Cocktail Watermarking for Digital Image Protection

Chun-Shien Lu, *Member, IEEE*, Shih-Kun Huang, Chwen-Jye Sze, and Hong-Yuan Mark Liao, *Member, IEEE*

***Abstract*—A novel image protection scheme called "cocktail watermarking" is proposed in this paper. We analyze and point out the inadequacy of the modulation techniques commonly used in ordinary spread spectrum watermarking methods and the visual model-based ones. To resolve the inadequacy, two watermarks which play complementary roles are simultaneously embedded into a host image. We also conduct a statistical analysis to derive the lower bound of the worst likelihood that the better watermark (out of the two) can be extracted. With this "high" lower bound, it is ensured that a "better" extracted watermark is always obtained. From extensive experiments, results indicate that our cocktail watermarking scheme is remarkably effective in resisting various attacks, including combined ones.**

***Index Terms*—Attacks, modulation, protection, robustness, watermarking.**

I. Introduction

TRANSFERRING digitized media via the Internet has become very popular in recent years. However, this frequent use of the Internet has created a need for security. As a consequence, to prevent information which belongs to rightful owners from being intentionally or unwittingly used by others, information protection is indispensable. A commonly suggested method is to insert watermarks into original information so that rightful ownership can be declared. This is the so-called watermarking technique. An effective watermarking procedure usually requires satisfaction of a set of typical requirements. These requirements include transparency, robustness, maximum capacity, universality, oblivious detection, solution of ownership deadlock and so on.

In the following paragraph, we will briefly review some existing watermarking methods. Other surveys regarding watermarking can also be found in [5], [9], [13], [14], [30], [35], [40]. In the literature, Koch and Zhao [16] transformed an image by using block-discrete cosine transform (block-DCT) and then utilized a pseudorandom number generator to select a subset of blocks. A triplet of blocks with midrange frequencies was slightly revised to yield a binary sequence watermark. This seems reasonable because low frequency components are perceptually important but easy to sense after modification, and high frequency components are easy to tamper with. Kundur and Hatzinakos [17] proposed to encode a watermark by a quantization operation. However, the watermark extracted by quantization is very sensitive to attacks. Cox *et al.*[5] proposed a global DCT-based spread spectrum approach to hide watermarks. They believed that the signal energy present in any frequency is undetectable if a narrowband signal is transmitted over a much broader bandwidth. Ideally, this will cause a watermark to spread over all frequencies so that the energy in any single frequency is very small and, thus, undetectable. Their watermark is of fixed length and is produced from a Gaussian distribution with zero mean and unit variance. They distribute as fairly as possible the watermark to the first 1000 largest ac coefficients. An objective measurement was proposed to evaluate the similarity between the original and the extracted watermarks. Hsu and Wu [15] used multiresolution representations for the host image and the binary watermark. The middle frequencies in the transformed wavelet domain were selected for modification using a residual mask. Their method has been shown to be effective for large images and for JPEG-based compression at higher bit rates. Bender *et al.* [3] also altered the intensities of a host image within a small range and hoped the updates were perceptually unnoticed. However, there are limitations in the above mentioned methods: 1) it is unclear where the watermark can be hidden and to what extent modification can be made to find the compromise between the transparency and the robustness requirements; (2) owing to inadequate robustness, these approaches are not suitable for practical use.

In order to improve the first drawback, the characteristics of the human visual system (HVS) have been incorporated into the watermark encoder design [8], [31], [35]. It is very meaningful and reasonable to take HVS into account because of its inherent features. If one can modify an image based on rules taken from the human visual system, then it will be easier to generate an imperceptible watermark with maximum modifications, and the length and strength of a watermark can be adaptive to the host image. Basically, a watermarking scheme that does not sufficiently utilize the capacity of a host image may cause the potential length and strength of a watermark to be bounded.

The second drawback mentioned above is, in fact, a major problem associated with current watermarking techniques. Generally speaking, current watermarking approaches are not strongly robust to attacks or combinations of several attacks, so that their use is limited [13]. In this paper, this problem will be seriously addressed. We shall begin by introducing two famous works [5], [31], which are frequently cited. The first one is the spread spectrum watermarking technique proposed by Cox *et al.*Their method has become very popular and has been employed by many researchers [14], [30]. The other one, proposed by Podilchuk and Zeng [31], is a human visual model-based watermarking scheme. However, the reasons why the two aforementioned methods are successful or not are still unclear. We shall investigate the modulation techniques used in [5], [31] and clearly point out their drawbacks. We assert

Manuscript received October 7, 1999; revised October 10, 2000. The associate editor coordinating the review of this paper and approving it for publication was Dr. M. Reha Civanlar.

The authors are with the Institute of Information Science, Academia Sinica, Taipei 11529, Taiwan, R.O.C (e-mail: lcs@iis.sinica.edu.tw; liao@iis.sinica.edu.tw).

Publisher Item Identifier S 1520-9210(00)11055-7.

1520–9210/00$10.00 © 2000 IEEE

that in order to obtain high detector responses, most of the transformed coefficients of the host image and the watermarked image have to be modulated along the same *direction*. This is the key concept needed to improve the previous approaches because a watermark detector can produce a high correlation value only when the above mentioned condition is satisfied. We have observed that an arbitrary attack usually tends to increase or decrease the magnitudes of the majority $(\geq 50\%)$ of the transformed coefficients. In other words, the chance that an attack will make the number of increased and of decreased coefficients equal is very low. In this paper, we propose an efficient modulation strategy, which is composed of positive modulation (increasing the magnitude of transformed coefficients) and negative modulation (decreasing the magnitude of transformed coefficients). The two modulation rules simultaneously hide two complementary watermarks in a host image so that at least one watermark survives under different attacks. Therefore, we call the proposed watermarking scheme "cocktail watermarking." The proposed cocktail watermarking scheme can embed watermarks firmly and make them hard to simultaneously remove. We have also conducted a statistical analysis to derive a lower bound, which provides the worst likelihood that the better watermark (out of the two) can be extracted. With this "high" lower bound, it is ensured that a "better" extracted watermark is always obtained. Experimental results confirm that our watermarking scheme can be robust to different kinds of attacks, including combined ones. In addition to the tests of several attacks in this paper, extensive tests had also been done in [4], [7].

The remainder of this paper is organized as follows. In Section II, we shall introduce the random modulation technique commonly used in conventional watermarking methods and propose a new modulation strategy called "complementary modulation" to satisfy the robustness requirement. In addition, statistical analysis is conducted to compute the lower bound of the worst likelihood that the embedded watermarks may be extracted. The combined and balanced attacks will be addressed in Section II-D. Our cocktail watermarking scheme, including encoding and decoding, will be presented in Sections III and IV, respectively. In Section IV-B, we shall provide false negative/positive analysis of bipolar watermark detection. Experimental results will be given in Section V, and concluding remarks will be made in Section VI.

II. Modulation Strategy

In the transformed domain, watermark modulation is an operation that alters the values of selected transformed coefficients using every selected coefficient's corresponding watermark value. In Section II-A, we shall introduce and analyze the modulation techniques commonly used in the existing watermarking methods and point out the inadequacy of random modulation. Section II-B will briefly analyze the behaviors of transformed coefficients when attacks are encountered. Section II-C will describe how to embed two watermarks which play complementary roles into a host image by means of the proposed "complementary modulation."

