An Asymmetric Subspace Watermarking Method for Copyright Protection

Jengnan Tzeng, Wen-Liang Hwang, and I-Liang Chern

***Abstract*—We present an asymmetric watermarking method for copyright protection that uses different matrix operations to embed and extract a watermark. It allows for the public release of all information, except the secret key. We investigate the conditions for a high detection probability, a low false positive probability, and the possibility of unauthorized users successfully hacking into our system. The robustness of our method is demonstrated by the simulation of various attacks.**

***Index Terms*—Asymmetric watermark, copyright protection.**

NOMENCLATURE

$\mathcal{W}$	Watermark space whose dimension is l.
$\mathcal{V}$	Orthogonal complement of $\mathcal{W}$ in feature space.
$\mathcal{H}$	Subspace of $\mathcal{W}$ whose dimension is $h < l$.
$\mathcal{G}$	Subspace of $\mathcal{W}$ whose dimension is $g < l$.
G	l by g matrix composed of an orthogonal basis of $\mathcal{G}$.
H	l by h matrix composed of an orthogonal basis of $\mathcal{H}$.
D	g by l detection matrix.
w	Watermark with length g.
B	Matrix of size g by h.
ϕ_o	Feature of the original image in the feature space.
ϕ_w	Feature of the watermarked image in the feature space.

I. INTRODUCTION

DIGITAL security information embedded in content, called watermarking, has many applications, including authentication, copyright protection, copy protection, fingerprinting, and broadcasting channel tracking [9], [21], [29], [30], [34]. In this paper, we only focus on asymmetric watermarking for the copyright protection of images. Copyright protection should not be confused with copy protection because, for the latter, one key is given to all recipients, while for copyright protection, each image has its own key. In a symmetric watermarking system, the keys necessary to embed and extract a watermark are secret and identical. The common secret key is a random sequence, which is embedded in an image by the spreading spectrum technique [6]. Notable security problems of the symmetric (one secret key) watermarking approach stem from the need to make the secret key available to owners and recipients, as well as from the need to identify which secret key is associated with which image in a large image database. Another problem is that the watermark is present as evidence of ownership, so it provides an attacker with the knowledge to remove the watermark [4]. The solution to the problem is a watermarking system that satisfies Kirckhoffs' principle [16], which states that a security system must assume that an adversary knows everything about the algorithm, except for the secret keys. Zero-knowledge watermark detection is one approach for resolving this problem [1], [4]. The basic idea is to replace the watermark detection process with a cryptographic protocol. Although this approach shows promise, it requires a great deal of bidirectional communication between owners and verifiers to prove ownership for copyright protection purposes.

Asymmetric watermarking is another approach that satisfies Kerckhoffs' principle. This system uses two sets of keys: one for embedding, and one for detecting. The detecting key is made public so that anyone has access to it and is permitted to use it to verify whether an image is watermarked or not. The public key of each watermarked image is usually stored in a safe place where a trusted third party can verify its integrity. This avoids the problem that anyone could produce a valid public key by his own asymmetric watermarking method. In an asymmetric system, the secret embedding keys are not used for verification. Therefore, no secret information is sent over the channel, nor can it be accessed in the database.

An asymmetric system must ensure that it is almost impossible, or at least computationally impossible, for those who know the entire system, except for the secret key, to successfully hack into the system [18]. Some interesting asymmetric schemes have been proposed for watermarking [10]–[13], [26]–[28]. Hartung and Girod [13] proposed the first asymmetric watermarking method. Furon and Duhamel [12] provide a useful survey of various methods, as well as an in-depth discussion of asymmetric watermarking. Their asymmetric watermarking method is applied to copy protection and their watermark detector must be embedded into detection devices, so that inverting the embedding process becomes computationally difficult. We propose an asymmetric watermarking method for copyright protection whereby all information, except the secret key, is released to the public.

In our previous study of symmetric watermarking, we proposed a robust subspace watermarking method in which our watermark was embedded into a subspace $\mathcal{W}$, where the watermark is resistant to conventional image operations. For our asymmetric watermarking method, we have further split our watermark space $\mathcal{W}$ into two subspaces, $\mathcal{G}$ and $\mathcal{H}$. The column vectors of the secret matrices G and H form an orthogonal basis of subspaces $\mathcal{G}$ and $\mathcal{H}$, respectively. Our published watermark w is embedded into subspace $\mathcal{G}$ by the secret matrix G, while

Manuscript received July 1, 2004; revised August 9, 2004. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Stefan Katzenbeisser.

J. Tzeng and I-L. Chern are with the Department of Mathematics, National Taiwan University, Taiwan, R.O.C.

W.-L. Hwang is with the Institute of Information Science, Academia Sinica, Taiwan, R.O.C. (e-mail: whwang@iis.sinica.edu.tw).

Digital Object Identifier 10.1109/TSP.2004.839921

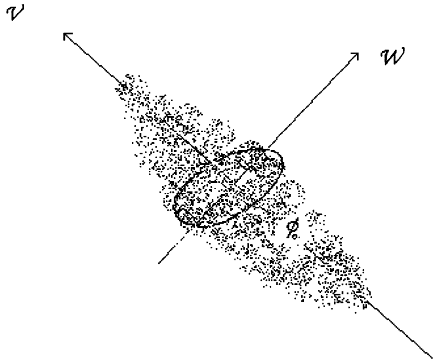

Fig. 1. Simplified schematic diagram of the subspace watermarking strategy. ϕ_0 is the original image feature. The distribution of the forged modifications of the watermarked image is plotted. The components of smaller eigenvalues from applying SVD to the distribution form the watmermark space $\mathcal{W}$, while the components of larger eigenvalues form the subspace $\mathcal{V}$.

our watermark detection method uses a publicly released matrix $D = G^T + BH^T$, whose domain is $\mathcal{W}$. With our published algorithm, a pirate can obtain our watermark space, but cannot obtain the basis of the space that we used to form G and H.

Section II provides a summary of our previous subspace symmetric watermarking method. Section III-A illustrates how we have extended it to our asymmetric method. In Section III-B, we analyze the security of our watermarking under projection attack. The robustness issue is analyzed in Section IV. Implementation and simulations of various attacks are demonstrated in Section V. Finally, in Section VI, we present our conclusions.

II. Symmetric Subspace Watermarking Method

A symmetric waterkmarking system embeds and detects a watermark using the same key, which is hidden from public access. One can refer to [6], [20], [32], [33] for information on many symmetric watermarking methods.

In [31], we proposed a subspace symmetric watermarking method, which models watermarking as a communication with side information [9], for copyright protection. The method makes the keys strongly dependent on the original image and on potential modifications of the watermarked image. The robustness of the approach lies in hiding a watermark in the subspace that is least susceptible to potential modifications. The distribution of the feature of the forged images is analyzed by singular value decomposition (SVD). According to the results of SVD, the optimal solution is to divide the feature space (such as DCT transform and wavelet transform) into two subspaces. A simple presentation of the subspace watermarking approach is given in Fig. 1.

