc.uvigo.es).

Publisher Item Identifier S 0733-8716(98)01107-X.

In [1], a watermarking procedure based on spread spectrum (SS) techniques is proposed for application to multimedia data. The watermark consists of a sequence of independent and identically distributed (i.i.d.) Gaussian random variables that are added to the perceptually most significant *DCT* coefficients. Placing the watermark in the perceptually relevant components of the original image provides a high level of robustness against many signal processing techniques aimed at eliminating noise from the image. However, the main limitation of this technique is the need for the original image in the ownership verification process.

A JPEG-based method for embedding labels into images is described in [2]–[4]. In this method, the original image is divided into 8×8 blocks. A triple is chosen among the DCT coefficients at the middle frequencies in each block, and its components are modified to encode one bit. This technique resembles a frequency-hopping SS scheme, but no perceptual constraint is imposed to the modifications introduced in the image. It is also sensitive to attacks such as cropping and affine transforms, that alter the spatial location of the 8×8 blocks with respect to the borders of the image, as well as additive noise concentrated in the middle frequencies.

In [5], a watermarking method is presented. It is based on the addition in the frequency (DCT) domain of an SS signal shaped by a perceptual mask that guarantees that the hidden signal is invisible. The watermarking process is performed blockwise, and the original image is required in the verification test. A data-hiding scheme based on a similar approach is described in [6]. In this case, an alternative spatial-domain watermarking technique is also proposed. The original image is segmented into blocks that are modified by single bits of the hidden message. For this reason, this data-hiding scheme is not robust against cropping. The original image is not required for information decoding. Similar techniques are applied for authentication and distortion measurement of images [7] and for watermarking of audio signals [8]. The scheme we analyze in this article is, in fact, similar to the spatial-domain data-hiding method described in [6].

In [9], a data-hiding scheme, called Patchwork, is proposed. In this method, one bit is encoded by randomly choosing a certain number of pairs of pixels and modifying the difference in luminance level of each pair. This method is sensitive to image cropping and affine transforms, because the spatial reference is fundamental for the correct operation of the

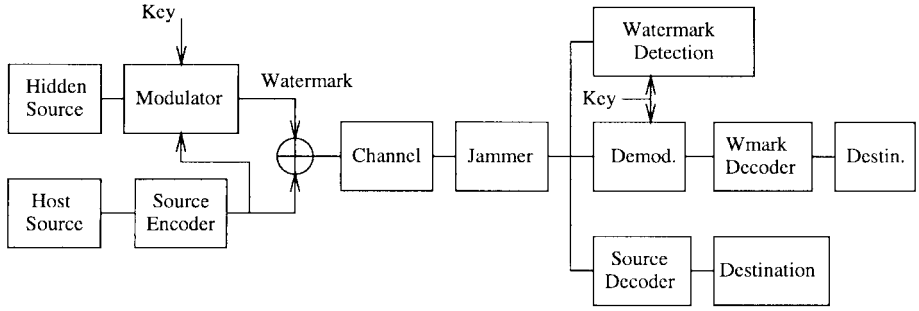


Fig. 1. General model of a data-hiding system.

decoding algorithm. It is also weak against random additive noise attacks.

Another method for data hiding is proposed in [10]. In this case, the image is divided into blocks that are transformed to a different domain (DCT, FFT, Daubechies wavelets). The coefficients in this new domain with the highest energy are altered to encode several bits. Each bit in fact modulates a single coefficient. This method is not resilient to cropping and affine transforms, because the performance of the detection algorithm relies on the correct segmentation of the image into blocks.

In [11], two data-hiding techniques are described. The first one is used with uncompressed video, and is based on direct-sequence (DS) SS techniques. The main weakness of this method, as it happens with all the techniques based on the use of spatial pseudorandom sequences, is the spatial synchronization. Attacks such as cropping, line removal, and affine transforms may shift and rotate the pseudorandom sequence; as a result, resynchronization is necessary before correlating.

In [12], the authors offer an overview of two data-hiding schemes based on classical SS techniques. One of them uses DS spectrum spreading and the other uses frequency hopping. An attempt is made to apply the concept of channel capacity to data hiding. These methods have the same weaknesses as other SS related schemes.

Even though different proposals for solving the copyright enforcement problem have been described and tested with diverse results, previous research in watermarking techniques has suffered from the absence of a theoretical approach to the study of limits in performance. In this paper a watermarking scheme is modeled and analyzed. Both watermarking and data hiding are unified into a single model and special emphasis is placed on the application of concepts from digital transmission theory and information theory. The technique we analyze is similar to schemes described by other authors [6], [11], [12].

Section II offers an overview of the main issues that appear in a watermarking system. In Section III-A, we describe the two-dimensional modulation scheme used to generate the watermark. Equivalent vector channels under different assumptions are derived in Sections III-B–III-D. In Section IV, a coding scheme and a detector structure are studied and analytical approximations to the bit error rate are obtained. In Section V, a watermark detection test structure is proposed and analytical approximations to the probability of false alarm and the probability of detection are derived. Finally, in Section VI, results from simulations are compared to analytical expressions.

II. General Model

In a watermarking scheme, an invisible signal carrying copyright information is added to the image to be protected. In this context, the watermarking approach may be considered as a steganographic technique [13]. In Fig. 1 a general model of a watermarking system is represented.

The signal at the output of the host source encoder corresponds to the image that must be watermarked. The hidden information source generates a message that identifies both the issuer and the recipient of the host data, and optionally, additional information. This message is then mapped onto a modulated waveform that is added to the image. One of the goals of the watermarking scheme is to make it difficult to guess the exact mapping between information and modulated waveforms. For this purpose the modulation process will have a secret key K as one of its parameters.

The channel models the transformations suffered by the image during distribution and authorized usage by the intended recipient. The delivered image may also be intercepted and manipulated by an unauthorized agent (or even by the intended recipient) to delete or corrupt the watermark and illegally redistribute the image. The attacks that the watermarked image may suffer can be categorized as follows.

1) Attacks aimed at deleting the watermark by extracting an estimate of the hidden signal from the watermarked image.
2) Attacks with the purpose of altering the extra information encoded in the hidden waveform. An example of this kind of attack is the use of additive random noise in order to increase the probability of error in the hidden data decoding test.

Two tests are available for ownership-verification purposes. The first one, which we will call the watermark detection test, is used to decide whether an image contains a watermark generated with a certain key. The second one, applied only if the watermark detection test has been passed, decodes the message carried by the hidden waveform.

For a given original image, the watermarked images obtained for different keys and messages can be considered points in a multidimensional space. This set of points must satisfy

certain conditions to be useful. First of all, the watermark must be imperceptible. This means that the set of possible watermarked images must lie inside a hypersphere defined by certain perceptually significant distortion metric and whose center is the original image.

In order to guarantee that the watermarking process is secure, the hidden signal must be inseparable from the original image. In other words, it must be difficult to estimate the original image from the watermarked image when the secret key is not known. Otherwise, an attacker could obtain a good estimate of the watermark and subtract it from the watermarked image. In addition, the set of signal points corresponding to valid watermarked images for a given original image must be sparse to achieve a low probability of generating a valid watermark when the secret key is not known. The watermark must also be robust against manipulations aimed at forcing a change in the result of the watermark detection test or the watermark decoding process. These are perhaps the most challenging requirements, since an unauthorized person who intercepts a watermarked image may store it and then apply any kind of processing technique in order to delete or corrupt the watermark.

SS techniques have been proposed to achieve security and robustness against manipulations. However, the main difference with respect to an SS communication system is that in a watermarking scheme, the jammer is not limited to additive noise attacks because he is the channel himself. Nevertheless, the hacker is limited to those manipulations that do not severely distort the image contents. The modulation scheme, therefore, must exploit this fact. Returning to the geometrical approach, the set of hidden message points in the signal space must lie in the subspace where the original image is defined. Forcing the hidden signal to be coupled to the original image makes it difficult to eliminate the watermark without altering the image. In addition, it guarantees that the watermark is resilient to compression.

Because a data-hiding system can be analyzed as a communication system, the concept of channel capacity can be applied. However, the computation of the channel capacity should take into account first the input constraints derived from the requirements the watermark must satisfy, then a statistical characterization of the host image, which acts as noise since it is assumed to be unknown. It should also consider the worst of the attacks that do not severely distort the image contents.

III. 2-D Multipulse Amplitude Modulation

A. Definitions

In the two-dimensional (2-D) multipulse amplitude modulation scheme, the signal carrying information is expressed as a linear combination of a set of L orthogonal functions $\{p_i[m,n]\}, i \in \{0,\cdots,L-1\}$

$$w[m,n] = \sum_{i=0}^{L-1} b_i p_i[m,n] \tag{1}$$

where $\{p_i[m,n]\}$ satisfy $\langle p_i, p_j\rangle = ||p_i||^2\delta_{ij}$. The signal $w[m,n]$ is added to the original image $x[m,n]$ to obtain the watermarked version $y[m,n] = x[m,n] + w[m,n]$. The coefficients $b_0,\cdots,b_{L-1}$ are used to encode a hidden message. Let $\mathcal{S} = \{\mathcal{S}_0,\cdots,\mathcal{S}_{L-1}\}$ be the collection of sets of points where the pulses $\{p_i\}$ take nonzero values

$$\mathcal{S}_i \triangleq \{(m,n)|p_i[m,n] \neq 0\}, \qquad i \in \{0,\cdots,L-1\}. \tag{2}$$

These sets define the spatial shape of the modulation pulses. The 8×8 blocks used in some watermarking methods [6], [7] can be considered as special cases of this model. For the sake of simplicity, we will assume in the sequel that pulses do not overlap, i.e. $\mathcal{S}_i \cap \mathcal{S}_j = \emptyset\ \forall i \neq j$. This assumption guarantees that pulses are orthogonal.