A. Random Modulation

Two very popular watermarking techniques, which take perceptual significance into account, were presented in [5], [31]. Cox *et al.* [5] used the spread spectrum concept to hide a watermark based on the following modulation rule:

$$I_i^m = I_i(1 + \alpha \cdot w_i) \tag{1}$$

where

- I_i and I_i^m significant DCT coefficients before and after modulation, respectively;
- w_i value of a watermark sequence;
- α is a weight that controls the tradeoff between transparency and robustness.

In [31], Podilchuk and Zeng presented two watermarking schemes based on a human visual model, i.e., the image adaptive-DCT (IA-DCT) and the image adaptive wavelet (IA-W) schemes. The watermark encoder designed for both IA-DCT and IA-W can be generally described as

$$I_{u,v}^m = \begin{cases} I_{u,v} + J_{u,v} \cdot w_{u,v}, & I_{u,v} > J_{u,v} \\ I_{u,v}, & \text{otherwise} \end{cases} \tag{2}$$

where

- $I_{u,v}$ and $I_{u,v}^m$ DCT or wavelet coefficients before and after modulation, respectively;
- $J_{u,v}$ masking value of a DCT or a wavelet based visual model;
- $w_{u,v}$ sequence of watermark values.

It is found from both embedding schemes that modulations take place in the perceptually significant coefficients with the modification quantity specified by a weight. The weight is either heuristically determined [5] or depends on a visual model [31]. Cox *et al.* [5] and Podilchuk and Zeng [31] both adopted a similar detector response measurement described by

$$\rho(W, W^e) = \frac{W \cdot W^e}{\sqrt{W^e \cdot W^e}} \tag{3}$$

where W and W^e are the original and the extracted watermark sequences, respectively. If the signs of a corresponding pair of elements in W and W^e are the same, then they contribute positively to the detector response. A higher value of $\rho(W, W^e)$ means there is stronger evidence that W^e is a genuine watermark. In (3), high correlation values can only be achieved if most of the transformed coefficients of the original image and the watermarked image are updated along the same *direction* during the embedding and the attacking processes, respectively. This is the key point if a watermark detector is to get a higher correlation value. However, we find that neither [5] nor [31] took this important factor into account. In fact, the modulation strategy they adopted is intrinsically random. Usually, a positive coefficient can be updated with a positive or a negative quantity, and a negative coefficient can be altered with a positive or a negative quantity as well. In other words, [5] and [31] did not consider the relationship between the signs of a *modulation pair*, which is composed of a selected transformed coefficient and its corresponding watermark value. This explains why many attacks can successfully defeat the above mentioned watermarking schemes.

B. Analyzing the Behaviors of Transformed Coefficients under Attacks

In the following analysis, we will assume that the watermark sequence W is embedded into a host image. For the random modulation techniques proposed in [5] and [31], there are four possible types of modulations: $\mathrm{Modu}(+,+), \mathrm{Modu}(+,-), \mathrm{Modu}(-,+)$, and $\mathrm{Modu}(-,-)$, where $\mathrm{Modu}(+/-,-/+)$ represents a positive/negative transformed coefficient modulated with a negative/positive watermark quantity. For a noise-style watermark with a Gaussian distribution of zero mean and unit variance, the probability of drawing a positive or a negative value is roughly equal to 0.5.

In the wavelet domain, the wavelet coefficients of a high-frequency band can be modeled as a generalized Gaussian distribution [1] with the mean close to zero; i.e., the probability of getting a positive or a negative coefficient is roughly equal to 0.5. The lowest frequency component is, however, only suitably modeled by a typical Gaussian distribution with the mean far away from zero. That is, the probability of obtaining a positive coefficient is extremely different from that of obtaining a negative coefficient. When wavelet decomposition is executed with many scales, the lowest frequency component is tiny. Therefore, the probability of getting a positive or a negative wavelet coefficient is still close to 0.5.

For the transformed coefficients in the DCT domain, the number of positive and that of negative global DCT coefficients are statistically very close to each other. Hence, no matter whether the DCT or the wavelet domain is employed, the probabilities of occurrence of the four types of modulations are all very close to 0.25 due to their characteristic of randomness. We have also observed the influence of a number of attacks to see how they update the magnitude of each transformed coefficient. The behaviors of attacks can be roughly classified into two categories. The first category contains those attacks like compression and blurring, which tend to decrease the magnitudes of most of the transformed coefficients of a watermarked image. Under these circumstances, it is hoped that every transformed coefficient can be modulated with a quantity that has different sign. The reason why the above modulation strategy is adopted is that it can adapt to compression-style attacks and enables more than 50% of the modulated targets to contribute a bigger positive value to the detector response. As a result, we can conclude that of the four types of modulations, only $\mathrm{Modu}(+,-)$ and $\mathrm{Modu}(-,+)$ will contribute positively to the detector response. On the other hand, the second category contains those attacks such as sharpening and histogram equalization, which have the tendency of increasing most of the magnitudes of transformed coefficients, then every constituent transformed coefficient should be modulated with a quantity that has a same sign. Under these circumstances, only $\mathrm{Modu}(+,+)$ and $\mathrm{Modu}(-,-)$ will contribute positively to the detector response. From our observations, we find that using the random modulation proposed in [5], [31], about 50% of the transformed coefficients can be increasingly modulated, and that the other half are decreasingly modulated. Therefore, it can be concluded that the random modulation strategy does not help the detector response value increase at all because it is simply an addition modulation disregarding the behaviors of attacks. We believe that a better modulation strategy should take the behaviors of attacks into account.

C. A New Modulation Strategy

In this section, we shall propose a new modulation scheme from the viewpoint of detection in order to obtain higher detector responses. It is noted that the detector response defined in (3) is a function of W and W^e. Basically, W is a hidden watermark and is, therefore, fixed once it is chosen. However, the values of W^e are dependent on the strength of an attack. Because we are concerned with preserving the consistency of modulation directions instead of the degree of changes, the watermark value is defined in the bipolar form $\{-1,1\}$, that is,

$$bipolar(t) = \begin{cases} 1, & t \geq 0 \\ -1, & t < 0, \end{cases} \tag{4}$$

where t is a real number. So, an extracted watermark value is determined from the sign of a piece of retrieved information using the bipolar test described in (4). It is noted that the following derivations are suitable for different types of watermarks (bipolar, noise, or gray-scale watermarks). The main difference is that the final detector response may reflect a totally different result.

If a watermark image has been attacked and the coordinate in the transformed domain is (x,y), then the extracted watermark value can be expressed as

$$\begin{aligned} w^e(i) &= w^e(map(x,y)) = bipolar(T^a(x,y) - T(x,y)) \\ &= bipolar((T^a(x,y) - T^m(x,y)) \\ &\quad + (T^m(x,y) - T(x,y))) \\ &= bipolar(\beta_1 + \beta_2), \quad i = 1,2,\ldots,W_L, \end{aligned} \tag{5}$$

where $T(x,y), T^m(x,y)$, and $T^a(x,y)$ represent the original, the modulated, and the attacked transformed coefficients, respectively. W_L and i denote the length and the index of a hidden watermark W, respectively. Note that the original image and its corresponding attacked image are perfectly registered if the watermarked image and the proposed relocation technique (will be described in Section IV-C) are used for registration or only geometric-free attacks (including slight geometric-distortion such as StirMark [28]) are considered. That is, there is a perfect correspondence between an attacked transformed coefficient and its original transformed coefficient. The mapping function *map* forms a one-to-one mapping (which will be described in Section III) which maps a selected transformed coefficient to its corresponding watermark index. From the analysis described in Section II-B, it is clear that in order to obtain a high detector response, the signs of $w(i)$ and $w^e(i)$ have to be the same. We can derive from (5) that there exist two possible conditions under which $w(i)$ and $w^e(i)$ will have the same sign.