One of the subspaces is the watermark space $\mathcal{W}$, in which the watermark is hidden. $\mathcal{W}$ is chosen because it is the least affected by most modifications of the image X_o. The orthogonal complement of $\mathcal{W}$ is denoted as $\mathcal{V}$, representing the subspace that is most susceptible to modifications of the image. This approach allows a copyright owner to custom select the watermark space that is most resistant to possible attacks. Let ϕ_o be the feature of the original image. Watermark w is embedded into $\mathcal{W}$ by:

$$\phi_w = \phi_o + Gw$$

where G is a secret matrix whose columns are a basis of $\mathcal{W}$, and ϕ_w is the feature of the watermarked image. Because watermark w is in $\mathcal{W}$, the watermark is robust against possible attacks. A pirate can simulate attacks on our watermarked image and obtain a good approximation of space $\mathcal{W}$, but he cannot disclose the secret matrix G from the space.

Our symmetric method uses the key G to embed, and its inverse G^T to extract, watermark w. By choosing G such that $G^T\phi_o = 0$, the method does not need a reference image to detect a watermark. Because the key is content-dependent, when the number of watermarked images is large, there are problems that copyright owners need to manage so that the correct key of a watermarked image can be located. It is also necessary to secretly communicate the keys to another party. In an asymmetric watermarking method, a verifier does not need exclusive permission to access a published key database. This reduces the key management effort. Also, anyone can prove copyright of a watermarked image without secret key communication.

III. The Asymmetric Watermarking Method

Our asymmetric watermarking method is based on the symmetric subspace watermarking method. We divide our feature space into subspace $\mathcal{W}$ and $\mathcal{V}$. The difference from our symmetric method is that we further divide $\mathcal{W}$ into two orthogonal subspaces $\mathcal{G}$ and $\mathcal{H}$. Let G and H denote the secret matrices whose columns form a basis of subspaces $\mathcal{G}$ and $\mathcal{H}$, respectively. We use the matrix G to embed our watermark w into subspace $\mathcal{G}$ and detect w by using the published keys (D, w), where matrix D is a weighted mixing of the matrices H and G.

A. Encoding and Decoding

Embedding our watermark w into subspace $\mathcal{G}$ is achieved by the function

$$\phi_w = \phi_o + Gw. \tag{1}$$

We require the watermark strength $||w||$ to be as large as possible, in order to obtain a high signal-to-noise ratio (SNR) of our watermark signal w to the original image feature ϕ_o. However, $||w||$ should not be so large that the perceptual quality of the watermarked image is destroyed. Finally, a feature reconstruction function is applied to ϕ_w to obtain a watermarked image X_w. The top subfigure of Fig. 2 shows the flowchart of our encoder.

Our detection is a hard decision function δ with a threshold ϵ. The decision function applies the detection matrix D to the extracted feature ϕ_e and then uses the sim function to measure the similarity between $D\phi_e$ and w. Our detector is

$$\delta(\phi_e) = \begin{cases} 1, & \text{if}|\text{sim}(w, D\phi_e)| \geq \epsilon \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where

$$\text{sim}(w, D\phi_e) = \frac{w^T D\phi_e}{||w|| ||D\phi_e||}. \tag{3}$$

The flowchart of our decoder is shown in the bottom subfigure of Fig. 2.

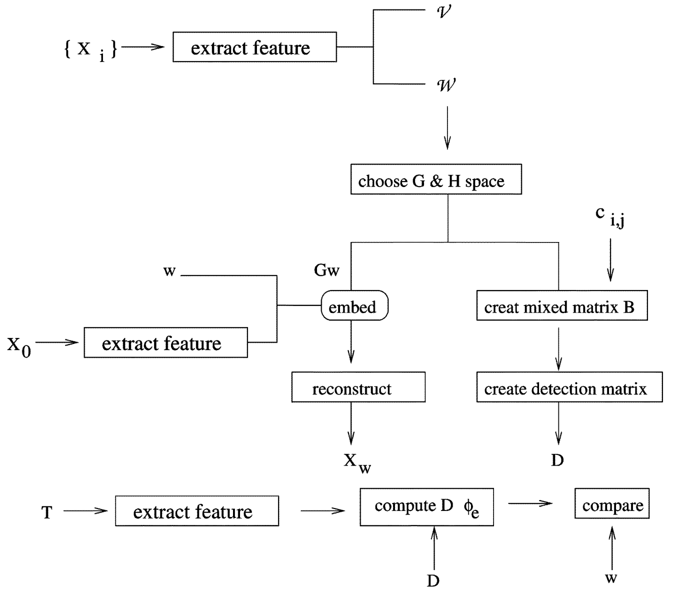

Fig. 2. Top: Our encoder: $\{X_i\}$ is a set of forged images of X_o. Bottom: Our decoder: T is a test image.

We give the matrix D the form

$$D = G^T + BH^T \tag{4}$$

where B is a matrix, H is a matrix, whose columns are a basis of $\mathcal{H}$, and $H^TG = 0$. Let

$$D\phi_o = m_o. \tag{5}$$

For the security reasons presented in part (a) of the theorem in Section III-B, m_o cannot be a zero vector.

Applying w^TD to our watermarked feature ϕ_w, we obtain

$$w^TD\phi_w = w^T(m_o + w). \tag{6}$$

Because m_o is the component from the original image feature, its magnitude $||m_o||$ is usually greater than $||w||$. To avoid the situation where $w^T(m_o + w) \approx w^Tm_o$ (when the detected value is dominated by the original image component), we impose w to be parallel to m_o. This also makes our asymmetric system robust against projection attack, as presented in part (e) of the theorem in Section III-B. Since m_o is dependent on the original image, choosing w to be parallel to m_o is an example of modeling watermarking as a communication with side information.

The embedding key of our asymmetric system is $\{G\}$. Our secret key is $\{G, H\}$, whereas the published detection key is $\{D, w\}$. Two important performance criteria of a watermarking system are security and robustness. Before we analyze the robustness of our asymmetric system, we will assess its security.

Comment: w and D can be hidden from the public and still preserve the integrity of the method by introducing the secret orthogonal matrix Q and releasing (QD, Qw) to the public. Because $Q^{-1} = Q^T$ and $||Q|| = 1$, the sim detected by (QD, Qw) is the same as that obtained by using (D, w).