To meet the security requirements, a different set of modulation pulses is generated for each value of the secret key K. The modulation pulses $p_i[m,n]$ are defined as follows:

$$p_i[m,n] \triangleq \begin{cases} \alpha[m,n]s[m,n], & \text{if } (m,n) \in \mathcal{S}_i \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where $s[m,n]$ is a key-dependent pseudorandom sequence that can be modeled as a zero-mean i.i.d. random sequence with marginal pdf $f_s(s)$ whose variance is constrained to be one. The sequence $\alpha[m,n]$ indicates the maximum allowable standard deviation at each pixel for the pulses to be invisible. We can infer that the coefficients b_i must satisfy $|b_i| \leq 1$ for the watermark to be invisible. The sequence $s[m,n]$ is i.i.d. to provide maximum uncertainty (entropy) for a given marginal distribution $f_s(s)$ when the secret key is unknown. This modulation technique is similar to a DS SS scheme. However, as we noted in the introduction, the main difference with respect to classical SS systems used in communications is that in our context, the jammer is not limited to additive noise attacks. He can in fact play the role of a worst-case channel especially designed to attack the hidden signal without perceptually degrading the image. In Sections III-B–III-D we discuss the impact of the choice of a marginal distribution $f_s(s)$ on the watermarking scheme.

B. Equivalent Vector Channel for Nonaltered Images

As we stated in the introduction, we assume that the original image $x[m,n]$ is not available in the watermark detection and decoding processes. Therefore, it must be taken as noise that introduces uncertainty in the detection problem. When $x[m,n]$ can be modeled as Gaussian noise, the correlation coefficients $r_i \triangleq \langle y, p_i\rangle$ are sufficient statistics for signal detection. However, the Gaussian model is not suitable for real world images. Because there is in fact a lack of good statistical models for common images [14], we will reduce the observation space to the projection onto the pulses $\{p_i\}$ and assume that the information in the subspace orthogonal to these pulses can be ignored.

We will now obtain a statistical characterization of the coefficients r_i for a given image $x[m,n]$. Our approach is to obtain detector structures based on these coefficients and aimed at optimizing the probability of error averaged over the set of keys K for a given image $x[m,n]$ (see Sections IV and V). We start by considering the case in which the watermarked image does not suffer any alteration, and fixed sets $\{\mathcal{S}_0,\cdots,\mathcal{S}_{L-1}\}$

are used for all the keys. We assume that if a perceptual analysis of the watermarked image $y[m,n]$ is performed, a good estimate of the perceptual mask $\alpha[m,n]$ used to generate the watermark is obtained, because the watermarked image and the original are perceptually equivalent. This assumption is in fact supported by simulation results. We will also assume that the spatial location of the pixels that compose the pulses $p_i[m,n]$ is precisely known when the secret key is available. This is not the case when the watermarked image has been cropped or has suffered a geometrical transformation. The issue of spatial synchronization is addressed in Section V. If the watermarked image $y[m,n]$ has not been manipulated, then

$$r_i = \langle y, p_i \rangle = b_i ||p_i||^2 + \langle x, p_i \rangle. \tag{4}$$

If we define $a_i = ||p_i||^2$ and $n_i = \langle x, p_i \rangle$, (4) can be written as

$$r_i = a_i b_i + n_i. \tag{5}$$

Both a_i and n_i are random variables that reflect the random nature of the modulation pulses. Since the key K is the only random variable in the model and $s[m,n]$ is i.i.d., each coefficient r_i is the sum of independent random variables. Therefore, if the pulse size is large enough, by the central limit theorem, r_i is approximately Gaussian. Let us define $\overline{a}_i = E[a_i]$ and $\tilde{a}_i = a_i - \overline{a}_i$. Now

$$r_i = \overline{a}_i b_i + n_{T_i} \tag{6}$$

where $n_{T_i} = \tilde{a}_i b_i + n_i$. The second-order moments of the zero-mean random variables $\tilde{a}_i$ and n_i are

$$\operatorname{var}(\tilde{a}_i) = \sum_{(m,n)\in\mathcal{S}_i} \alpha^4[m,n]\left(E[s^4] - 1\right) \tag{7}$$

$$\operatorname{var}(n_i) = \sum_{(m,n)\in\mathcal{S}_i} \alpha^2[m,n] x^2[m,n] \tag{8}$$

$$E[n_i n_j] = 0, \quad E[\tilde{a}_i \tilde{a}_j] = 0 \quad \forall i \neq j \tag{9}$$

$$E[\tilde{a}_i n_j] = \delta_{ij} \sum_{(m,n)\in\mathcal{S}_i} \alpha^3[m,n] E[s^3] x[m,n] \tag{10}$$

where δ_{ij} is the Kronecker delta function. We will assume that $f_s(s)$ is chosen to be symmetrical about the origin. Under this assumption, $E[s^3] = 0$ and, as a result, $\tilde{a}_i$ and n_j are uncorrelated for all i and j. Therefore, the variance of the aggregate noise is $\operatorname{var}(n_{T_i}) = b_i^2 \operatorname{var}(\tilde{a}_i) + \operatorname{var}(n_i)$. Let us define the vectors $\boldsymbol{b} \triangleq (b_0, \cdots, b_{L-1})^T$, $\boldsymbol{r} \triangleq (r_0, \cdots, r_{L-1})^T$, $\boldsymbol{n}_T \triangleq (n_{T_0}, \cdots, n_{T_{L-1}})^T$ and the matrices $\overline{\boldsymbol{A}} \triangleq [\overline{a}_{ij}]$ and $\boldsymbol{\Gamma} = [\gamma_{ij}] \triangleq E[\boldsymbol{n}_T \boldsymbol{n}_T^T]$. Then

$$\boldsymbol{r} = \overline{\boldsymbol{A}}\boldsymbol{b} + \boldsymbol{n}_T \tag{11}$$

$$\overline{a}_{ij} = \delta_{ij} \sum_{(m,n)\in\mathcal{S}_i} \alpha^2[m,n] \tag{12}$$

$$\gamma_{ij} = \delta_{ij} \sum_{(m,n)\in\mathcal{S}_i} \alpha^2[m,n] x^2[m,n] + \delta_{ij} b_i^2 \sum_{(m,n)\in\mathcal{S}_i} \alpha^4[m,n]\left(E[s^4] - 1\right). \tag{13}$$

These equations describe the equivalent Gaussian vector channel represented graphically in Fig. 2. Note that $E[s^4] \geq 1$ for any pdf $f_s(s)$ since $E[s^2]$ is fixed to one, and that equality holds only for the two-level discrete distribution in which $s[m,n]$ takes ± 1 values. Therefore, when this distribution is used, $\operatorname{var}(\tilde{a}_{ii}) = 0$ and the aggregate noise power is minimum for a given fixed image and a given set $\mathcal{S}$. However, this two-level discrete distribution has less uncertainty than other discrete and continuous pdf's. There is a tradeoff between noise variance (and hence, probability of error) in the equivalent vector channel and uncertainty about the pulse amplitude when the secret key is not known. This means that if we want the hidden signal to be more difficult to intercept, a penalty in performance must be paid. For a given value $E[s^4] \neq 1$, the maximum entropy distribution $f_s(s)$ has the form [15], [16]

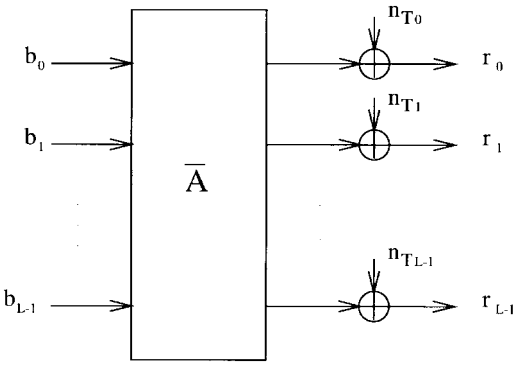


Fig. 2. Equivalent channel for a 2-D multipulse modulation scheme.

$$f_s(s) = A e^{-\lambda_1 s^4 - \lambda_2 s^2}. \tag{14}$$

The distribution which maximizes the entropy subject only to the second order moment constraint $E[s^2] = 1$ is the Gaussian distribution $f_s(s) = (2\pi)^{-1/2} e^{-s^2/2}$, for which $E[s^4] = 3$. Therefore, this distribution achieves the maximum possible entropy over all the values of $E[s^4]$. For other values, the maximum entropy distribution is not Gaussian but follows expression (14).

Matrices $\overline{\boldsymbol{A}}, \boldsymbol{\Gamma}$ are strongly dependent on the spatial pulse shape. Block pulses, for example, induce highly variable values of the matrix coefficients because the statistics of the image in different blocks can significantly differ. If we spread the modulation pulses over the whole image, the matrix coefficients associated with each vector element r_i will gather contributions from all the regions of the image, and, therefore, more homogeneous matrices will result.