First, if β_1 and β_2 have the same sign, then $bipolar(\beta_1 + \beta_2)(=w^e(i))$ and $bipolar(\beta_2)(=w(i))$ will be the same (scenario 1 in Fig. 1). We will propose a complementary modulation strategy in Section II-C1 to achieve the first condition. The second condition is that β_1 and β_2 have different signs, but that $|\beta_1| < |\beta_2|$. Under these circumstances, the modulated

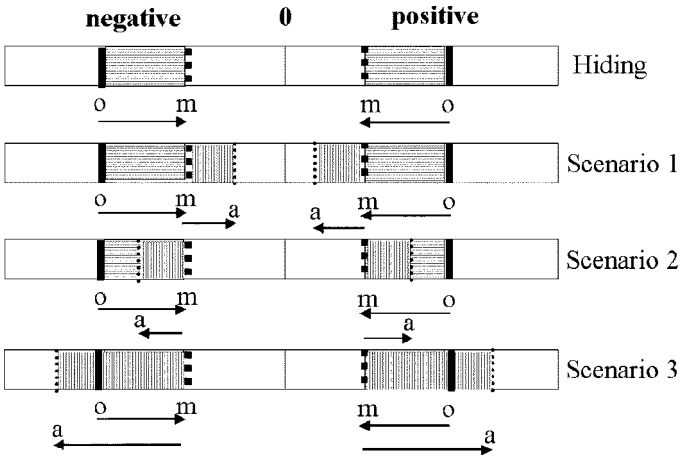


Fig. 1. Scenarios in the attacking process for negative modulation. "o" denotes the original wavelet coefficient, "m" represents the wavelet coefficient after modulation, and "a" is the coefficient after attacks; positive/negative denote the portion of positive/negative wavelet coefficients; the horizontal/vertical area represents the hiding/attacking quantity: (top figure) hiding using negative modulation; (scenario 1) the behaviors of the hiding and the attacking processes are the same; (scenario 2/scenario 3) the behaviors of the hiding and the attacking processes are different, but the strength of the attack is smaller/larger than that of negative modulation.

amount is larger than the amount altered by an attack. In other words, the applied attack is not strong enough to influence the sign change created by the modulation process. Introduction of the second condition is necessary to obtain a higher detector response because the use of a human visual model will maximize the hiding capacity. Scenario 2 in Fig. 1 illustrates the above mentioned phenomenon. In this paper, the human visual model is introduced to help determine the maximum capacity allowed to embed watermarks. More specifically, masking, the effect of a visual model, refers to the fact that a component in a given visual signal may become imperceptible in the presence of another signal, called a masker. This refers to a situation where a signal raises the visual *threshold* for other signals around it. For a given visual distance and display resolution, it is possible to determine the just noticeable distortion (JND) for each spatial frequency from specified wave functions. Psychologists have experimented with several contrast sensitivity functions (CSF) from some specific wave functions, such as the DCT basis function [27] and wavelet [38]. Since wavelet transform is very powerful in image representation, we shall use the wavelet-based visual model [38] to determine the maximum capacity that is allowed for a watermark encoder.

1) Complementary Modulation: In what follows, a complementary modulation strategy will be presented. The proposed scheme embeds two watermarks, which play complementary roles in resisting various kinds of attacks. The values of the two watermarks are drawn from the same watermark sequence. The difference is that they are embedded using two different modulation rules: **positive modulation** and **negative modulation**. If a modulation operates by adding a negative quantity to a positive coefficient $(\mathrm{Modu}(-,+))$ or by adding a positive quantity to a negative coefficient $(\mathrm{Modu}(+,-))$, then we call it "negative modulation." Otherwise, it is called "positive modulation" if the sign of the added quantity is the same as that of the corresponding wavelet coefficient $(\mathrm{Modu}(+,+)$ or $\mathrm{Modu}(-,-))$. Fig. 1 illustrates the relationship between the original coefficient, the modulated coefficient, and the attacked coefficient. These relationships will be used in explaining the proposed complementary modulation strategy. Higher detector response can always be obtained since at least one of the two watermarks is able to capture the behavior of the wavelet coefficients with respect to any attacks under the assumption that the original image is available in the detection process.

Let L^m_{NM} be a set of locations in the wavelet domain whose corresponding wavelet coefficients are to be decreased in magnitude, and let $H_{s,o}(x,y)$ and $H^m_{s,o}(x,y)$ be the original and the modulated wavelet coefficients, respectively, at (x,y). The subscripts s and o represent, respectively, scale and orientation. The explicit form of L^m_{NM} can be expressed as follows:

$$\begin{aligned} L^m_{\mathrm{NM}} &= \left\{(x,y) \,\middle|\, \left|H^m_{s,o}(x,y)\right| < \left|H_{s,o}(x,y)\right|\right\} \\ &= \left\{(x,y) \,\middle|\, \left(H^m_{s,o}(x,y) - H_{s,o}(x,y)\right) \right. \\ &\qquad \left. \cdot\, H_{s,o}(x,y) < 0\right\} \\ &= \{(x,y) \,|\, w(map(x,y)) \cdot H_{s,o}(x,y) < 0\}. \end{aligned} \tag{6}$$

Note that using negative modulation and the human visual model, $|H^m_{s,o}(x,y)| < |H_{s,o}(x,y)|$ is guaranteed to hold because the modification of $H_{s,o}(x,y)$ is bounded. This same rule is applied for position modulation. The embedding rule that specifies the condition $w(map(x,y)) \cdot H_{s,o}(x,y) < 0$ is called "**negative modulation (NM)**." The set L^m_{NM} is altered and becomes a new set, L^{m*}_{NM}, after an attack. Let L^{m*}_{NM} be expressed as

$$\begin{aligned} L^{m*}_{\mathrm{NM}} = &\left\{(x,y) \,\middle|\, \left|H^a_{s,o}(x,y)\right| < \left|H^m_{s,o}(x,y)\right|\right\} \\ &\cup \left\{(x,y) \,\middle|\, \left|H^a_{s,o}(x,y)\right| \geq \left|H^m_{s,o}(x,y)\right|\right\}. \end{aligned}$$

The set of elements L^a_{NM}, which indicates the locations where the embedding and the attacking processes behave *consistently*, should be identified. This set can be expressed as follows:

$$\begin{aligned} L^a_{\mathrm{NM}} &= L^m_{\mathrm{NM}} \cap L^{m*}_{\mathrm{NM}} \\ &= \left\{(x,y) \,\middle|\, \left|H^m_{s,o}(x,y)\right| < \left|H_{s,o}(x,y)\right|\right\} \\ &\quad \cap \left\{(x,y) \,\middle|\, \left|H^a_{s,o}(x,y)\right| < \left|H^m_{s,o}(x,y)\right|\right\} \\ &= \left\{(x,y) \,\middle|\, \left|H^a_{s,o}(x,y)\right| < \left|H^m_{s,o}(x,y)\right|\right\} \\ &= \{(x,y) \,|\, w(map(x,y)) \cdot w^e(map(x,y)) > 0\} \end{aligned} \tag{7}$$

where $H^a_{s,o}(x,y)$ is the attacked wavelet coefficient. Since the modulation and the attack processes behave in the same way at $(x,y), w(map(x,y)) \cdot w^e(map(x,y)) > 0$ holds and contributes positively to the detector response. On the other hand, a "**positive modulation (PM)**" event for watermark encoding can be defined as $w(map(x,y)) \cdot H_{s,o}(x,y) > 0$. Similarly, the set of locations whose corresponding coefficients are increasingly modulated in magnitude, L^m_{PM}, and the set L^a_{PM}, which contains locations where the wavelet coefficients are increasingly modulated in magnitude by an attack given that a positive modulation event has occurred, can be defined as in the negative modulation case.