B. Security Assessment

Watermark security requires that there is an extremely low risk that unauthorized users will be able to successfully hack into the system. Security assessment determines whether Gw can remain secure under malicious attack. A list of published threats to the security of a watermark system is given in [2], [3], [5], [8], [14], [15], and [17]. Among these threats, the oracle attack, which estimates what is secret in the detection process from observations of detector outputs, is not as serious a security threat to asymmetric watermarking as it is to symmetric watermarking. If the oracle attack on an asymmetric watermarking method is successful, it will disclose the detection key, which is already public. An analysis of the oracle attack on the proposed asymmetric method is given in Appendix B. The unsuccessful oracle attack on our asymmetric watermarking method is further evidence of the conclusion in [2], which indicates that the more public information there is in a watermarking scheme, the fewer the potential malicious attacks.

We evaluate the security threats of malicious attacks on our watermarking system. The projection attack is analyzed and presented in this section. The other two attacks, namely jamming a random noise into our watermark space and the copy attack, are evaluated by simulations, and presented in Section V.

A projection attack tries to find the feature $\tilde{\phi}$ that satisfies

$$\min_{\phi} ||\phi - \phi_w||^2$$

with the constraint $w^TD\phi = 0$.[1] This means $\tilde{\phi}$ is the feature without a watermark that is closest to ϕ_w. As a projection attack is extremely effective in removing a watermark, we pay particular attention to this kind of attack.

From $w^TD\phi = 0, \phi$ must be on the hyperplane passing through the origin. The normal vector of the hyperplane is $a = D^Tw$. The solution $\tilde{\phi}$ that minimizes $||\phi - \phi_w||^2$ is the projection of ϕ_w to the hyperplane. Therefore, we have

$$\tilde{\phi} = \phi_w - \frac{a^T\phi_w}{||a||^2}a. \tag{7}$$

Let

$$\phi_o = \psi_o + \sigma_o \tag{8}$$

where ψ_o and σ_o are the components of the original image feature ϕ_o in $\mathcal{W}$ and $\mathcal{V}$ subspaces, respectively. Also let

$$\psi_o = Gs + Ht \tag{9}$$

where $s \neq 0$ and $t \neq 0$ are the coefficient vectors of ψ_o in G and H, respectively. According to (1), the component of the watermarking image feature ϕ_w in $\mathcal{W}$ is

$$\psi_w = \psi_o + Gw. \tag{10}$$

The following theorem shows that we can construct a matrix D so that the projection attack yields σ_o. Because ψ_o is the perceptually robust feature of the original image, the image reconstructed from $\sigma_o \in \mathcal{V}$ has a high probability of being a perceptually distorted image.

[1] The real constraint is $|\text{sim}(W, D\phi)| < \epsilon$, where ϵ is the threshold. We use a simplified constraint so that the attack can be analyzed.

Theorem: Given G, H, and ϕ_o.

a) To be robust to the projection attack, the detection matrix D [defined in (4)] must be chosen such that $D\phi_o \neq 0$.

b) If D is chosen such that

$$D^T w = \lambda \psi_w \tag{11}$$

where $\lambda \neq 0$ and ψ_w is defined by (10), then applying the projection attack to ϕ_w [defined in (1)] obtains σ_o.

c) If D and w satisfy (11), then

$$w = \frac{\lambda}{1-\lambda} s \tag{12}$$

where $\lambda \neq 1$ and s as defined in (9).

d) If B is constructed as

$$B = \frac{(1-\lambda)}{\|s\|^2} st^T + \sum_{i,j} c_{i,j} u_i v_j^t \tag{13}$$

where $u_i \perp s, v_j \perp t$ (defined in (9)), $c_{i,j}$ is a real number, and w satisfies (12), then D satisfies (11).

e) If w and D are chosen according to (12) and (13), then w is parallel to m_o [defined in (5)].

f) If B satisfies (13), and $\lambda = 1 + (\|s\|^2)/(\|t\|^2)$, then $D\phi_o = 0$.

Proof: See Appendix A.

Corollary: In order to be robust against projection attack, the watermark w must be parallel to m_o, and λ cannot be equal to either $0, 1$, or $1 + (\|s\|^2)/(\|t\|^2)$.

Fig. 3 shows that the projection attack on our watermarked image makes the image perceptually unacceptable. Thus, our watermarking is secure under projection attack. Fig. 4 shows a diagram of the oracle attack on our watermarked image. Having made a security accessment of our watermarking system, we now analyze its robustness.

IV. Robustness Analysis

The performance criteria of a robust watermarking system are high detection probability and low false positive probability. The latter measures the probability that a watermark is detected in a test image, even though the test image is not watermarked. An analysis of false positive probability by using a sim detector is given in [19].

Detection probability measures the probability that a detector correctly determines that a manipulated watermarked image contains a watermark. A feature extracted from this watermarked image is composed of ϕ_w and a noise n_w. That is

$$\phi_e = \phi_w + n_w. \tag{14}$$

Applying D to ϕ_e and using (1), (4), and (14), we obtain

$$D\phi_e = (m_o + w) + Dn_w.$$

Because we choose w to be parallel to m_0, we have

$$\begin{aligned}\operatorname{sim}(w, D\phi_e) &= \frac{w^T(m_o + w + Dn_w)}{\|w\|\|m_o + w + Dn_w\|}\\ &= \frac{(m_o + w)^T(m_o + w + Dn_w)}{\|m_o + w\|\|m_o + w + Dn_w\|}\\ &\geq \cos\sin^{-1}\frac{\|Dn_w\|}{\|m_o + w\|}.\end{aligned} \tag{15}$$