C. Equivalent Vector Channel Under Linear Filtering

A similar analysis can be made to include in the vector channel model the effects of possible transformations suffered by the watermarked image. One such transformation, interesting to study because of its power as a signal processing tool, is the finite-impulse response (FIR) space-variant linear filtering. Let $h_{k,l}[m,n]$ be the coefficients of a space-variant linear filter that is applied to the watermarked image. Let $z[m,n]$ be the resulting filtered image

$$z[m,n] = \sum_{k,l} h_{k,l}[m,n] y[m-k, n-l]. \tag{15}$$

Let us define $x^{k,l}[m,n] = x[m-k, n-l]$ and let us define in the same way $y^{k,l}[m,n]$ and $p_i^{k,l}[m,n]$. The correlation coefficient r_i can now be written as

$$r_i = \langle z, p_i \rangle = \sum_{j=0}^{L-1} b_j \sum_{k,l} \left\langle h_{k,l} p_j^{k,l}, p_i \right\rangle + \sum_{k,l} \langle h_{k,l} x^{k,l}, p_i \rangle. \tag{16}$$

Now we can see that for a given key, i.e., a given outcome of $s[m,n]$, intersymbol interference (ISI) is present because $\langle h_{k,l} p_j^{k,l}, p_i \rangle \neq 0$ in general when $(k,l) \neq (0,0)$. The Gaussian approximation is still applicable if the pulse size is large enough since r_i can be seen as the sum of independent random variables. The equivalent ISI channel in matrix form is

$$\boldsymbol{r} = \boldsymbol{A}\boldsymbol{b} + \boldsymbol{n} \tag{17}$$

$$a_{ij} = \sum_{k,l} \left\langle h_{k,l} p_j^{k,l}, p_i \right\rangle \tag{18}$$

$$n_i = \sum_{k,l} \langle h_{k,l} x^{k,l}, p_i \rangle. \tag{19}$$

Let $x_f[m,n]$ be the image filtered by $h_{kl}[m,n]$. The covariance matrix $\boldsymbol{N}$ of the noise vector $\boldsymbol{n}$ has the following entries

$$N_{ij} = \delta_{ij} \sum_{(m,n)\in\mathcal{S}_i} \alpha^2[m,n] x_f^2[m,n]. \tag{20}$$

We can decompose the random matrix $\boldsymbol{A}$ as we did in Section III-B

$$\boldsymbol{A} = \overline{\boldsymbol{A}} + \tilde{\boldsymbol{A}} \tag{21}$$

where

$$\overline{\boldsymbol{A}} \triangleq E[\boldsymbol{A}] \tag{22}$$

$$\overline{a}_{ij} = \delta_{ij} \sum_{(m,n)\in\mathcal{S}_i} h_{0,0}[m,n]\alpha^2[m,n] \tag{23}$$

$$\tilde{a}_{ij} = \sum_{k,l} \left\langle h_{k,l} p_j^{k,l}, p_i \right\rangle - \overline{a}_{ij}. \tag{24}$$

Therefore, $\overline{\boldsymbol{A}}$ is deterministic and $\tilde{\boldsymbol{A}}$ is a zero-mean random matrix. The second-order moments of the elements of $\tilde{\boldsymbol{A}}$ are

$$\begin{aligned}\operatorname{var}(\tilde{a}_{ij}) = {} & \delta_{ij} \sum_{(m,n)\in\mathcal{S}_i} h_{0,0}^2[m,n]\alpha^4[m,n](E[s^4]-1) \\ & + \sum_{(k,l)\neq(0,0)} \sum_{(m,n)\in\mathcal{S}_i\cap\mathcal{S}_j^{k,l}} h_{k,l}^2[m,n] \\ & \cdot \alpha^2[m,n]\alpha^2[m-k,n-l]\end{aligned} \tag{25}$$

$$E[\tilde{a}_{ij}\tilde{a}_{kl}] = 0 \quad \forall (i,j) \neq (k,l) \tag{26}$$

$$E[\tilde{a}_{ij} n_k] = 0 \quad \forall (i,j), k \tag{27}$$

where $\mathcal{S}_j^{k,l} \triangleq \{(m,n) | (m-k, n-l) \in \mathcal{S}_j\}$. The product $\tilde{\boldsymbol{A}}\boldsymbol{b}$ (that we will denote by $\boldsymbol{n}_A$) is a zero-mean random vector whose components are uncorrelated since the entries of the noise matrix $\tilde{\boldsymbol{A}}$ are zero-mean, uncorrelated and independent of $\boldsymbol{b}$. Furthermore, this random vector is uncorrelated with the noise vector $\boldsymbol{n}$. Let $\boldsymbol{\Phi} = E[\boldsymbol{n}_A \boldsymbol{n}_A^T]$ be the covariance matrix of $\boldsymbol{n}_A$. Its components are

$$\phi_{ij} = \delta_{ij} \sum_{k=0}^{L-1} b_k^2 \operatorname{var}(\tilde{a}_{ik}). \tag{28}$$

Now we can define the aggregate noise $\boldsymbol{n}_T = \boldsymbol{n}_A + \boldsymbol{n}$ whose covariance matrix is the sum of two diagonal matrices

$$\boldsymbol{\Gamma} \triangleq E[\boldsymbol{n}_T \boldsymbol{n}_T^T] = \boldsymbol{\Phi} + \boldsymbol{N}. \tag{29}$$

Again we can observe that for a given image and fixed sets $\{\mathcal{S}_i\}$, the elements of the covariance matrix $\boldsymbol{\Phi}$ can be minimized by choosing $f_s(s)$ to be a two-level discrete distribution defined at $\{-1,+1\}$, since in this case $E[s^4] = 1$. If we use a different distribution in order to increase uncertainty when the secret key is not known, then a penalty in noise variance will result. The equivalent channel, after grouping the noise contributions together in a single vector is

$$\boldsymbol{r} = \overline{\boldsymbol{A}}\boldsymbol{b} + \boldsymbol{n}_T. \tag{30}$$

It is interesting to note that if the sets $\{\mathcal{S}_i\}$ are fixed, both matrices $\overline{\boldsymbol{A}}$ and $\boldsymbol{\Gamma}$ are diagonal even when the watermarked image has been linearly filtered.

D. Key-Dependent Pulse Spatial Location

The formulas we have obtained so far are conditioned to fixed sets $\mathcal{S}_0, \cdots, \mathcal{S}_{L-1}$. As we stated in previous paragraphs, we propose the use of key-dependent pulses spread over the image in order to provide high spatial uncertainty when the secret key is not known, as well as to achieve higher resilience to cropping. Spread pulses can be generated by randomly assigning each pixel of the image to one of the sets $\mathcal{S}_0, \cdots, \mathcal{S}_{L-1}$. Maximum uncertainty is achieved if the probability of assigning the pixel (m,n) to each set $\mathcal{S}_i$ is $1/L$ and this probability is independent of the assignments performed on the rest of the pixels. In Appendix A, we prove that the matrices $\overline{\boldsymbol{A}}$ and $\boldsymbol{\Gamma}$ corresponding to the resulting equivalent vector channel are

$$\overline{a}_{ij} = \delta_{ij} \frac{1}{L} \sum_{m,n} h_{0,0}[m,n]\alpha^2[m,n] \tag{31}$$

$$\begin{aligned}\gamma_{ii} = {} & \frac{1}{L} \sum_{m,n} \alpha^2[m,n] x_f^2[m,n] \\ & + b_i^2 \frac{1}{L} \sum_{m,n} h_{0,0}^2[m,n]\alpha^4[m,n]\left(E[s^4]-1\right) \\ & + \frac{\sum_j b_j^2}{L^2} \sum_{(k,l)\neq(0,0)} \sum_{m,n} h_{k,l}^2[m,n] \\ & \cdot \alpha^2[m,n]\alpha^2[m-k,n-l] \\ & + b_i^2 \frac{L-1}{L^2} \sum_{m,n} h_{0,0}^2[m,n]\alpha^4[m,n]\end{aligned} \tag{32}$$

$$\begin{aligned}\gamma_{ij} = {} & -b_i b_j \frac{1}{L^2} \sum_{m,n} h_{0,0}^2[m,n]\alpha^4[m,n], \\ & i \neq j.\end{aligned} \tag{33}$$

Note that even though the covariance matrix of the noise conditioned to a fixed set $\mathcal{S}$ is diagonal, the covariance matrix considering all the possible sets $\mathcal{S}$ is no longer diagonal. However, simulations show that the cross-covariance terms are small compared to the terms in the diagonal, and therefore, they can be neglected.

E. Application to Wiener Filtering

Linear filtering can be used as an attack but it can also be useful for improving the performance of the watermark detection and decoding processes. In fact, if a minimum mean square error (MMSE) linear estimate of the original image is subtracted from the watermarked image, the signal-to-noise ratio (SNR) in the coefficients r_i is substantially improved. Assume that the adaptive one-tap Wiener filter presented in [17], normally used for image restoration, is applied. Then, the preprocessed signal is

$$z[m,n] = y[m,n] - \left[\frac{\sigma_x^2[m,n]}{\sigma_x^2[m,n]+\sigma_w^2[m,n]} \cdot (y[m,n] - E[y[m,n]]) + E[y[m,n]]\right]. \tag{34}$$

The expectation $E[y[m,n]]$ is actually estimated by means of a moving average filter. Hence, it can be included as part of the resulting filter. The variance $\sigma_x^2[m,n]$ is also estimated from the watermarked image as $\hat{\sigma}_x^2[m,n] = \hat{\sigma}_y^2[m,n] - \alpha^2[m,n]$. Therefore, the resulting filter kernel is

$$h_{i,j}[m,n] = \frac{\alpha^2[m,n]}{\hat{\sigma}_y^2[m,n]}\left(\delta[i,j] - \frac{1}{N}\sum_{k,l}\delta[i-k,j-l]\right) \tag{35}$$

where the summation corresponds to the moving average and is defined on a N-pixels region around the origin. Note that this is not the optimal MMSE linear estimate of $w[m,n]$ because it implicitly assumes that $x[m,n]$ is white, and this is not true. However, simulations show that this simple filter improves the performance of the detection process. The statistics of the resulting equivalent channel can be computed by substituting the filter kernel (35) into (31)–(33)

$$\overline{a}_{ij} = \delta_{ij}\frac{1}{L}\frac{N-1}{N}\sum_{m,n}\frac{\alpha^4[m,n]}{\hat{\sigma}_x^2[m,n]+\alpha^2[m,n]} \tag{36}$$

$$\gamma_{ii} = \frac{1}{L}\sum_{m,n}\frac{\alpha^4[m,n]}{(\hat{\sigma}_x^2[m,n]+\alpha^2[m,n])^2} \cdot \left[\left(\frac{N-1}{N}\right)^2\alpha^2[m,n]\left(E[s^4]-\frac{1}{L}\right) + \left(x[m,n]-\frac{1}{N}\sum_{k,l}x[m-k,n-l]\right)^2 + \frac{1}{N^2}\sum_{(k,l)\neq(0,0)}\alpha^2[m-k,n-l]\right] \tag{37}$$

$$\gamma_{ij} = -\frac{1}{L^2}\left(\frac{N-1}{N}\right)^2\sum_{m,n} \cdot \frac{\alpha^8[m,n]}{(\hat{\sigma}_x^2[m,n]+\alpha^2[m,n])^2}, \qquad i \neq j. \tag{38}$$