Notice that only one watermark is hidden with respect to each modulation rule (event) under this complementary modulation

strategy. It is obvious that the two sets L^m_{NM} and L^m_{PM} are disjointed. That is

$$L^m_{\mathrm{NM}} \cap L^m_{\mathrm{PM}} = \emptyset.$$

For an attack that favors negative modulation, most ($\geq$50%) of the wavelet coefficients will decrease in magnitude. Let P^a_{NM} be the probability that wavelet coefficients will be decreasingly modulated in magnitude by an attack provided that the embedding rule "**negative modulation**" has been employed. So, P^a_{NM} is defined as (8), shown at the bottom of the page, where $||S||$ denotes the number of elements in the set S. It is not hard to realize that $P(|H^m_{s,o}(x,y)| < |H_{s,o}(x,y)|) = ||L^m_{\mathrm{NM}}||/W_L$ holds. Ideally, the condition $P^a_{\mathrm{NM}} = 1$ only holds for an attack whose behavior completely matches negative modulation. That is, all the coefficients of the original image and the watermarked image decrease. In fact, it is difficult for an attack to match the behavior of negative modulation completely. Therefore, the relation $||L^a_{\mathrm{NM}}|| \leq ||L^m_{\mathrm{NM}}||$ holds. Furthermore, under the assumption that the attack favors negative modulation, $1/2||L^m_{\mathrm{NM}}|| \leq ||L^a_{\mathrm{NM}}||$ holds. That is

$$\frac{1}{2}\left||L^m_{\mathrm{NM}}\right|| \leq \left||L^a_{\mathrm{NM}}\right|| \leq \left||L^m_{\mathrm{NM}}\right||, \tag{9}$$

and

$$P^a_{\mathrm{NM}} \in [0.5\ 1]. \tag{10}$$

From (10), we know that more than or exactly 50% of the pairs of $(w(\cdot), w^e(\cdot))$ will have the same sign and, thus, will contribute positively to the detector response. These pairs result from the fact that more than or exactly 50% of the wavelet coefficients' magnitudes decrease. Similar procedures can be deduced to compute P^a_{PM} given that positive modulation has occurred. One may ask what will happen if we do not know the tendency of an attack in advance. Fortunately, since our approach hides two complementary watermarks in a host image, at least one modulation will match the behavior of an arbitrary attack with the probability, P^a, guaranteed to be larger than or equal to 0.5; i.e.,

$$P^a = \max\{P^a_{\mathrm{NM}}, P^a_{\mathrm{PM}}\} \geq 0.5. \tag{11}$$

D. Complementary Modulation under Combined Attack and Balanced Attack

As discussed in Section II-C1, our complementary modulation scheme can tolerate a great number of attacks. However, robustness against a combined attack or a balanced attack has not been addressed. In this section, we shall explain how our scheme can survive under a combined attack or a balanced attack. First of all, we must define what a combined attack is. In this paper, a combined attack is defined as an attack composed of several (more than one) attacks of the same type or of different types.

Recall that watermarks are encoded in a host image using the positive/negative modulation rules so as to yield so-called positively/negatively modulated watermarks. If one can positively/negatively modulate almost or more than 50% of the transformed coefficients of the negatively/positively modulated hidden watermark, then the embedded watermarks are said to have been successfully removed. Practically speaking, this is the only way to make our cocktail watermarking scheme fail. However, it is extremely difficult to correctly guess most of the positions of the two embedded watermarks even if an attack is organized in a combined form.

On the other hand, a balanced attack is an attack which is able to either increase or decrease the modified image pixels within a close approximation. One may argue that such an attack will successfully remove most of our hidden watermarks. However, one can find that results obtained after a balanced attack are similar to those obtained after performing a combined attack. We shall describe some experiments which were conducted to check the robustness of our scheme under combined attacks and balanced attacks in Section V. The overall performance analysis will be discussed in Section IV-B.

III. Cocktail Watermark Encoding

The cocktail watermark encoding algorithm was developed based on the assumption that the original image (host image) is gray-scale. The wavelet transform adopted in this paper is constrained such that the size of the lowest band is 16 $\times$ 16. Here, the hidden watermark is either a noise-style watermark or a bipolar watermark. Gray-scale watermark hiding can be found in our previous work [21]. A noise-style watermark is Gaussian distributed with zero mean and unit variance. On the other hand,

$$\begin{aligned}
P^a_{\mathrm{NM}} &= P(\textbf{coefficients that are decreasingly modulated by an attack}\,|\,\mathbf{NM}) \\
&= \frac{P\left(\left(\left|H^a_{s,o}(x,y)\right| < \left|H^m_{s,o}(x,y)\right|\right) \cap \left(\left|H^m_{s,o}(x,y)\right| < |H_{s,o}(x,y)|\right)\right)}{P\left(\left|H^m_{s,o}(x,y)\right| < |H_{s,o}(x,y)|\right)} \\
&= \frac{P\left(\left|H^a_{s,o}(x,y)\right| < \left|H^m_{s,o}(x,y)\right|\right)}{P\left(\left|H^m_{s,o}(x,y)\right| < |H_{s,o}(x,y)|\right)} \\
&= \frac{||L^a_{\mathrm{NM}}||/W_L}{||L^m_{\mathrm{NM}}||/W_L} \\
&= \frac{||L^a_{\mathrm{NM}}||}{||L^m_{\mathrm{NM}}||}
\end{aligned} \tag{8}$$

a bipolar watermark value is defined as the sign of a noise-style watermark value, and the magnitudes of the Gaussian sequence are used as the weights for modulation. In this paper, the payload of the current cocktail watermarking system is just one bit, i.e., the presence or absence of a watermark is reported.

A. Selection of Wavelet Coefficients

The region used to hide watermarks is divided into two parts, i.e., the lowest frequency part and a part that covers the remaining frequencies. It is noted that the lowest frequency wavelet coefficients correspond to the smallest portion of a decomposition. Hence, different weights may be assigned to achieve a compromise between transparency and robustness. Similar to [31], only the frequency masking effect of the wavelet-based visual model [38] is considered here. Owing to the lack of wavelet-based image-dependent masking effects, heuristic weight assignment needs to be used.