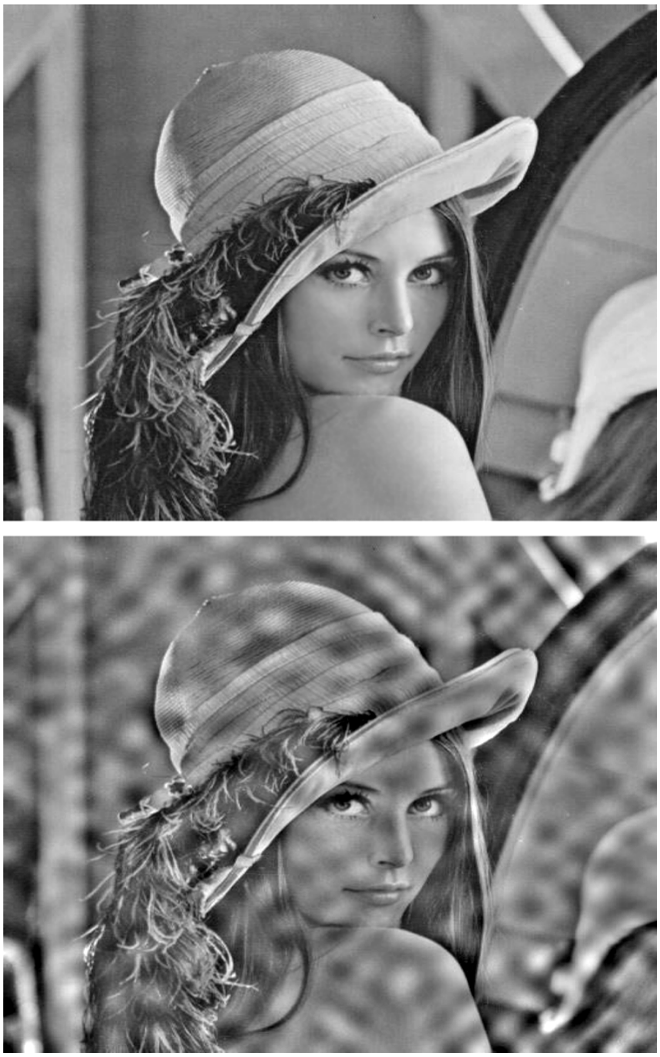

Fig. 3. Top: Our watermarked image. Bottom: The image obtained from the feature extracted by applying a projection attack to the watermarked image. The PSNR of the noise image, obtained by subtracting the bottom image from the top image, is 17 dB.

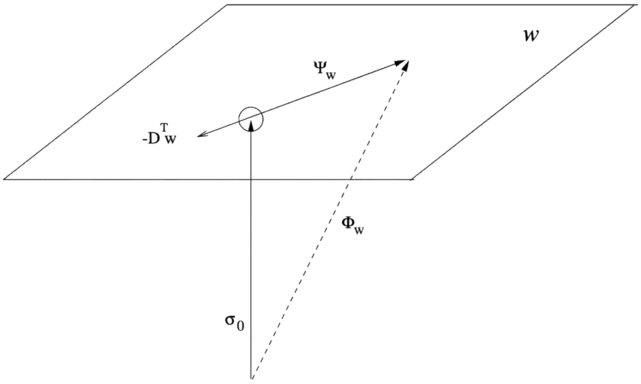

Fig. 4. Feature obtained by applying the oracle attack is circled, where $D^T w$ is the successfully estimated vector of the attack. This is the same feature as that obtained by the projection attack.

The proof of the last inequality is given in Fig. 5. Thus, in order to have a high detection probability, we need to design a matrix D so that $(\|Dn_w\|)/(\|m_o + w\|)$ is as small as possible.

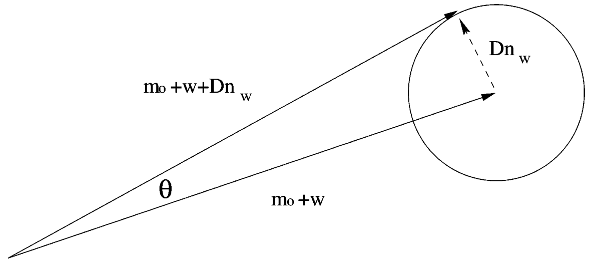

Fig. 5. Plane contains $m_o + w$ and Dn_w. The largest angle θ between the two vectors occurs when $m_o + w + Dn_w$ is perpendicular to Dn_w. As $\sin\theta = (\|Dn_w\|)/(\|m_o + w\|)$, so $\theta = \sin^{-1}(\|Dn_w\|)/(\|m_o + w\|)$. Therefore, the last inequality of (15) is satisfied.

1) ROC Curves: In the theorem, the matrix B in the detection matrix $D = G^T + BH^T$ has the general form

$$B = \frac{(1-\lambda)}{\|s\|^2} st^T + \sum_{i,j} c_{i,j} u_i v_j^t \tag{16}$$

where $u_i \perp w, v_j \perp t$, and $c_{i,j}$ is a real number. Here, we present the simulation results of the detection probability and false positive probability as functions of a threshold with different values of $c_{i,j}$ and demonstrate our method's performance by using receiver operating characteristic curves, or ROC curves [24].

We let $c_{i,j}$ be a uniformly distributed random variable with range $[-c, c]$, where c is a parameter. To measure the detection probability for each c, we performed 1000 attacks on a watermarked Lena image. These attacks included shifting (at most ten pixels in either a horizontal or vertical direction), scaling, blurring, JPEG compression, adding white noise, sharpening, rotation (at most $\pm 5^\circ$), stirmark,[2] as well as combinations of the attacks. The top figure of Fig. 6 shows the detection probability, versus a threshold of $|\mathrm{sim}|$, of the attacked watermarked images.

To measure the false positive probability, we watermarked the Lena image with 16 different watermarks. For each watermarked Lena and for each c, we obtained a pair of public detection keys (D_i, w_i). We used the keys to 1000 unwatermarked images to measure the $|\mathrm{sim}|$ values. The curves in the bottom figure of Fig. 6 are the false positive probabilities, versus a threshold of $|\mathrm{sim}|$, with different c values.

Fig. 6 shows that both the detection probability and false positive probability are smaller at a given threshold for a larger c value. If we apply D to ϕ_e (the feature extracted from a test image), then from (16), we have

$$\begin{aligned} D\phi_e &= G^T\phi_e + BH^T\phi_e \\ &= \frac{(1-\lambda)^2}{\lambda\|s\|^2}(t^T H^T \phi_e)w + G^T\phi_e + \sum_{i,j} c_{i,j} u_i v_J^T H^T \phi_e \\ &= \tilde{k}w + G^T\phi_e + \sum_i \tilde{c}_i u_i. \end{aligned}$$

Because $\sum_i \tilde{c}_i u_i$ is normal to w, this term does not contribute to the value in the norminator of $\mathrm{sim}(w, D\phi_e)$, but it does contribute to the value in the denominator. Hence, as c increases, the detected values, of both a watermarked or an un-watermarked image, decrease.