IV. Channel Coding

A. Detector Structure

Once we have obtained a statistical characterization of the observation vector $\boldsymbol{r}$, we can study detector structures to decode hidden information. Assume that each message i in an alphabet of size $M = 2^L$ is mapped onto a vector $\boldsymbol{b}(i) = (b_0(i)\cdots b_{L-1}(i))$, where $b_j(i) = \pm 1, j = \{0,\cdots,L-1\}, i = \{1,\cdots,M\}$. Therefore, L-bit messages can be hidden. The ML detector for the vector channel analyzed in Sections III-B–III-D chooses the codeword $\boldsymbol{b}(i)$ which maximizes the Gaussian pdf

$$f(\boldsymbol{r}|\boldsymbol{b}(i)) = (2\pi)^{-L/2}|\boldsymbol{\Gamma}|^{-1/2} \cdot \exp\left\{-\tfrac{1}{2}(\boldsymbol{r}-\boldsymbol{c}(i))^T\boldsymbol{\Gamma}^{-1}(\boldsymbol{r}-\boldsymbol{c}(i))\right\} \tag{39}$$

where $\boldsymbol{\Gamma}$ is the covariance matrix of $\boldsymbol{n}_T$, and $\boldsymbol{c}(i) = \overline{\boldsymbol{A}}\boldsymbol{b}(i), i = \{1,\cdots,M\}$. This is equivalent to maximize

$$\boldsymbol{r}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{c}(i) - \tfrac{1}{2}\boldsymbol{c}^T(i)\boldsymbol{\Gamma}^{-1}\boldsymbol{c}(i). \tag{40}$$

The resulting decision regions $\{\mathcal{D}_i\}$ are

$$\mathcal{D}_i = \{\boldsymbol{r}|\boldsymbol{r}^T\boldsymbol{\Gamma}^{-1}\overline{\boldsymbol{A}}(\boldsymbol{b}(i)-\boldsymbol{b}(j) > (\boldsymbol{b}(i)+\boldsymbol{b}(j))^T\overline{\boldsymbol{A}}^T\boldsymbol{\Gamma}^{-1}\overline{\boldsymbol{A}}(\boldsymbol{b}(i)-\boldsymbol{b}(j) \quad \forall j \neq i\}. \tag{41}$$

As we have seen in previous sections, when the pulse locations defined by the sets $\{\mathcal{S}_i\}$ are fixed and independent of the key K, the matrices $\boldsymbol{\Gamma}$ and $\overline{\boldsymbol{A}}$ are diagonal, and the decision regions of the ML detector for the simple code above proposed are defined by the coordinate hyperplanes. The detector is thus quite simple, since it is equivalent to a bit-by-bit hard decoder.

When key-dependent pulse locations are assumed, $\boldsymbol{\Gamma}$ is no longer diagonal. Therefore, the decision regions are intricate and the computational complexity increases. We can, alternatively, use a bit-by-bit hard decoder. The use of this detector structure is reasonable because, even though it is suboptimal in this case, it does not considerably degrade the performance since the cross-covariance terms in $\boldsymbol{\Gamma}$ are in practice small compared to the terms in the diagonal. Furthermore, it is simple to implement and independent of $\boldsymbol{\Gamma}$ and $\overline{\boldsymbol{A}}$.

Let $\overline{a}, \gamma$ be any of the elements in the diagonal of $\overline{\boldsymbol{A}}$ and $\boldsymbol{\Gamma}$, respectively. Considering the structure of the binary antipodal code proposed above, we can prove that in both the key-dependent and the key-independent pulse location cases the probability of bit error averaged over all the keys for a given image and the bit-by-bit detector is given by

$$P_b = Q\left(\frac{\overline{a}}{\sqrt{\gamma}}\right) \tag{42}$$

where

$$Q(x) \triangleq \frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-(t^2/2)}\,dt. \tag{43}$$

If we analyze (31) and (32) we can see that for a fixed image size, the ratio $\overline{a}^2/\gamma$ (which may be regarded as an SNR), decreases with L. In other words, P_b decreases with the pulse size. Therefore, in order to achieve a certain P_b a minimum pulse size is required, or equivalently, a maximum number of bits per pixel is allowed.

B. Attacks

We can use (31)–(33) to analyze the effects of different kinds of attacks on the watermarking scheme. When the watermarked image is cropped, for example, some points of the modulation pulses will be lost. Hence, the SNR for each r_i will decrease and the probability of bit error will degrade. If spread pulses with key-dependent pulse locations are used, then the probability of bit error will increase in the same amount for all bits and errors affecting different bits are approximately independent. These characteristics facilitate the design of good binary coding schemes at bit level.

The watermarked image could also be attacked by adding zero-mean white noise. If the noise variance at pixel (m,n) is $\sigma_n^2[m,n]$ and the noise is added before the linear filtering operation, then we can analyze the effect of this attack just by adding to (32) the term

$$\frac{1}{L}\sum_{m,n}\alpha^2[m,n]\sum_{k,l}h_{k,l}^2[m,n]\sigma_n^2[m-k,n-l]. \tag{44}$$

A worst-case attack of this kind is the addition of noise shaped by the perceptual mask used to generate the watermark, i.e. $\sigma_n^2[m,n]=\alpha^2[m,n],\forall(m,n)$. This attack is studied in Section VI.

V. Sync Recovery and Watermark Detection

A. Detector Structure

Throughout the analysis in Sections III and IV we have assumed that the exact location of the pulses was known. However, several kinds of attacks such as cropping and affine transforms may change the spatial location of the watermark. The synchronization recovery algorithm is in fact intimately related to the watermark detection test. When it succeeds/fails to acquire synchronization, we can infer that the image is watermarked/not watermarked with the given key. In the sequel we will consider both tests as equivalent processes.

Suppose that the watermarked image may have suffered a geometric transformation $T(\cdot,\boldsymbol{\xi})$ with unknown parameters $\boldsymbol{\xi}$ for which we do not assume any *a priori* distribution. The watermark detection test can be formulated as the binary hypothesis test

$$\begin{aligned} H_1&:\ z[m,n]=T(x[m,n]+w[m,n],\boldsymbol{\xi})\\ H_0&:\ z[m,n]=T(x[m,n],\boldsymbol{\xi}). \end{aligned} \tag{45}$$

The performance of this test is measured by the probability of false alarm (P_F) and the probability of detection (P_D). The former is the probability of an arbitrary nonwatermarked image yielding a positive result in the watermark detection test and the latter is the probability of getting a positive result with an image that is in fact watermarked. It is crucial for the credibility of the watermarking system to fix a very low P_F. In fact, the goodness of the system can be measured in terms of the P_D guaranteed for a certain P_F.

As we have already stated, we will reduce the observation space to the projection onto the subspace spanned by the pulses $\{p_i\}$. Therefore, the watermark detection test will be based on the coefficients r_i and we will use the models defined in previous sections in our derivations. A uniformly most powerful test (UMP) [18] does not exist for the binary hypothesis test in (45). Alternatively, we can perform the test independently for each value of $\boldsymbol{\xi}$ and finally decide H_1 if the test yielded a positive result for at least one of those values. This technique can be expressed as the following test:

$$\Lambda_g(z)=\max_{\boldsymbol{\xi}}\ \frac{f(\boldsymbol{r}|\boldsymbol{\xi},H_1)}{f(\boldsymbol{r}|\boldsymbol{\xi},H_0)}\underset{H_0}{\overset{H_1}{\gtrless}}H_1 \tag{46}$$

where $f(\boldsymbol{r}|\boldsymbol{\xi},H_1)$ is the pdf of $r_i(\boldsymbol{\xi})=\langle z,T(p_i,\boldsymbol{\xi})\rangle$ assuming that the image is watermarked and has suffered the transformation $T(\cdot,\boldsymbol{\xi})$. When a geometric transformation $T(\cdot,\boldsymbol{\xi})$ is applied to the watermarked image, the perceptual mask suffers approximately the same transformation, i.e., $T(p_i,\boldsymbol{\xi})\simeq T(\alpha,\boldsymbol{\xi})T(s,\boldsymbol{\xi})$. For this reason, if $\alpha[m,n]$ is obtained from the image under test, the components of $\boldsymbol{r}(\boldsymbol{\xi})$ are actually computed as $r_i(\boldsymbol{\xi})=\langle z,\alpha T(s,\boldsymbol{\xi})\rangle$.

As we did in previous sections, we assume that the key K is the only random variable in the watermarking model and that all the keys are equiprobable. Therefore, we will measure P_D as the probability of getting a key that yields a positive result in the watermark detection test when the image has been watermarked with that key, and P_F as the probability of getting a key that yields a positive result when the image has not been watermarked. A detection test will be designed specifically for each image, fixing a threshold value that guarantees a desired P_F. However, the P_D achievable in each case depends on the characteristics of the image under test. In fact, P_D indicates the suitability of each image for being watermarked. Small images, for instance, will lead to poor P_D values and will be bad candidates for watermarking. For any image under test, the decision should be accompanied with the corresponding P_D, which should be considered as a measure of the confidence level for that decision, assuming that a certain P_F is guaranteed. Our goal in this section is to provide expressions that can be used both to fix thresholds to achieve a desired P_F for any image and to measure from either the original or a watermarked image the P_D that can be expected in each test.