Before the wavelet coefficients of a host image are modulated, locations for embedding must be selected. A set of wavelet coefficients is selected if their magnitudes are larger than their corresponding JND thresholds. Because two complementary watermarks need to be hidden, the length of each watermark should be one half the amount of the total of the selected coefficients. Therefore, the watermark designed using our approach is image-adaptive [31]. For the sake of security, the two hidden watermarks must randomly spread in the wavelet domain. We use a secret key to generate a Gaussian sequence, G, with zero mean with its length equal to the number of selected wavelet coefficients. The relationship between the selected wavelet coefficients and the drawn Gaussian sequence is a one-to-one mapping. The mapping function is defined as

$$map(x,y) = \begin{cases} 1, & G(i) \geq 0 \\ -1, & G(i) < 0 \end{cases} \tag{12}$$

where (x, y) is the coordinate in the wavelet domain and i is the index of the Gaussian sequence, G. Ideally, the mean of *map* will be zero. If it is not, then it is forced to be zero. The locations in the wavelet domain which correspond to positive/negative *map* values will be assigned to employ positive/negative modulation rules. In the remaining of this paper, the coordinates $(x_p, y_p)/(x_n, y_n)$ will be used to denote the location selected with respect to positive/negative *map* value, respectively. In what follows, we shall describe in detail the proposed complementary modulation rules.

B. Complementary Modulation Rules

As discussed in Section II-C1, the sign of a selected wavelet coefficient and its corresponding watermark value are very important in our complementary modulation scheme. To modulate wavelet coefficients for complementary watermark hiding, the watermark sequence W is sorted in increasing order according to their magnitudes. After sorting, let $\tilde{w}(top(i))/\tilde{w}(bottom(i))$ refer to a watermark value, which is retrieved from the first/last i position (usually negative/positive value) of the sorted sequence $\tilde{W}$. That is, $top(i) + bottom(i) = W_L + 1$, where $1 \leq i \leq W_L$. The watermark embedding process proceeds as follows. For each pair of wavelet coefficients, $H_{s,o}(x_p, y_p)$ and $H_{s,o}(x_n, y_n)$, which come from the selected coefficient sequence with $map(x_p, y_p) = 1$ and $map(x_n, y_n) = -1$, are modulated to become $H^m_{s,o}(x_p, y_p)$ and $H^m_{s,o}(x_n, y_n)$ with $|H^m_{s,o}(x_p, y_p)| > H_{s,o}(x_p, y_p)|$ and $|H^m_{s,o}(x_n, y_n)| < H_{s,o}(x_n, y_n)|$, respectively, according to the positive modulation and the negative modulation rules. There is also possible that we cannot guarantee to take the watermark value $\tilde{w}(top(i))$ or $\tilde{w}(bottom(i))$ as what we want. But this doesn't matter because the watermark sequence W is Gaussian distributed so that the number of positive watermark values will be almost equal to that of negative watermark values. Under this circumstance, the resulting watermark errors will be very small (<1% percent). The noise-style watermark hiding and the bipolar watermark hiding are, respectively, described in Sections III-B1 and III-B2.

1) Noise-Style Watermark Hiding:

Positive Modulation: See (13), shown at the bottom of the page, where $J_{s,o}(\cdot,\cdot)$ represents the JND values of a wavelet-based visual model [38]. α is a weight used to control the maximum possible modification that will lead to the least image quality degradation. It is defined as

$$\alpha = \begin{cases} \alpha_L, & H_{s,o}(\cdot,\cdot) \in \text{ lowest frequency band} \\ \alpha_H, & \text{others.} \end{cases} \tag{14}$$

α_L and α_H refer to the weights imposed on the lowest and the remaining frequency coefficients, respectively. If both of them are set to be one, they are diminished as in [31].

Negative Modulation: See (15), shown at the bottom of the next page.

2) Bipolar Watermark Hiding: For bipolar watermark hiding, the complementary strategy, like the above noise-style watermark hiding, can be expressed as the following condensed single one:

Complementary Modulation:

$$\begin{aligned} H^m_{s,o}(x,y) = H_{s,o}(x,y) + J_{s,o}(x,y) \\ \cdot bipolar(\tilde{w}(i)) \cdot |\tilde{w}(i) \cdot \alpha|, \\ |H_{s,o}(x,y)| > J_{s,o}(x,y) \end{aligned} \tag{16}$$

where $bipolar(\cdot)$ serves as a bipolar watermark value and has been defined in (4). If $map(x, y) > 0$, then positive modulation is applied and the embedded watermark value $\tilde{w}(i)$ is extracted starting from the bottom/top of the sorted watermark sequence $\tilde{W}$ when $H_{s,o}(x,y) \geq / < 0$ to guarantee $|H^m_{s,o}(x,y)| > |H_{s,o}(x,y)|$. On the contrary, if $map(x, y) < 0$, then negative modulation is applied and the embedded watermark value

$$H^m_{s,o}(x_p,y_p) = \begin{cases} H_{s,o}(x_p,y_p) + J_{s,o}(x_p,y_p) \cdot \tilde{w}(bottom(i)) \cdot \alpha, & H_{s,o}(x_p,y_p) > J_{s,o}(x_p,y_p) \\ H_{s,o}(x_p,y_p) + J_{s,o}(x_p,y_p) \cdot \tilde{w}(top(i)) \cdot \alpha, & H_{s,o}(x_p,y_p) < -J_{s,o}(x_p,y_p) \end{cases} \tag{13}$$

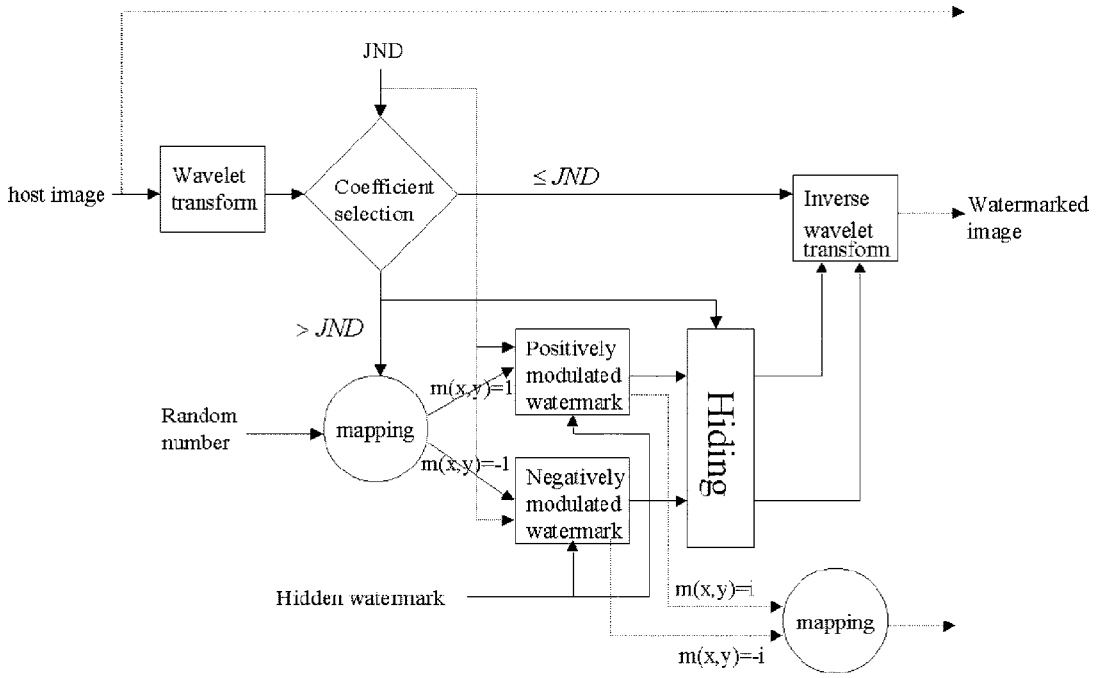


Fig. 2. Watermark embedding process of our cocktail watermarking scheme.