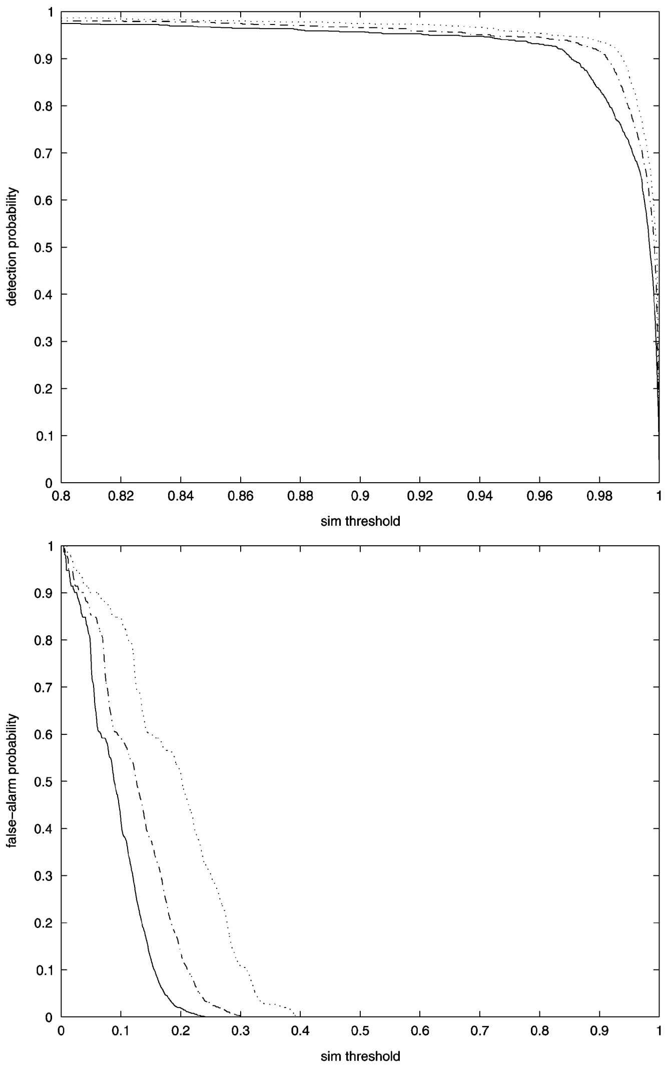

Fig. 6. Probabilities of detection and false-positives when $|\mathrm{sim}|$ is used as our threshold with different c values. Top: Detection probability. Bottom: Mean false positive probability. The horizontal axis is the $|\mathrm{sim}|$ threshold. Solid curves: $c = 0.15$. Dashed-dotted curves: $c = 0.1$. Dotted curves: $c = 0.05$.

The plots of the empirical data of the detection and false positive probabilities, shown in Fig. 6, do not overlap. This implies that many $|\mathrm{sim}|$ values can be used as our threshold because any of them will yield a high detection probability and a low false positive probability.

Although there is no overlap between the detection probability and the false positive probability, according to the law of large numbers, if we had enough images, both the detection probability and the false positive probability would have become a Gaussian distribution. As proposed in [7], we model the detection probability and false positive probability as Gaussian distributions. We compute the mean and the standard deviation of the Gaussian distribution of the false positive probability from the detected values of the unwatermarked images, in the same way, we compute the parameters of the detection probability for the watermarked images. From the Gaussian distributions, we can draw the ROC curve of our empirical data. Fig. 7 shows the ROC curves of different c values, obtained

[2] [Online] Available at: http://www/petitcolas.net/fabien/watermarking/stirmark

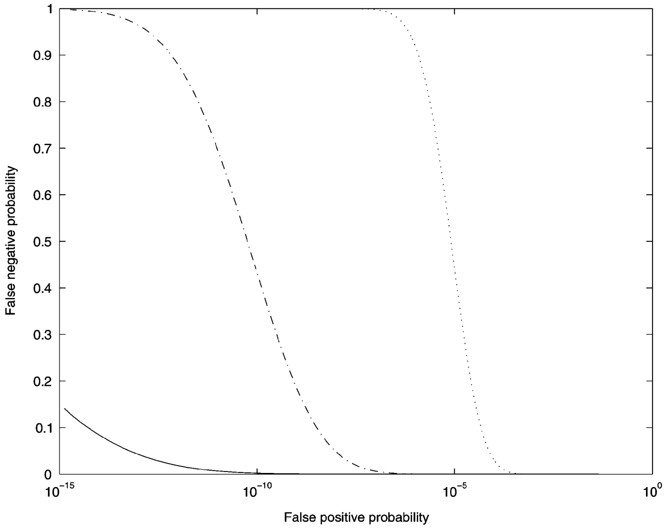

Fig. 7. ROC curves of our empirical data. The curves plot the false positive probability in the logarithmic (base 10) scale against the false negative probability, which is defined as one minus the detection probability, as a function of the threshold and c value. Solid curve: $c = 0.15$. Dashed-dotted curve: $c = 0.1$. Dotted curve: $c = 0.05$.

in this manner. For numerical precision, the figure shows only the part of the curves whose false positive probability is above 10^{-15}. The intersections of the curves and the axis of false positive probability (x-axis) are 0.34, 0.50, and 0.56 for curves of $c = 0.15, c = 0.1$, and $c = 0.05$, respectively. In the following simulations, we choose $c = 0.1$ and use 0.5 as our threshold. This corresponds to the false positive probability below 10^{-5} in our simulation.

V. Implementation and Simulation Results

A. Implementation

The major implementation issues were to find the watermark space $\mathcal{W}$ and the matrix D. Our image database had a total of 61 images of womens' faces. One of them was a Lena image, the other 60 were down-loaded from the Google Search Engine. Their sizes ranged from 223 by 342 pixels to 512 by 512 pixels. We applied the full frame DCT to each image, and then selected DCT coefficients from their upper left 32 by 32 corners to form our feature. The coefficients corresponded to 32 horizontal low-frequency bands and 32 vertical low-frequency bands. Thus, our feature space had a dimension of 1024 frequency bands.

From each image in our database, we obtained a set of 100 forged images by means of image operations. Our operations included: blurring (with B-spline kernel), JPEG compression, scaling, rotations (with $-5^\circ \leq \theta \leq 5^\circ$), translations (by shifting at most ten pixels either up, down, left or right), adding random noise, stirmark, various image operations from the Matlab image toolbox, as well as combinations of all these operations. For each image, we computed the covariance matrix from the collection of features obtained from the forged images of an image. Using SVD on the covariance matrix, we chose our watermark space $\mathcal{W}$ to be the space spanned by 900 eigenvectors (of the covariance matrix) whose corresponding eigenvalues were small.

Let W be the matrix of the 900 eigenvectors and ϕ_o be the feature of the original image. Using a notebook computer with a CPU 1.8 GHz, it takes about 5 min to generate a watermark space and its orthogonal basis. From an orthogonal basis of $\mathcal{W}$ space, we randomly choose 300 vectors as an orthogonal basis of $\mathcal{G}$, and form the columns of the secret matrix G. The remaining vectors form the columns of the secret matrix H.

We project ϕ_o to $\mathcal{W}$ and obtain ψ_o, which is then represented as $\psi_o = Gs + Ht$, where s and t are the coefficient of ψ_o in $\mathcal{G}$ and $\mathcal{H}$, respectively. We choose our watermark w to be parallel to s. Let $w = ks$, where k is a scalar. In order to have a high SNR of our watermark signal w to ϕ_o, k should be as large as possible. However, it cannot be too large, or it will decrease the perceptual quality of our watermarked image. The average PSNR of our watermarked image is 44 dB. After w is chosen, we have $\lambda = (1/1 + k)$. We then find $\{u_i\}_{i=1}^{299}$ with each $u_i \perp w$, and $\{v_j\}_{j=1}^{599}$ with each $v_j \perp t$. We randomly choose $c_{i,j} \in [-1, 1]$ and use them to construct a matrix B to satisfy (16). From B, we obtain the detection matrix $D = G^T + BH^T$.