The pdf under hypothesis H_1 can be decomposed as

$$f(\boldsymbol{r}|\boldsymbol{\xi},H_1)=\frac{1}{M}\sum_{i=1}^{M}f(\boldsymbol{r}|\boldsymbol{b}(i),\boldsymbol{\xi},H_1). \tag{47}$$

We will limit our analysis to transformations consisting in integer shifts (e.g., cropping). Then, for every $\boldsymbol{\xi}$ each conditional pdf can be approximated by a Gaussian pdf with mean $\overline{\boldsymbol{A}}\boldsymbol{b}(i)$ and covariance matrix $\boldsymbol{\Gamma}$. The pdf under hypothesis H_0 is approximated by a zero-mean Gaussian vector pdf with

covariance matrix $\boldsymbol{N} = \sigma^2 \boldsymbol{I}_L$, where

$$\sigma^2 = \frac{1}{L}\sum_{m,n} \alpha^2[m,n] x_f^2[m,n]. \tag{48}$$

We will assume that L_s pulses are reserved for synchronization purposes and are thus modulated by known coefficients (assume $+1$)

$$w[m,n] = \sum_{i=0}^{L_s-1} p_i[m,n] + \sum_{i=L_s}^{L-1} b_i p_i[m,n]. \tag{49}$$

If we neglect the cross-covariance terms in $\boldsymbol{\Gamma}$, we get, after some algebra, the following expression for the log maximum likelihood function $l(z) \triangleq \ln \Lambda_g(z)$

$$\begin{aligned} l(z) = \max_{\boldsymbol{\xi}} & \frac{L}{2} \ln \frac{\sigma^2}{\gamma} - \frac{\overline{a}^2 L}{2\gamma} \\ & - \frac{1}{2}\left(\frac{1}{\gamma} - \frac{1}{\sigma^2}\right) \sum_{i=0}^{L-1} r_i^2(\boldsymbol{\xi}) + \frac{\overline{a}}{\gamma} \sum_{i=0}^{L_s-1} r_i(\boldsymbol{\xi}) \\ & + \sum_{i=L_s}^{L-1} \ln\left(\cosh\left(\frac{\overline{a} r_i(\boldsymbol{\xi})}{\gamma}\right)\right) \underset{H_0}{\overset{H_1}{\gtrless}} \eta \end{aligned} \tag{50}$$

where $\overline{a}$ and γ are the same as in (42). We will approximate P_D by the probability of exceeding the threshold assuming H_1 is true and the estimate of $\boldsymbol{\xi}$ is correct. When P_F is very low, the probability of exceeding the threshold for other values of $\boldsymbol{\xi}$ when H_1 is true is negligible. Therefore, the approximation is reasonable. We will define P_F as the probability of a nonwatermarked image exceeding the threshold for $\boldsymbol{\xi} = \boldsymbol{\theta}$ i.e., when no transformation is assumed. When the image is not watermarked, any of the points $\boldsymbol{\xi}$ examined during the maximization in (46) can lead to a false alarm event. Hence, the actual P_F can be approximated by further multiplying the P_F obtained in the derivations by the size of the search space. Let $\mu(s) = \ln E[e^{s\boldsymbol{l}(\boldsymbol{r})}|H_0]$. Then

$$\begin{aligned} \mu(s) = & \left(\frac{L}{2}\ln\frac{\sigma^2}{\gamma} - \frac{\overline{a}^2 L}{2\gamma}\right)s + (L - L_s) \\ & \cdot \ln E\left[e^{-s(r^2/2)((1/\gamma)-(1/\sigma^2))} \cosh^s\left(\frac{r\overline{a}}{\gamma}\right)\right] \\ & + L_s \ln E\left[e^{-s(r^2/2)((1/\gamma)-(1/\sigma^2))} e^{s(r\overline{a}/\gamma)}\right] \end{aligned} \tag{51}$$

where $r \sim N(\overline{a}, \sigma^2)$. It is assumed that $b_i = 1$ for $i \in \{0, \cdots, L_s - 1\}$. The Chernoff bound for probabilities P_F and P_D is [18]

$$P_F \leq e^{\mu(s) - s\dot{\mu}(s)}, \qquad s > 0 \tag{52}$$

$$P_D \geq 1 - e^{\mu(s) + (1-s)\dot{\mu}(s)}, \qquad s < 1 \tag{53}$$

and $\dot{\mu}(s) = \eta$, where $\dot{\mu}(s)$ is the first derivative of $\mu(s)$ with respect to s. Using central limit theorem arguments a tighter approximation is [18]

$$P_F \simeq \frac{1}{\sqrt{2\pi s^2 \ddot{\mu}(s)}} e^{\mu(s) - s\dot{\mu}(s)} \tag{54}$$

$$P_D \simeq \frac{1}{\sqrt{2\pi (1-s)^2 \ddot{\mu}(s)}} e^{\mu(s) + (1-s)\dot{\mu}(s)} \tag{55}$$

where $\ddot{\mu}(s)$ is the second derivative of $\mu(s)$ with respect to s. When only affine transforms consisting of integer shifts are considered, the maximization in (46) can be implemented by using a brute force search algorithm. This is actually a computationally complex technique, considering that the pulses are spread over the whole image. The use of sequential detection algorithms in the synchronization recovery process is left as an open research line.

When affine transforms that include scaling and rotations are considered, other issues appear. If $s[m,n]$ is i.i.d., its autocorrelation function is a delta function. This means that the peak in the function (50) is very narrow and may be very difficult to find by a brute force searching algorithm. The peak can be smoothed if $s[m,n]$ is nonwhite or, in other words, it has some redundancy. A smoother function reduces the uncertainty when the key is not known. However, it allows the use of a synchronization recovery algorithm in two steps: first, during acquisition, a sequential search is performed over a grid defined in the space of unknown parameters; then a fine adjustment of these parameters is performed by means of an iterative algorithm (e.g., a gradient algorithm). The design of sequences $s[m,n]$ beneficial to the synchronization algorithm is an open research line.

B. Attacks

When an image is cropped, the ratios $\overline{a}^2/\gamma$ and $\overline{a}^2/\sigma^2$ decrease. Therefore, P_D decreases for a fixed P_F. For a maximum cropping factor, it is possible to obtain a minimum pulse size (maximum number of pulses) in order to guarantee resilience to cropping.

Line (or column) removal can considerably reduce the performance of the watermark detection test with little effort. If one line is removed, for instance, two peaks will appear in the function $l(\boldsymbol{r})$ as a function of $\boldsymbol{\xi}$ (see Section V-A). In fact, the amplitude of the original peak will be distributed between those two peaks. One solution is doubling the pulse size in order to guarantee the required P_D in watermark detection and BER in data decoding. Work is in progress to develop synchronization schemes and pulse shapes robust against this kind of attacks.

Additive noise will clearly decrease P_D for a given threshold. The noise power that a hacker can add to a watermarked image is limited. Hence a minimum pulse size can be fixed in order to guarantee a minimum P_D for a required P_F.

VI. Experimental Results and Comparisons

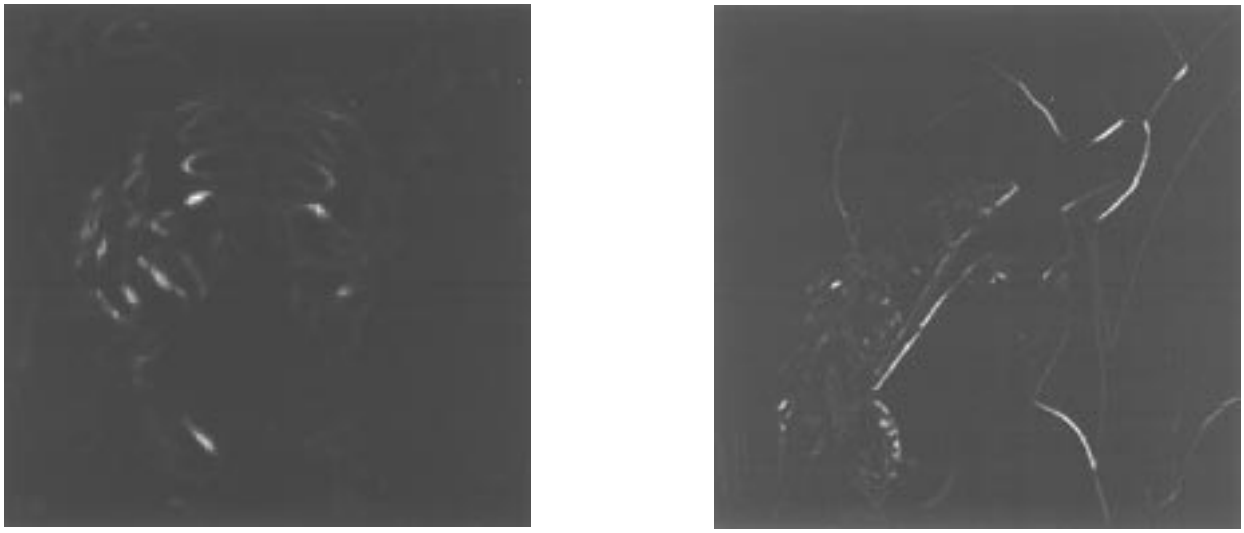
In this section we compare results from simulations to the analytical expressions derived in previous sections. We have performed all the experiments using the gray-level images of different sizes shown in Fig. 3. In Fig. 4, we can see examples of watermarked images. The perceptual mask $\alpha[m,n]$ is based on a visibility function defined in [17]. In Fig. 5, we show the perceptual mask corresponding to the images under study. The marginal distribution of $s[m,n]$ used in the experiments is a symmetrical discrete distribution taking values in $\{\pm\frac{1}{2}\sqrt{8/5}, \pm\sqrt{8/5}\}$.

(a) (b)

Fig. 3. Original images used in the experiments.

(b) (a)

Fig. 4. Watermarked images.



(a) (b)

Fig. 5. Perceptual masks.