$\tilde{w}(i)$ is extracted starting from the top/bottom of the sorted watermark sequence $\tilde{W}$ when $H_{s,o}(x,y) \geq / < 0$ to guarantee $|H^m_{s,o}(x,y)| < |H_{s,o}(x,y)|$.

Based on the above mentioned positive and negative modulations, the mapping relationship between the position of a selected wavelet coefficient and the index of its corresponding watermark value can be finally established as

$$map(x,y) = \begin{cases} i, & G(i) \geq 0 \\ -i, & G(i) < 0. \end{cases} \tag{17}$$

These mapping results will be stored for watermark detection and kept secret such that pirates cannot easily remove the hidden watermarks. As a result, in the watermark detection process, we search for the positive/negative *signs* of $map(x,y)$ to detect watermarks embedded based on positive/negative modulation rules. Furthermore, the positive/negative *values* of $map(x,y)$ determine the index i of hidden watermarks. Fig. 2 illustrates our watermark hiding process.

IV. Cocktail Watermark Decoding

In the literature, a number of authors [2], [10], [12], [17], [19], [36] have proposed extracting a watermark without access to the original image, but the correlation values detected using their methods are not high enough, especially under strong attacks. Basically, the above mentioned methods used a prediction technique for watermark detection. Currently, the original image, in this paper, is still needed to extract watermarks due to the lack of a reliable oblivious watermarking technique. The need for a host image is suitable for destination-based watermarking [31].

A. Watermark Detection

1) Noise-Style Watermark Detection: From the watermark modulation procedures described in (13) and (15), the extracted noise-style watermark, W^e, is generated by means of a demodulation process as

$$w^e(|map(x,y)|) = \frac{H^a_{s,o}(x,y) - H_{s,o}(x,y)}{J_{s,o}(x,y) \cdot \alpha} \tag{18}$$

where *map* is a mapping function, and $H_{s,o}(x,y)$ and $H^a_{s,o}(x,y)$ are the original and the distorted wavelet coefficients, respectively. Note that we will extract two watermarks, respectively, according to the signs of *map*. The detector response is then calculated using the similarity measurement described in (3).

2) Bipolar Watermark Detection: The extracted bipolar watermark value, $w^e(\cdot)$, is expressed as

$$w^e(|map(x,y)|) = bipolar\left(H^a_{s,o}(x,y) - H_{s,o}(x,y)\right). \tag{19}$$

To calculate the detector response for bipolar watermarks, the used correlator is

$$\rho(W, W^e) = \frac{\sum w(i) w^e(i)}{W_L} \tag{20}$$

where $w(i)$ $(i = 1, 2, \ldots, W_L)$ is the sequence of embedded watermark values, $w^e(i)$ is the extracted watermark values, and W_L is the length of the hidden watermark.

$$H^m_{s,o}(x_n, y_n) = \begin{cases} H_{s,o}(x_n, y_n) + J_{s,o}(x_n, y_n) \cdot \tilde{w}(top(i)) \cdot \alpha, & H_{s,o}(x_n, y_n) > J_{s,o}(x_n, y_n) \\ H_{s,o}(x_n, y_n) + J_{s,o}(x_n, y_n) \cdot \tilde{w}(bottom(i)) \cdot \alpha. & H_{s,o}(x_n, y_n) < -J_{s,o}(x_n, y_n). \end{cases} \tag{15}$$

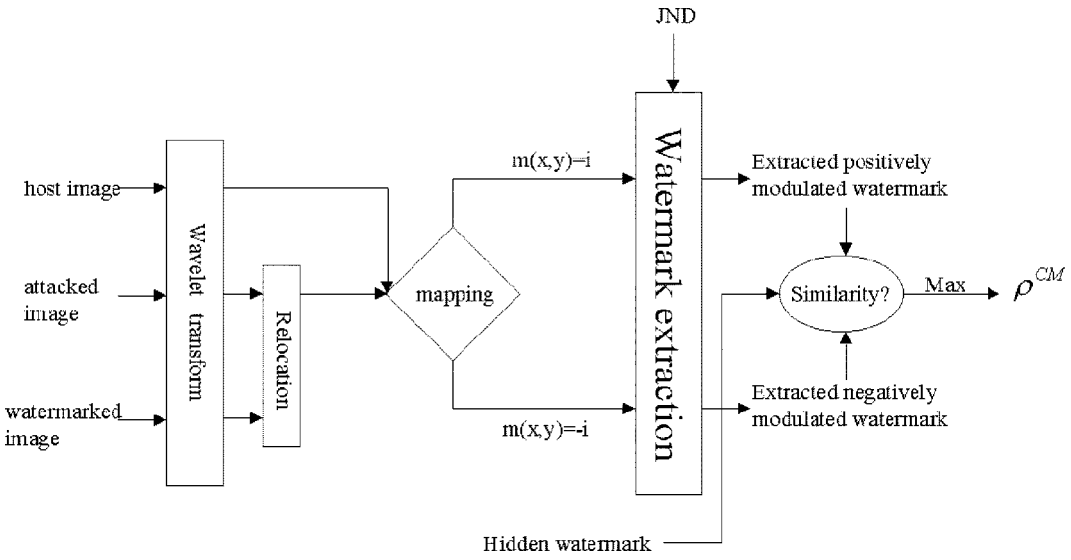


Fig. 3. Watermark detection process of our cocktail watermarking scheme.

3) Choice of a Higher Detector Response: According to the mapping function, the detector responses resulting from positive modulation and negative modulation are represented by $\rho^{\text{pos}}(\cdot,\cdot)$ and $\rho^{\text{neg}}(\cdot,\cdot)$, respectively. The final detector response, $\rho^{\text{CW}}(\cdot,\cdot)$, is thus defined as

$$\rho^{\text{CW}}(\cdot,\cdot) = \max(\rho^{\text{pos}}(\cdot,\cdot), \rho^{\text{neg}}(\cdot,\cdot)) \tag{21}$$

where CW is an abbreviation of cocktail watermarking. Furthermore, if the relocation step (which will be detailed in Section IV-C) is applied, then the detector response is denoted as $\rho^{\text{CW}}_{\text{Re}}(\cdot,\cdot)$; otherwise, it is denoted as $\rho^{\text{CW}}_{N\,\text{Re}}(\cdot,\cdot)$. A better detector response can be determined by calculating the maximum value of $\rho^{\text{CW}}_{\text{Re}}(\cdot,\cdot)$ and $\rho^{\text{CW}}_{N\,\text{Re}}(\cdot,\cdot)$, that is

$$\rho^{\text{CW}}(\cdot,\cdot) = \max\left(\rho^{\text{CW}}_{\text{Re}}(\cdot,\cdot), \rho^{\text{CW}}_{N\,\text{Re}}(\cdot,\cdot)\right). \tag{22}$$

Fig. 3 illustrates the complete procedure used in our watermark detection process.