B. Simulations

We now demonstrate the resistance of our asymmetric watermarking method to the following attacks. See [22] and [23] for further discussion of various attacks.

1) Applying a Copy Attack: A copy attack uses noise reduction methods to extract an approximation of a watermark from a watermarked image X_w, and hides the feature in another image, Y [17]. This increases the false positive probability.

Let (D_x, w_x) be the public keys of the watermarked image X_w, and ϕ_y be a feature of image Y. Let $G_x w_x + \tilde{n}$ be the estimated mark from X_w. We replace ϕ_y with $\phi_y + k(G_x w_x + \tilde{n})$ as Y's feature. A copy attack is successful if one can use w_x and D_x to obtain a high sim from $\phi_y + kG_x w_x + k\tilde{n}$

$$\mathrm{sim}(w_x, D_x(\phi_y + kG_x w_x + k\tilde{n})) = \frac{w_x^T D_x(\phi_y + k\tilde{n}) + k\|w_x\|^2}{\|w_x\|\|D_x(\phi_y + kG_x w_x + k\tilde{n})\|}.$$

We performed the following simulations of copy attacks. We obtained a noise image N by subtracting the watermarked Lena image from a denoised image, using a mask which is a tensor product of the B-spline filter with coefficients $[(1/16)(4/16)(6/16)(4/16)(1/16)]$. We then extracted an estimated feature of $G_x w$ from the image N. The estimated feature was multiplied by k, and embedded into each non-Lena image in our database. After applying Lena's detection matrix D_x to each image, we compared the results with w_x. We found that the mean and the standard deviation of the 60 copy attacks were 0.0346 and 0.0287, respectively. Comparing these results with our threshold 0.5, the probability of successfully copy attacking our method is very low.

2) Spreading Noise Into Watermark Space: The simulation results shown in [31] indicate that a pirate can simulate attacks on our watermarked image and obtain a good approximation of space $\mathcal{W}$, but he cannot obtain the secret matrix G from the space. In this scenario, we evaluate the efficiency of a pirate who attacks our watermark space $\mathcal{W}$ by jamming it with

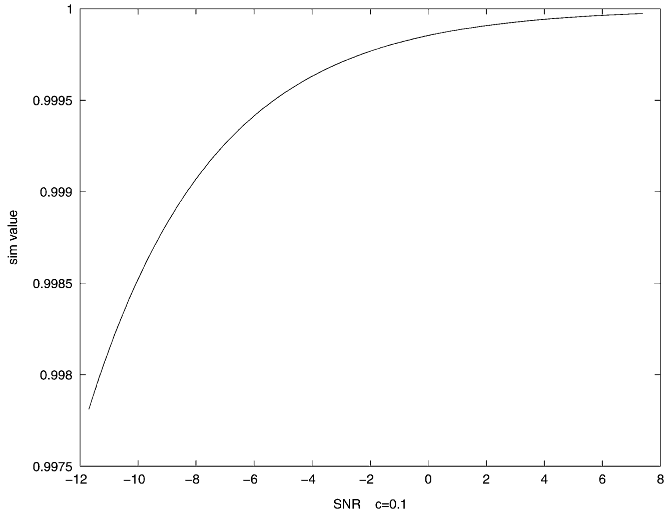

Fig. 8. White noise spreading attack on $\mathcal{W}$. The watermark space of the watermarked Lena image was attacked by random noises that had various energy levels. The SNR was measured by $20\log_{10}(\|w\|/\|n\|)$.

number	method	sim mean	sim std
1	shift	0.978	0.244
2	blur	0.999	0.001
3	JPEG	0.998	0.002
4	sharpen	0.999	0.002
5	rotation (±5°)	0.957	0.085
6	stirmark	0.984	0.030
7	combinations	0.934	0.092

Fig. 9. Blind attacks: The mean and standard deviation of various attacks on our watermarked images. Method Key: (1) Shift: a random shift of at most 10 pixels. (2) Blur: a smoothing of images by a B-spline of order chosen from 3, 5, 7, and 9. (3) Spreading noise: the adding of white noise to images. (4) JPEG: the compression of images using JPEG with a quality factor between 30 and 100. (5) Sharpen: the sharpening of images using Microsoft Photo Editor (parameter 5). (6) Rotation: the rotation of images by at most $\pm 5^\circ$. (7) Stirmark: a random seed from 1 to 100. (8) Combination: a combination of the rotation [θ in (6)], shift [in (1)], and a rotation back $(-\theta)$ operations.

random noise. We embed 64 random noises that have various levels of energy into the watermark space of a watermarked Lena image. Performance results for this attack are shown in Fig. 8. We plot the mean, obtained by averaging the detection values of the 64 random noise attacks on the $\mathcal{W}$ space, versus the SNR that is measured by $20\log_{10}(\|w\|/\|n\|)$, where n is our random noise. One can observe from the figure that even at a very low SNR, the detection value is still quite high compared to our threshold. This proves that our method is robust against this type of attack.

3) Blind Attacks: Blind attacks are carried out with the intention of removing a watermark when the attacker does not know the watermarking method. For each of the 61 images in our database, we produced 32 watermarked images and performed an average of 100 attacks on each image. These attacks included: shifting, blurring, JPEG compression, sharpening, rotation, stirmarking, as well as combinations of the above attacks. Fig. 9 shows the mean and the standard deviation of the $|\mathrm{sim}|$ values. Although the rotation attack (i.e., rotating our watermarked image without rotating it back) has a slightly lower detection value, our method is still robust against these attacks.

VI. Conclusion

To resolve the weaknesses of current symmetric watermarking methods, we have designed an asymmetrical watermarking method for copyright protection that satisfies the zero knowledge principle. All of our watermarking, except the secret matrices G and H, have been released and are publicly available. Our asymmetric design is robust because it enhances the watermark space concept of our previous symmetric watermarking method. As our watermark is highly dependent on the original image, it cannot be removed without the watermarked image being perceptually distorted. Our method is secure, since we embed secret information Gw within a subspace of $\mathcal{W}$, and provide the public with a key $(D = G^T + BH^T)$ to detect Gw. Because the secret basis of $\mathcal{G}$ is hidden from the public, estimating Gw is extremely difficult. Further development of our method will reduce the size of our detection matrix, and extend our method to copy protection.

Appendix A

Here, we provide the proof of our theorem.

Proof of a): If D is chosen such that $D\phi_o = 0$, then

$$\|\tilde{\phi} - \phi_w\| \leq \|w\|.$$

Because $D\phi_o = 0, \phi_o$ meets the constraint $w^T D\phi_o = 0$, and since $\tilde{\phi}$ is the solution that minimizes $\|\phi - \phi_w\|^2$, we have $\|\tilde{\phi} - \phi_w\| \leq \|\phi_o - \phi_w\| = \|Gw\| = \|w\|$.