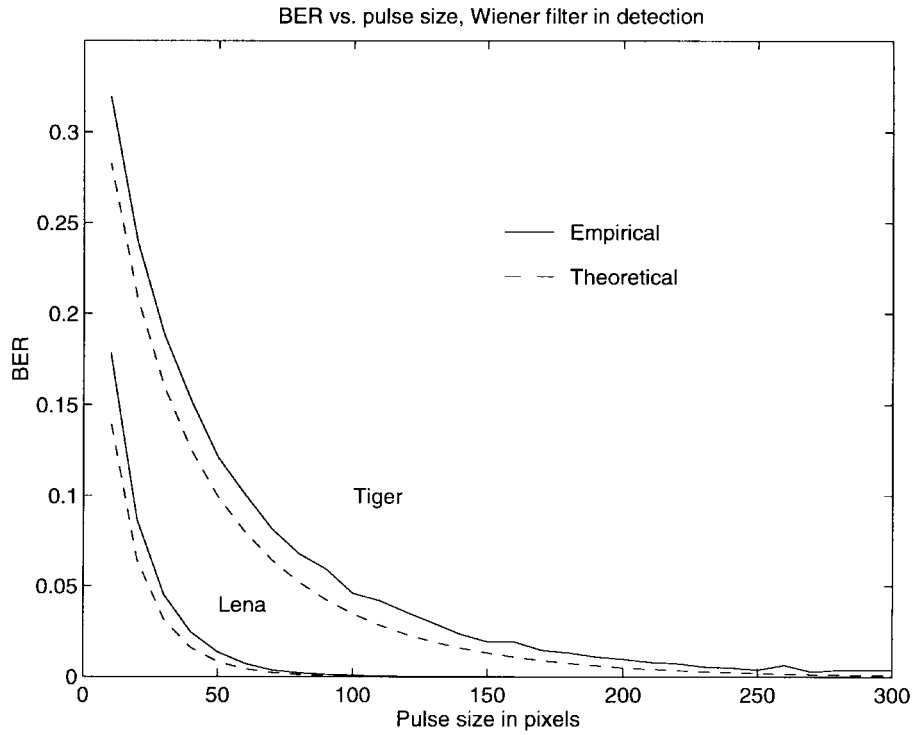


Fig. 6. BER versus modulation pulse size when estimation of the original image using an adaptive Wiener filter is performed prior to detection.

A. Data Decoding

The first group of figures evaluate the performance of the data-hiding scheme using the detector discussed in Section IV. In all the cases studied we have obtained the empirical curves taking 100 keys at random. In Fig. 6, we can see plots of the BER obtained both empirical and analytically for different pulse sizes when the watermarked image is unaltered and Wiener filtering is applied for noise reduction as a preprocessing step (Section III-E). The empirical BER is slightly above the theoretical BER due to the additional variance term that in practice appears because the coefficients of the Wiener filter are computed from the watermarked image, which is key-dependent. Work is in progress for contemplating this effect and obtaining better estimates of the statistics of the equivalent channel.

In Fig. 7, we show similar plots when low pass filtering rather than Wiener filtering is performed before detection for estimation of the original image. We can see that the performance has substantially degraded with respect to the previous plots. In this case, the filter used in demodulation is independent of the key, and therefore the additional variance discussed in the previous paragraph does not appear.

In Fig. 8, we can observe the effect of an attack based on worst case additive Gaussian noise shaped by the perceptual mask. Wiener filtering is performed prior to detection for noise reduction. The BER has increased slightly with respect to the first plots. These curves can be used to choose a conservative pulse size that provides robustness against additive noise. The increase in variance due to the key-dependent Wiener filter also appears in these plots.

In Fig. 9, plots of the BER are shown for a Wiener filter attack that can be considered as a worst case linear filtering attack aimed at deleting the watermark. As we can see, the BER has not substantially degraded with respect to the nonattacked case (see Fig. 6). This effect is due to the fact that in ideal conditions, linear MMSE estimation of the hidden signal leads to the same results when it is done after performing linear MMSE estimation of the original image (which is actually the attack). We can see that the difference between the theoretical BER and the empirical BER has increased because there is a greater additional variance term not considered in the theoretical derivations, resulting from the use of two consecutive key-dependent filtering operations.

B. Sync Recovery and Watermark Detection

In the last group of figures, we show bounds and approximations to the receiver operating characteristic (ROC) when the watermarked images do not suffer any attack. Curves obtained from simulation results are also shown. The P_F is so small that it cannot be estimated through experimentation. Hence the empirical curves actually represent the empirical P_D and the analytical approximation to P_F evaluated over a range of threshold values. In all the cases, the experiments have been performed taking 400 keys at random. The parameters $\overline{a}, \gamma, \sigma^2$ computed and used in the tests correspond to pessimistic SNR values. Hence, the theoretical P_F obtained from these parameters is an upper bound with respect to the actual P_F.

In Fig. 10, we plot the Chernoff bound and an approximation to the theoretical ROC for the Tiger image for 20 pulses, none of them reserved for synchronization purposes, when low-pass filtering is used to estimate the original image before detection. Note that the approximation for P_D is very close to the empirical P_D.

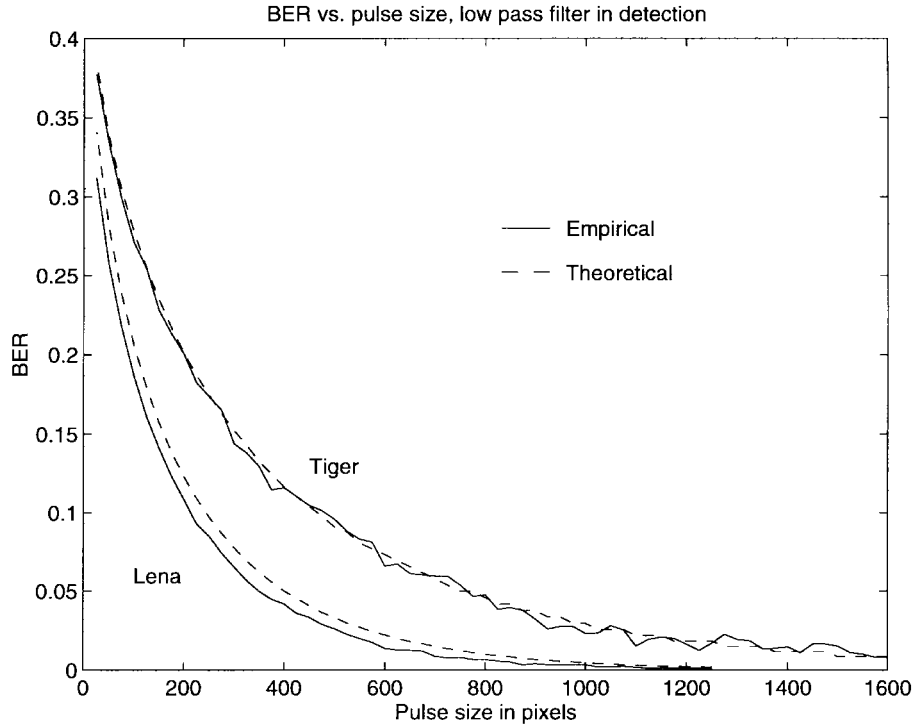


Fig. 7. Bit error rate versus modulation pulse size when estimation of the original image using a low pass filter is performed prior to detection.

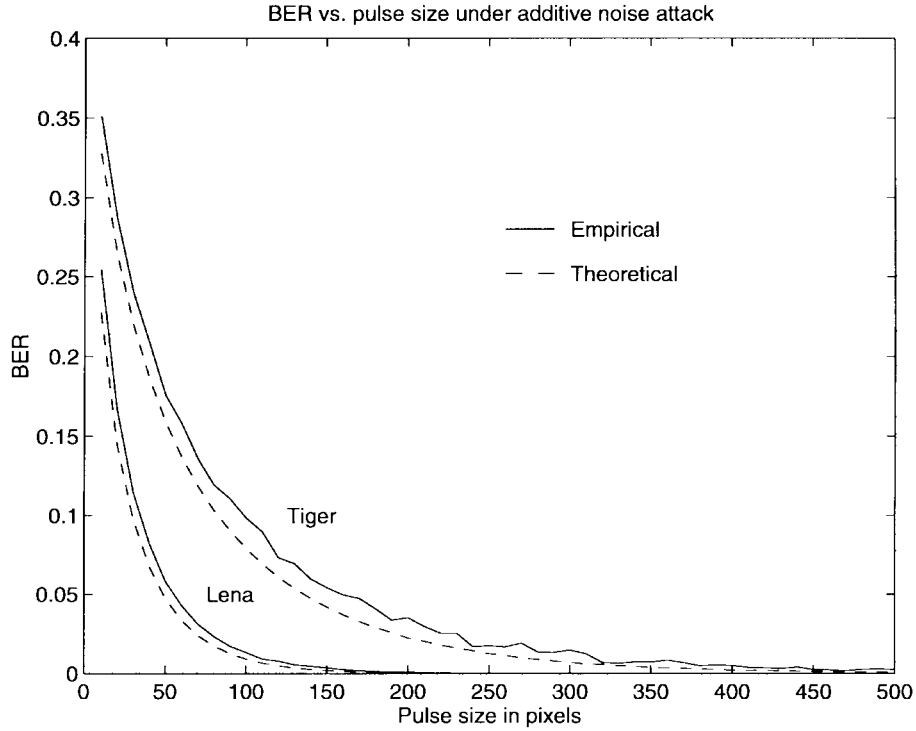


Fig. 8. Bit error rate versus modulation pulse size when the watermarked image is attacked with worst case additive noise and Wiener filtering is performed prior to detection.