B. Performance Analysis of Bipolar Watermark Detection

The probabilities of false negative (miss detection, failure to detect an existing watermark) and false positive (false alarm) can be estimated to support the proposed watermarking method. Here, we use a bipolar watermark as an example to compute all necessary estimations. In general, the probability of false negative using our cocktail watermarking can be derived as

$$\begin{aligned} P^{\text{CW}}_{fn} &= P\left\{\rho\left(W_{\text{pos}}, W^e_{\text{pos}}\right) \right. \\ &\quad \left. < \epsilon \,\&\, \rho\left(W_{\text{neg}}, W^e_{\text{neg}}\right) < \epsilon \,|\, W\right\} \\ &= P\left\{\rho\left(W_{\text{pos}}, W^e_{\text{pos}}\right) < \epsilon \,|\, W\right\} \\ &\quad \cdot P\left\{\rho\left(W_{\text{neg}}, W^e_{\text{neg}}\right) < \epsilon \,|\, W\right\} \\ &= P^{\text{pos}}_{fn} \cdot P^{\text{neg}}_{fn} \end{aligned} \tag{23}$$

where ϵ is the threshold used to decide the existence of an extracted watermark. Equation (23) is derived based on the fact that the two events, $\rho(W_{\text{pos}}, W^e_{\text{pos}}) < \epsilon$ and $\rho(W_{\text{neg}}, W^e_{\text{neg}}) < \epsilon$, are independent. It should be noted that if multiple watermarks are embedded using the same modulation rule, then all the events will be the same. Index *pos/neg* denotes that the watermarks are embedded using the positive/negative modulation rule and W/W^e represents the original/extracted watermark. Since the hidden watermark value is bipolar, the original and the extracted watermark values either have the same sign (i.e., $w_t(i)w^e_t(i) = 1$) or have different signs (i.e., $w_t(i)w^e_t(i) = -1$), where $t \in \{pos, neg\}$. It can be shown that $\sum w_t(i)w^e_t(i)$ belongs to the set $\{-W_L, -W_L+2, \ldots, W_L-2, W_L\}$ or to $\sum w_t(i)w^e_t(i) = W_L - 2j$, where $j \in [0\ W_L]$. Let p_1 be the probability of $w_t(i)w^e_t(i) = 1$; it is equal to P^a_{NM} or P^a_{PM}, depending on the type of attack encountered. Then, we can derive P^{pos}_{fn} as

$$\begin{aligned} P^{\text{pos}}_{fn} &= P\left\{\rho\left(W_{\text{pos}}, W^e_{\text{pos}}\right) < \epsilon \,|\, W\right\} \\ &= P\left\{\sum w_{\text{pos}}(i)w^e_{\text{pos}}(i) < W_L \cdot \epsilon \,|\, W\right\} \\ &= \sum_{j=\lceil \frac{W_L(1-c)}{2} \rceil}^{W_L} P\left\{\sum w_{\text{pos}}(i)w^e_{\text{pos}}(i) = W_L - 2j \,|\, W\right\} \\ &= \sum_{j=\lceil \frac{W_L(1-c)}{2} \rceil}^{W_L} \binom{W_L}{j} p_1^{W_L} \left(\frac{1-p_1}{p_1}\right)^j. \end{aligned} \tag{24}$$

P^{neg}_{fn} can be derived in the same way.

The derivation of P^{pos}_{fn} or P^{neg}_{fn} is similar to that of Kundur and Hatzinakos [17], but the result is extremely different since p_1 is found using a different modulation strategy. If p_1 is predicted to be 0.5 such as in [17] or other methods which use random modulation [5], [31], then the probability of false negative is

$$P_{fn} = \sum_{j=\lceil \frac{W_L(1-c)}{2} \rceil}^{W_L} \binom{W_L}{j} 0.5^{W_L}. \tag{25}$$

However, it should be noted that the probability, p_1, in our scheme is lower bounded by 0.5. It can be expected that our

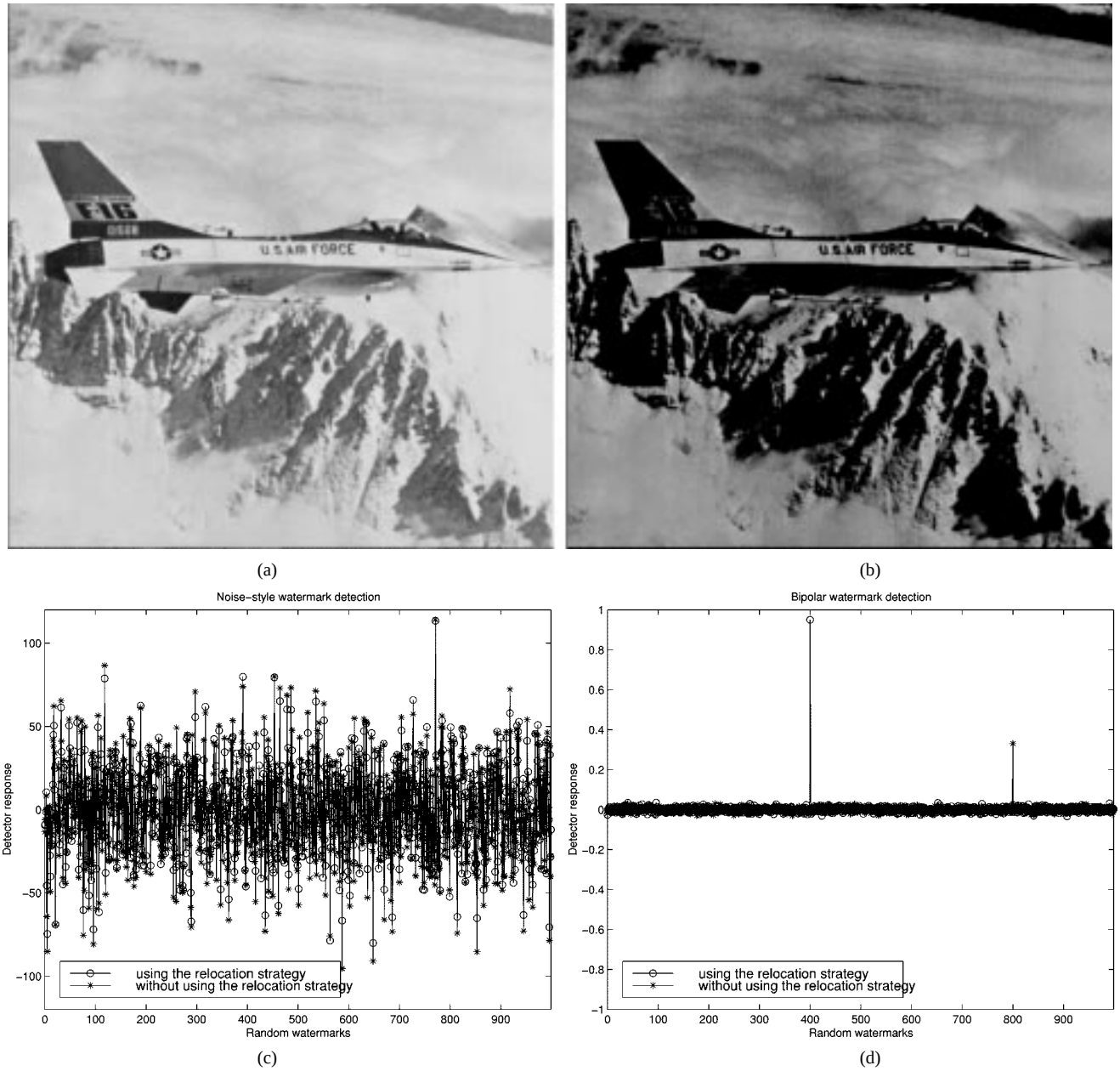


Fig. 4. Comparisons between noise-style and bipolar watermark detection: (a) watermarked image; (b) brightness/contrast attacked image; (c) and (d) detector responses of noise-style watermark/bipolar watermark with respect to 1000 random marks. The resultant detector responses corresponding to the correct watermarks 400 (using the relocation strategy) and 800 (without using the relocation strategy) are indistinguishable [shown in (c)], and are uniquely distinguished [shown in (d)] from the others.