As this proof indicates that for $D\phi_o = 0$, the image reconstructed from $\tilde{\phi}$ (i.e., the image that does not have a watermark) is probably perceptually undistorted. ■

Proof of b): Let $a = D^T w = \lambda\psi_w$, where $\lambda \neq 0$. Applying the projection attack to a watermarked image obtains $\tilde{\phi}$, which is (from (7))

$$\begin{aligned}\tilde{\phi} &= \phi_w - \frac{a^T\phi_w}{\|a\|^2}a \\ &= \phi_w - \frac{\lambda\|\psi_w\|^2}{\lambda^2\|\psi_w\|^2}\lambda\psi_w \\ &= \phi_w - \psi_w \\ &= \sigma_o\end{aligned}$$

where $\tilde{\phi}$ satisfies $w^T D\tilde{\phi} = 0$ and the detected value is zero. The image reconstructed using $\tilde{\phi}$ is probably a distorted image. ■

Proof of c): If $D^T w = \lambda\psi_w$, then according to (4) and (10), we obtain

$$(G + HB^T)w = \lambda G(s + w) + \lambda Ht.$$

Because the columns of G and H are orthonormal and $G^T H = 0$, we have

$$w = \lambda(s + w) \tag{17}$$

and

$$B^T w = \lambda t. \tag{18}$$

From (17), when $\lambda \neq 1$, we obtain

$$w = \frac{\lambda}{1-\lambda}s. \tag{19}$$

■

Proof of d): We want to construct a matrix B such that $D^T w = \lambda \psi_w$ is satisfied.

Substituting (19) into (18), we have

$$B^T s = (1-\lambda) t.$$

The general form of B that solves the above equation is

$$B = \frac{1-\lambda}{\|s\|^2} s t^T + \sum_{i,j} c_{i,j} u_i v_j^T \tag{20}$$

where $c_{i,j}$ are real numbers and $u_i \perp s, v_j \perp t$, for every i and j.

Substituting B in (20) into $D = G^T + BH^T$ and using the fact that $w = (\lambda/1-\lambda)s$, we obtain

$$\begin{aligned} D^T w &= (G + HB^T) w \\ &= Gw + H \left(\frac{(1-\lambda)}{\|s\|^2} t s^T + \sum_{i,j} c_{i,j} v_j u_i^T \right) w \\ &= Gw + \lambda H t \\ &= \frac{\lambda}{1-\lambda} Gs + \lambda H t \\ &= \lambda \left(G \left(\frac{\lambda}{1-\lambda} s + s \right) + Ht \right) \\ &= \lambda (G(s+w) + Ht) = \lambda \psi_w. \end{aligned} \tag{21}$$

■

Proof of e): In a), we show that for our method to be robust against the projection attack, $D\phi_o$ cannot be zero. In the following, we show that $D\phi_o$ becomes zero for a particular λ.

$$\begin{aligned} D\phi_o &= (G^T + BH^T)(Gs + Ht) \\ &= s + Bt \\ &= s + \frac{(1-\lambda)}{\|s\|^2} s \|t\|^2 \\ &= \left(\frac{\|t\|^2}{\|s\|^2} (1-\lambda) + 1 \right) s. \end{aligned} \tag{22}$$

If $D\phi_o = 0$, then we have

$$\frac{\|t\|^2}{\|s\|^2}(1-\lambda) + 1 = 0.$$

That is

$$\lambda = 1 + \frac{\|s\|^2}{\|t\|^2}. \tag{23}$$

■

Proof of f): From (22) and (12), we have $D\phi_o$ parallel to w.

Appendix B

Here, we discuss the oracle attack on our asymmetric watermarking method. The attack estimates what is secret in the detection process and uses what is estimated to remove the embedded watermark. This attack can successfully estimate the secret key of a subspace-based symmetric watermarking system [31], where the detection matrix is the secret matrix $D = G^T$. Although the oracle attack estimates the detection key $D^T w = Gw + HB^T w$ of our asymmetric watermarking method, it cannot obtain the secret information Gw. In an asymmetric watermarking system, the detection key is public, thus a successful oracle attack obtains publicly available information that is of no use for secret key estimation.

The next phase of the attack is to add the vector $-\eta D^T w$ to the watermark feature so that

$$w^T D(\phi_w - \eta D^T w) = 0. \tag{24}$$

According to our Theorem, we have $D^T w = \lambda \psi_w$. If we substitute this into (24), we have

$$\psi_w^T (\phi_w - \eta \psi_w) = 0. \tag{25}$$

From (1), (8), and (10), the solution of (25) is $\eta = 1$. Thus, the oracle attack and projection attack result in the same feature σ_o. As shown previously, Fig. 4 shows a diagram of the oracle attack on our watermarked image.

Acknowledgment

The authors would like to express their gratitude to Reviewers 1 and 2 for their insightful suggestions. They give special thanks to Reviewer 1, whose invaluable suggestions—especially about the projection attack—significantly improved their manuscript.

References

[1] A. Adelsbach, S. Katzenbeisser, and A. R. Sadeghi, "Watermark detection with zero-knowledge disclosure," in *ACM Multimedia Syst.*, 2003, vol. 9.

[2] M. Barni, F. Bartolini, and T. Furon, "A general framework for robust watermarking security," *Signal Processing*, vol. 83, pp. 2069–2084, 2003.

[3] J. A. Bloom, I. J. Cox, T. Kalker, J. P. Linnartz, M. L. Miller, and C. B. S. Traw, "Copy protection for DVD video," *Proc. IEEE*, vol. 87, no. 7, pp. 1267–1276, Jul. 1999.

[4] S. Craver, "Zero knowledge watermark detection," in *Proc. 3rd Int. Workshop on Information Hiding*, vol. 1768, Spring 2000, pp. 101–116. Lecture Notes in Computer Science.

[5] S. Craver, N. Memon, B. L. Yeo, and M. M. Yeung, "Resolving rightful ownership with invisible watermarking teachniques: Limitations, attacks, and implications," *IEEE J. Select. Areas Commun.*, vol. 16, no. 9, pp. 573–587, Dec. 1998.

[6] I. Cox, J. Kilian, T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for multimedia," *IEEE Trans. Image Processing*, vol. 6, no. 12, pp. 1673–1687, Dec. 1997.

[7] I. Cox, M. Miller, and J. Bloom, *Digital Watermarking*. New York: Morgan Kaufmann, 2002, pp. 173–177.

[8] I. Cox and J. P. Linnartz, "Some general methods for tampering with watermarks," *IEEE J. Select. Areas Commu.*, vol. 16, no. 4, pp. 587–593, May 1998.