In Fig. 11, we show an approximation to the theoretical ROC for the Tiger image when only one pulse covering the whole image is used as a watermark, and assuming that Wiener filtering is used to estimate the original image before detection. In this simple case the theoretical P_F and P_D can be computed exactly for the parameters $\overline{a}, \gamma, \sigma^2$ estimated using the analytical expressions, because $\boldsymbol{r}$ is one dimensional. Even though the distance between the theoretical and empirical curves has increased due to the additional variance term introduced by the Wiener filter (Section VI-A), the theoretical

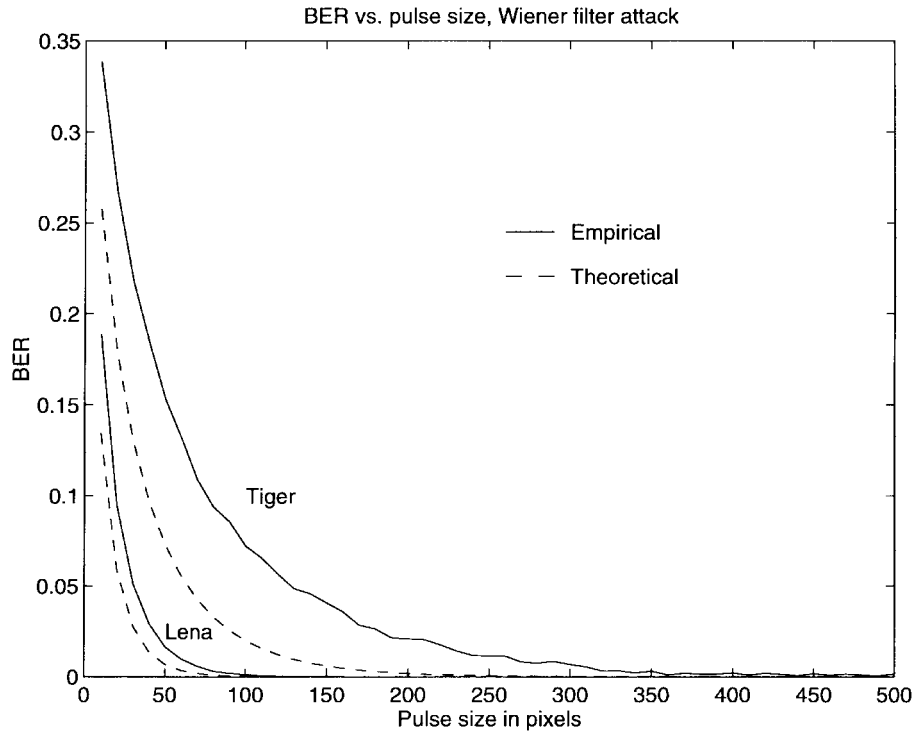


Fig. 9. BER versus modulation pulse size when the watermarked image is attacked with Wiener filtering and Wiener filtering is performed prior to detection.

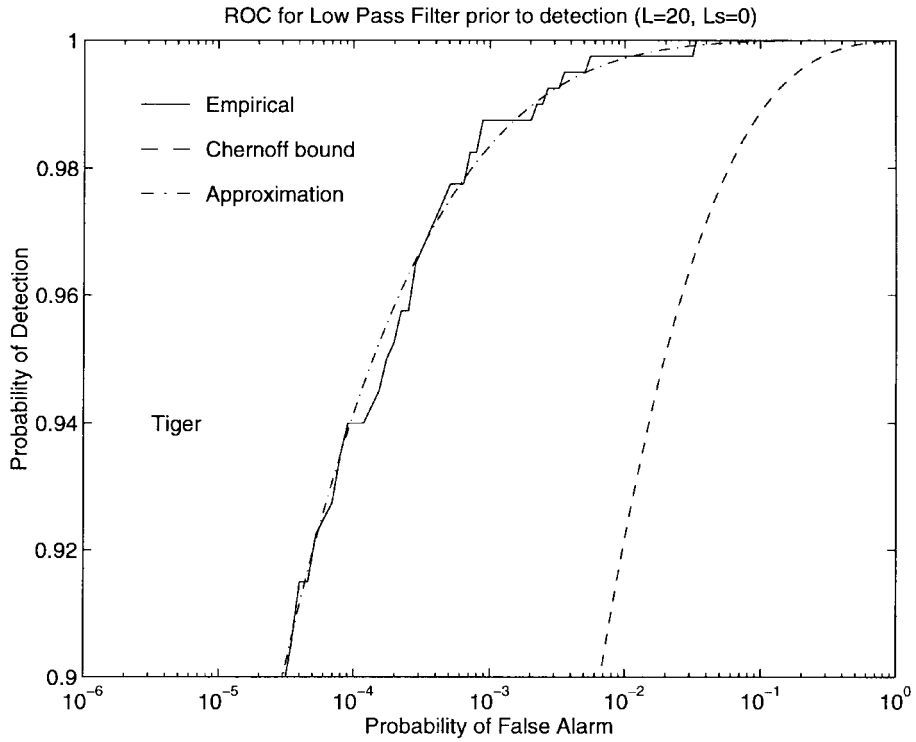


Fig. 10. Bound and approximation to the ROC of the watermark detection test with low-pass filtering before detection.

P_F is still an upper bound, since the parameters $\overline{a}, \gamma, \sigma^2$ are computed pessimistically. We can see that the P_F has substantially improved because a better estimation of the original image is performed and only one pulse with no additional information is used.

Using the approximations to the ROC, it is possible to fix adequate values of the threshold for a desired P_F. Similar approximations can be used to obtain values of P_D and P_F achievable under different kinds of attacks. In fact, since the attacker should not considerably degrade the image, a worst-

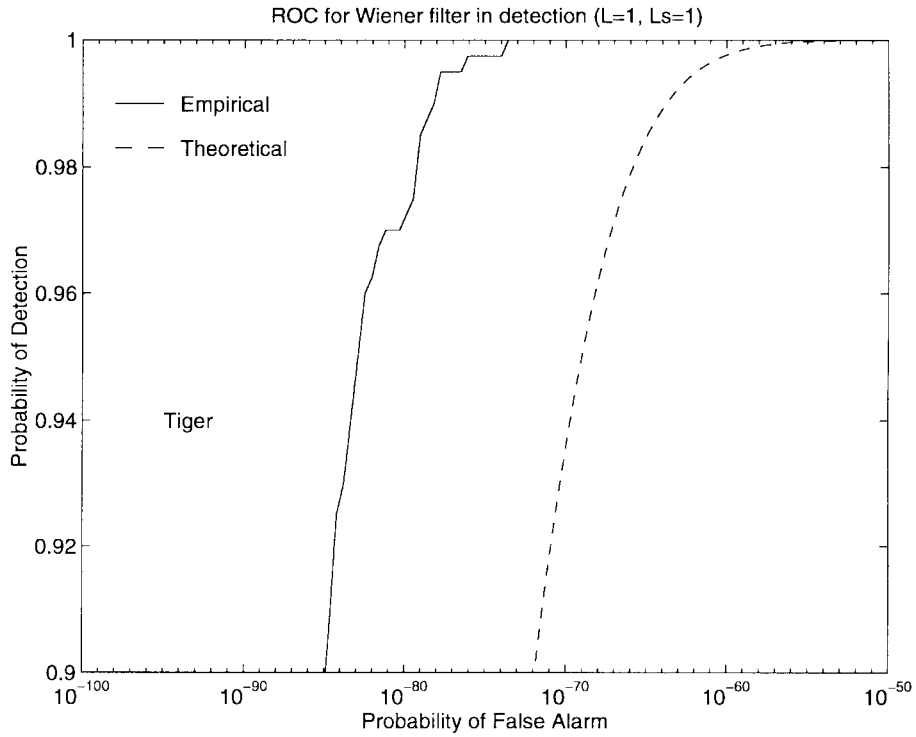


Fig. 11. Approximation to the ROC of the watermark detection test with pure watermarking and Wiener filtering before detection.

case ROC exists that can be considered as the achievable performance for any attack.

VII. Conclusions and Further Work

In this paper we have introduced the theoretical analysis of a data hiding and watermarking system. As a result of this analysis, we have derived detector structures and expressions for performance measures such as the BER and the ROC associated with a given original image when alterations such as additive noise, cropping, and linear filtering are possible. These expressions can be used to fix parameters such as the number of pulses and the watermark detection threshold necessary to achieve a desired level of performance. Moreover, performance measures can be obtained prior to watermarking and can thus be used to estimate the capacity of an image for information hiding purposes.

Promising lines of research are the design of channel codes for small pulse sizes in order to approach the overall information capacity of the image, the use of sequential detection techniques [19] and new pulse shapes to improve the watermark detection algorithm in terms of computational complexity and robustness against scaling and rotations, and the application of the analytical approach discussed in this paper to transformed domains (DCT, DFT, wavelets, etc.) and color images.

The most challenging research line is the definition of a theoretical framework in terms of Shannon information theory. In a watermarking system, problems related to source coding, channel coding and cryptographic security appear. Hence, a careful analysis should combine these three fields.

Appendix

A. Equivalent Channel for Key-Dependent Pulse Locations

In this appendix we derive the first- and second-order moments of the equivalent vector channel when spread pulses with key-dependent pulse locations are used and the watermarked image is linearly filtered. Assume that the codeword $(b_0, \cdots, b_{L-1})$ is hidden. Then

$$E[r_i] = E_{\mathcal{S}}[E_{\boldsymbol{r}}[r_i|\mathcal{S}]] \tag{56}$$

$$\operatorname{var}(r_i) = E_{\mathcal{S}}[\operatorname{var}_{\boldsymbol{r}}(r_i|\mathcal{S})] + \operatorname{var}_{\mathcal{S}}(E_{\overline{\boldsymbol{r}}}[r_i|\mathcal{S}]) \tag{57}$$

$$\begin{aligned}\operatorname{cov}(r_i, r_j) &= E_{\mathcal{S}}[\operatorname{cov}_{\boldsymbol{r}}(r_i, r_j|\mathcal{S})] \\ &\quad + \operatorname{cov}_{\mathcal{S}}(E_{\boldsymbol{r}}[r_i|\mathcal{S}], E_{\boldsymbol{r}}[r_j|\mathcal{S}]) \\ &= E_{\mathcal{S}}[E_{\boldsymbol{r}}[r_i|\mathcal{S}]E_{\boldsymbol{r}}[r_j|\mathcal{S}]] \\ &\quad - E_{\mathcal{S}}[E_{\boldsymbol{r}}[r_i|\mathcal{S}]]E_{\mathcal{S}}[E_{\boldsymbol{r}}[r_j|\mathcal{S}]]\end{aligned} \tag{58}$$

where $E_{\boldsymbol{r}}[r_i|\mathcal{S}]$ and $\operatorname{var}_{\overline{\boldsymbol{r}}}(r_i|\mathcal{S})$ correspond to the formulas for $\overline{a}_{ii}$ and γ_{ii} shown in Section III-C. The expected value of the ith component of the received vector $\boldsymbol{r}$ is

$$\begin{aligned}E_{\mathcal{S}}[E_{\boldsymbol{r}}[r_i|\mathcal{S}]] &= \sum_{\mathcal{S}} E_{\boldsymbol{r}}[r_i|\mathcal{S}]\Pr\{\mathcal{S}\} \\ &= b_i \sum_{m,n} h_{0,0}[m,n]\alpha^2[m,n]\Pr\{(m,n) \in \mathcal{S}_i\}.\end{aligned} \tag{59}$$