[9] I. Cox, M. Miller, and A. Mckellips, "Watermarking as communication with side information," *Proc. IEEE*, vol. 87, pp. 1127–1141, Jul. 1999.

[10] J. Eggers, J. Su, and B. Girod, "Public key watermarking by eigenvectors of linear transforms," in *Proc. Eur. Signal Process. Conf.*, Tampere, Finland, Sep. 2000.

[11] T. Furon, I. Venturini, and P. Duhamel, "An unified approach of asymmetric watermarking schemes," in *Proc, SPIE: Security and Watermarking of Multimedia Contents III*, P. W. Wong and E. Delp, Eds., San Jose, CA, Jan. 2000.

[12] T. Furon and P. Duhamel, "An asymmetric watermarking method," *IEEE Trans. Signal Processing*, vol. 51, no. 4, pp. 981–995, Apr. 2003.

[13] F. Hartung and B. Girod, "Fast public-key watermarking of compressed video," in *Proc. IEEE Int. Conf. Image Processing*, Santa Barbara, CA, Oct. 1997, pp. 528–531.

[14] T. Kalker, "Watermark estimation through detector observations," in *Proc. IEEE Benelux Signal Processing Symp.*, Leuven, Belgium, Mar. 1998.

[15] ——, "A security risk for publicly available watermark detectors," in *Proc. Benelux Information Theory Symp.*, Veldhoven, The Netherlands, May 1998.

[16] A. Kerckhoffs, "La cryptographie militaire," *J. Sci. Militaires*, vol. 9, pp. 5–38, Jan. 1883.

[17] M. Kutter, S. Voloshynovskiy, and A. Herrigel, "Watermark copy attack," in *SPIE Security and Watermarking of Multimedia Contents III*, P. W. Wong and E. Delp, Eds., San Jose, CA, Jan. 2000.

[18] J. C. A. Van Der Lubbe, *Basic Methods of Cryptography* Cambridge, MA, 1998.

[19] M. L. Miller and J. A. Bloom, "Computing the probability of false watermark detection," in *Proc. Workshop on Information Hiding*, Dresden, Germany, Sep. 1999.

[20] P. Moulin and E. Delp, "A mathematical appraoch to watermarking and data hiding," in *IEEE ICIP*. Thessaloniki, Greece, 2001.

[21] P. Moulin and A. Ivanovic, "The zero-rate spread-spectrum watermarking game," *IEEE Trans. Signal Processing*, vol. 51, no. 4, pp. 1098–1117, Apr. 2003.

[22] F. A. P. Petitcolas, R. J. Anderson, and M. G. Kuhn, "Attacks on copyright marking systems," *Lecture Notes on Computer Science 1525: Information Hiding*, pp. 218–238, 1998.

[23] ——, "Information hiding—A survey," in *Proc. IEEE*, 1999, pp. 1062–1078.

[24] H. V. Poor, *An Introduction to Signal Detection and Estimation*. New York: Springer, 1994.

[25] J. G. Proakis, *Digital Communications*. New York: McGraw-Hill, 1995.

[26] R. Van Schyndel, A. Tirkel, and I. Svalbe, "Key independent watermark detection," in *Proc. Int. Conf. Multimedia Comput. Syst.*, vol. 1, Florence, Italy, Jun. 1999.

[27] J. Smith and C. Dodge, "Developments in steganography," in *Proc. 3rd Int. Workshop on Information Hidding*, A. Pfitzmann, Ed., Dresden, Germany, Sep. 1999, pp. 77–87.

[28] J. Stern and J. P. Tillich, "Automatic detection of a watermarked document using a private key," in *Proc. 4th Int. Workshop on Information Hiding*, vol. 2137, I. S. Moskowitz, Ed., Pittsburgh, PA, Apr. 2001.

[29] M. Swanson, B. Zhu, A. Tewfik, and L. Boney, "Robust audio watermarking using perceptual masking," *Signal Processing*, vol. 66, no. 3, pp. 337–355, May 1998.

[30] W. Trapper, M. Wu, Z. Wang, and K. J. R. Liu, "Anti-Collusion fingerprintng for multimedia," *IEEE Trans. Signal Processing*, vol. 51, no. 4, pp. 1069–1087, Apr. 2003.

[31] J. Tzeng, W. L. Hwang, and I. Liang Chern, "Enhancing image watermarking methods with/without reference images by optimization on second order statistics," *IEEE Trans. Image Processing*, vol. 11, pp. 771–782, Jul. 2002.

[32] S. Voloshynovskiy, A. Herrigel, N. Baumgaertner, and T. Pun, "A stochastic approach to content adaptive digital image watermarking," in *Proc. 3rd Int. Workshop on Information Hiding*, Dresden, Germany, Sep. 1999, pp. 211–236.

[33] R. B. Wolfgang, C. I. Podilchuk, and E. J. Delp, "Perceptual watermarks for digital images and video," in *Proc. IEEE*, vol. 87, Jul. 1999, pp. 1108–1126.

[34] M. M. Yeung, "Digital watermarking," *Commun. ACM*, vol. ACM 41, no. 7, pp. 31–33, Jul. 1998.

Jengnan Tzeng received the B.S. degree from National Chengchi University, Taiwan, R.O.C., the M.S. degree from National Central University, Taiwan, and the Ph.D. degree from National Taiwan University in 2004, all in mathematics.

He is currently a Postdoctoral Researcher in the Genomics Research Center, Academia Sinica, Taipei, Taiwan. His research interests include digital watermarking, wavelet analysis, and PDE methods in image processing and genomic research.

Wen-Liang Hwang received the B.S. degree in nuclear engineering from National Tsing Hua University, Hsinchu, Taiwan, R.O.C., the M.S. degree in electrical engineering from the Polytechnic Institute of New York, and the Ph.D. degree in computer science from New York University in 1993.

He was a Postdoctoral Researcher with the Department of Mathematics, University of California, Irvine, in 1994. In January 1995, he became a member of the Institute of Information Science, Academia Sinica, Taipei, Taiwan, where he is currently an Associate Research Fellow. He is co-author of the book Practical Time-Frequency Analysis (New York: Academic, 1998). His research interests include wavelet analysis, signal, and image processing, multimedia transmission, and computer vision.

Dr. Hwang was awarded the Academia Sinica Research Award for Junior Research in 2001.

I-Liang Chern received the B.S. and M.S. degrees in mathematics from National Taiwan University, Taiwan, R.O.C., and the Ph.D. degree in mathematics from New York University in 1983.

He was with Academia Sinica, Taipei, Taiwan, the Courant Institute, New York University, and Argonne National Laboratory, Argonne, IL. He has been a Professor in the Mathematics Department of National Taiwan University since 1991. His research interests include multiscale scientific computing, wavelet analysis, and partial differential equations.