Hence

$$\overline{a}_{ij} = \delta_{ij}\frac{1}{L}\sum_{m,n} h_{0,0}[m,n]\alpha^2[m,n]. \tag{60}$$

The variance of r_i is

$$\begin{aligned}\operatorname{var}(r_i) &= \sum_{\mathcal{S}} \operatorname{var}(r_i|\mathcal{S})\Pr(\mathcal{S}) + \sum_{\mathcal{S}} E^2_{\boldsymbol{r}}[r_i|\mathcal{S}]\Pr(\mathcal{S}) \\ &\quad - \left(\sum_{\mathcal{S}} E_{\boldsymbol{r}}[r_i|\mathcal{S}]\Pr(\mathcal{S})\right)^2 \\ &= \sum_{m,n} \alpha^2[m,n]x_f^2[m,n]\Pr\{(m,n)\in\mathcal{S}_i\} \\ &\quad + b_i^2 \sum_{m,n} h_{0,0}^2[m,n]\alpha^4[m,n]\big(E\big[s^4\big]-1\big) \\ &\quad \cdot \Pr\{(m,n)\in\mathcal{S}_i\} + \sum_{j=0}^{L-1} b_j^2 \sum_{(k,l)\neq(0,0)} \sum_{m,n} h_{kl}^2[m,n] \\ &\quad \cdot \alpha^2[m,n]\alpha^2[m-k,n-l]\Pr\Big\{(m,n)\in\mathcal{S}_i\cap\mathcal{S}_j^{k,l}\Big\} \\ &\quad + b_i^2 \sum_{m_1,n_1}\sum_{m_2,n_2} \alpha^2[m_1,n_1]h_{0,0}[m_1,n_1] \\ &\quad \cdot \alpha^2[m_2,n_2]h_{0,0}[m_2,n_2] \\ &\quad \cdot \Pr\{(m_1,n_1)\in\mathcal{S}_i,(m_2,n_2)\in\mathcal{S}_i\} \\ &\quad - b_i^2\left(\sum_{m,n} h_{0,0}[m,n]\alpha^2[m,n]\right. \\ &\qquad \left.\cdot \Pr\{(m,n)\in\mathcal{S}_i\}\right)^2 \end{aligned} \tag{61}$$

hence

$$\begin{aligned}\gamma_{ii} &= \frac{1}{L}\sum_{m,n}\alpha^2[m,n]x_f^2[m,n] \\ &\quad + b_i^2\frac{1}{L}\sum_{m,n} h_{0,0}^2[m,n]\alpha^4[m,n]\big(E\big[s^4\big]-1\big) \\ &\quad + \frac{\sum_j b_j^2}{L^2}\sum_{(k,l)\neq(0,0)}\sum_{m,n} h_{k,l}^2[m,n] \\ &\quad \cdot \alpha^2[m,n]\alpha^2[m-k,n-l] \\ &\quad + b_i^2\frac{L-1}{L^2}\sum_{m,n} h_{0,0}^2[m,n]\alpha^4[m,n]. \end{aligned} \tag{62}$$

Following similar arguments, the cross-covariance terms are

$$\begin{aligned}&\operatorname{cov}(r_i,r_j) \\ &\quad = b_i b_j \sum_{m_1,n_1}\sum_{m_2,n_2} h_{0,0}[m_1,n_1]\alpha^2[m_1,n_1] \\ &\qquad \cdot h_{0,0}[m_2,n_2]\alpha^2[m_2,n_2] \\ &\qquad \cdot \Pr\{(m_1,n_1)\in\mathcal{S}_i,(m_2,n_2)\in\mathcal{S}_j\} \\ &\qquad - b_i b_j \sum_{m_1,n_1}\sum_{m_2,n_2} h_{0,0}[m_1,n_1]\alpha[m_1,n_1] \\ &\qquad \cdot h_{0,0}[m_2,n_2]\alpha[m_2,n_2]\Pr\{(m_2,n_2)\in\mathcal{S}_i\} \\ &\qquad \cdot \Pr\{(m_2,n_2)\in\mathcal{S}_j\} \end{aligned} \tag{63}$$

therefore

$$\gamma_{ij} = -b_i b_j \frac{1}{L^2}\sum_{m,n} h_{0,0}^2[m,n]\alpha^4[m,n], \qquad i\neq j. \tag{64}$$

REFERENCES

[1] I. J. Cox, J. Kilian, T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for multimedia," NEC Res. Inst., Princeton, NJ, Tech. Rep. 95-10, 1995.

[2] E. Koch, J. Rindfrey, and J. Zhao, "Copyright protection for multimedia data," in *Digital Media and Electronic Publishing*. New York: Academic, 1996, pp. 203–213.

[3] E. Koch and J. Zhao, "Toward robust hidden image copyright labeling," in *Proc. 1995 IEEE Workshop on Nonlinear Signal and Image Processing*, Neos Marmaras, Greece, June 1995, pp. 452–455.

[4] 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*, R. Oldenbourg, Ed., Vienna, Austria, Aug. 21–25, 1995, pp. 242–251.

[5] M. D. Swanson, B. Zhu, and A. H. Tewfik, "Transparent robust image watermarking," in *Proc. IEEE Int. Conf. on Image Processing*, vol. III, Lausanne, Switzerland, Sept. 1996, pp. 211–214.

[6] M. D. Swanson, B. Zhu, and A. H. Tewfik, "Robust data hiding for images," in *Proc. IEEE Digital Signal Processing Workshop*, Loen, Norway, Sept., 1996, pp. 37–40.

[7] B. Zhu, M. D. Swanson, and A. H. Tewfik, "A transparent robust authentication and distortion measurement technique for images," in *Proc. IEEE Digital Signal Processing Workshop*, Loen, Norway, Sept., 1996, pp. 45–48.

[8] 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.

[9] W. Bender, D. Gruhl, and N. Morimoto, "Techniques for data hiding," in *Proc. SPIE*, San Jose, CA, Feb. 1995, pp. 2420–2440.

[10] F. M. Boland, J. J. K. O. Ruanaidh, and C. Dautzenberg, "Watermarking digital images for copyright protection," in *IEE Int. Conf. on Image Processing and its Applications*, Edinburgh, Scotland, 1995, pp. 326–330.

[11] F. Hartung and B. Girod, "Digital watermarking of raw and compressed video," in *Digital Compression Technologies and Systems for Video Communications*, N. Ohta, Ed., vol. 2952, SPIE Proceedings Series, Oct. 1996, pp. 205–213.

[12] J. R. Smith and B. O. Comiskey, "Modulation and information hiding in images," in *Proc. Int. Workshop on Information Hiding*, Cambridge, UK, May 1996, pp. 207–226.

[13] R. Anderson, "Stretching the limits of steganography," in *Proc. Int. Workshop on Information Hiding*, Cambridge, U.K., May 1996, pp. 39–48.

[14] A. N. Netravali and B. G. Haskell, *Digital Pictures. Representation, Compression and Standards*. New York: Plenum, 1995.

[15] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[16] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*. New York: McGraw-Hill, 1984.

[17] J. S. Lim, *Two-Dimensional Signal and Image Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1990.

[18] H. L. V. Trees, *Detection, Estimation and Modulation Theory*, Pt. I. New York: Wiley, 1968.

[19] M. D. Srinath, P. K. Rajasekaran, and R. Viswanathan, *Introduction to Statistical Signal Processing with Applications*. Englewood Cliffs, NJ: Prentice-Hall, 1996.

Juan R. Hernández (S'97) was born in Salamanca, Spain, on February 12, 1970. He received the Ingeniero de Telecomunicación degree from the University of Vigo, Spain, in 1993 and the M.S. degree in electrical engineering from Stanford University, Stanford, CA, in 1996. Since 1996 he has been working toward the Ph.D. degree at Stanford University, Stanford, CA.

From 1993 to 1995, he was a member of the Department of Communication Technologies, University of Vigo, Spain, where he worked on hardware for digital signal processing and access control systems for digital television. His research interests include digital communications and copyright protection in multimedia.

Fernando Pérez-González (M'93) received the Ingeniero de Telecomunicación degree and the Ph.D. degree in telecommunications engineering from the Universities of Santiago, Spain, in 1990, and Vigo, Spain in 1993.

He joined the Faculty of the School of Telecommunications Engineering as an Assistant Professor in 1990 and is currently Associate Professor in the same institution. He has visited the University of New Mexico, Albuquerque, 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 co-editor of the book *Intelligent Methods in Signal Processing and Communications* (Cambridge, MA: Birkhauser, 1997) and has been guest editor of a special section of *Signal Processing*, devoted to Signal Processing for Communications.

José Manuel Rodríguez was born on April 30, 1973, in Spain. He is currently working toward the Ingeniero de Telecomunicación degree at the University of Vigo, Spain.

His current research interest is in image watermarking.

Gustavo Nieto was born on January 9, 1973 in Spain. He is currently working toward the Ingeniero de Telecomunicación degree at the University of Vigo, Spain.

His current research interest is in image watermarking.