false negative probability will definitely be smaller than those obtained using other methods. Furthermore, we would like to emphasize that it does not help reduce false negative to embed multiple watermarks with the same property [5], [31]. On the other hand, the false positive (false alarm) probability using our cocktail watermarking scheme can also be derived as

$$\begin{aligned} P_{fp}^{\mathrm{CW}} &= P\left\{\rho\left(W_{\mathrm{pos}}, W_{\mathrm{pos}}^{e}\right)\right. \\ &\quad \left.\geq \epsilon \,\|\, \rho\left(W_{\mathrm{neg}}, W_{\mathrm{neg}}^{e}\right) \geq \epsilon \,|\, not\ W\right\} \\ &= P\left\{\rho\left(W_{\mathrm{pos}}, W_{\mathrm{pos}}^{e}\right) \geq \epsilon \,|\, not\ W\right\} \\ &\quad + P\left\{\rho\left(W_{\mathrm{neg}}, W_{\mathrm{neg}}^{e}\right) \geq \epsilon \,|\, not\ W\right\} \\ &= P_{fp}^{\mathrm{pos}} + P_{fp}^{\mathrm{neg}}, \end{aligned} \tag{26}$$

where $\|$ is the "OR" operation, and P_{fp}^{pos} and P_{fp}^{neg} can be similarly derived as in (24).

The threshold ϵ can be set automatically using (23) if a desired false negative probability is given. Under the condition that the watermark length W_L and the threshold ϵ are fixed, our false negative probability is the lowest among the existing methods using random modulation. If we want to reduce the false negative probability, ϵ has to be decreased but at the expense of increasing the false positive probability.

C. Relocation for Attacks That Generate Asynchronous Phenomena

In this section, we shall present a relocation strategy for solving the asynchronous phenomena caused by attacks if an

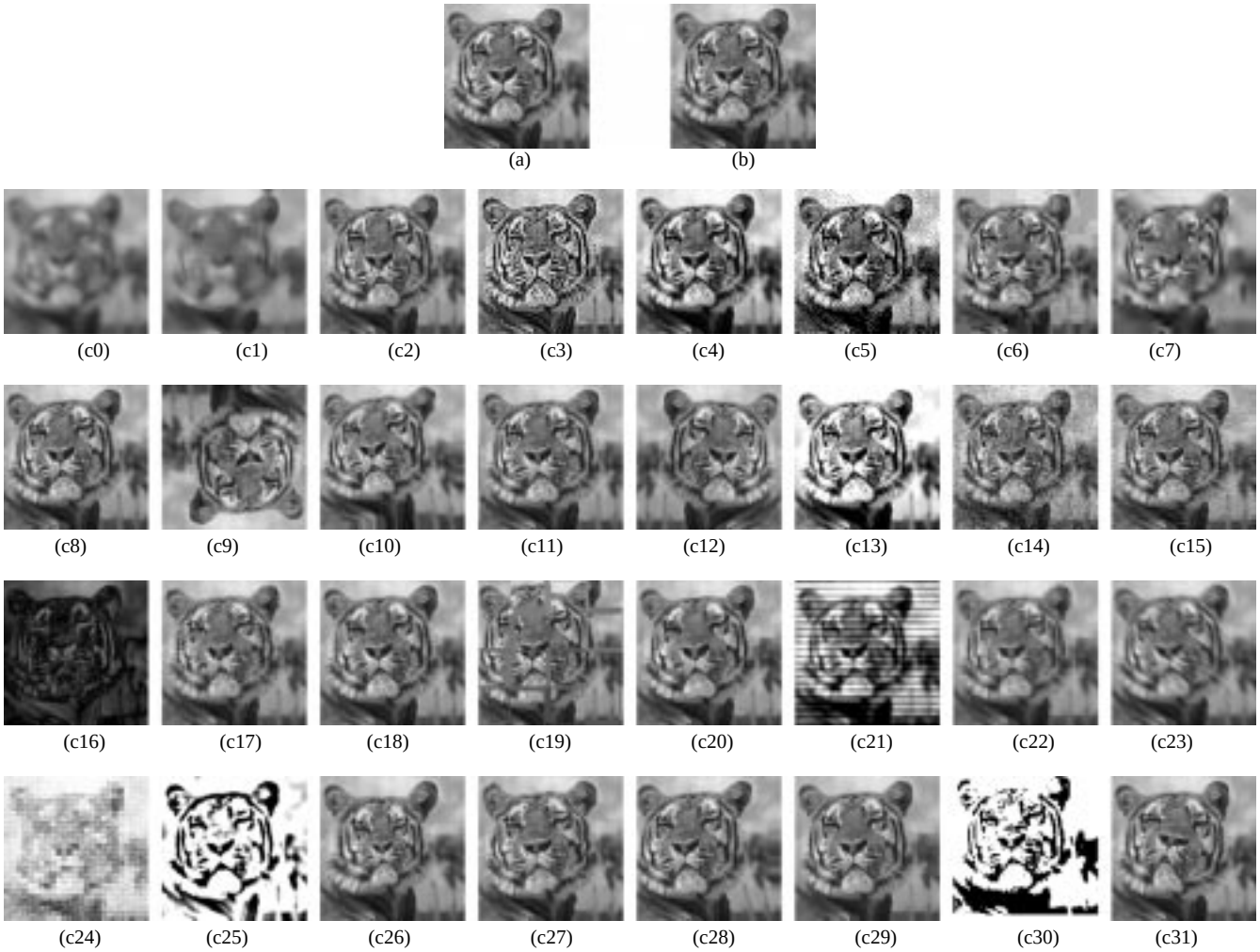


Fig. 5. (a) Host image; (b) watermarked image of (a); (c0)–(c31) attacked watermarked images: (c0) blurred (mask size 15×15); (c1) median filtered (mask size 11×11); (c2) rescaled; (c3) sharpened (with a factor 85 of *XV*); (c4) histogram equalized; (c5) dithered; (c6) JPEG compressed (with a quality factor of 5%); (c7) SPIHT [33] (at a compression ratio of 64:1); (c8) StirMark attacked (1 time with all default parameters); (c9) StirMark+Rotated 180°; (c10) StirMark attacked (5 times with all default parameters); (c11) jitter attacked (five pairs of columns were deleted/duplicated); (c12) flip; (c13) brightness/contrast adjusted; (c14) Gaussian noise added; (c15) texturized; (c16) difference of clouds; (c17) diffused; (c18) dusted; (c19) extruded; (c20) faceted; (c21) halftoned; (c22) mosaiced; (c23) motion blurred; (c24) patchworked; (c25) photocopied; (c26) pinched; (c27) rippled; (c28) sheared; (c29) smart blurred; (c30) thresholded; (c31) twirled.

watermarked image can be used. For those oblivious watermarking methods robust to geometric attacks without referring to any prior information in recovering geometric effects, readers should refer [20], [26], [32]. In what follows, we shall introduce some attacks of this sort. StirMark [28] is a very strong type of attack that defeats many existing watermarking techniques. Analysis of StirMark [28] has shown that it introduces unnoticeable quality loss in an image with some simple geometrical distortions. Jitter [29], which leads to spatial errors in images that are perceptually invisible, is another example. Basically, these attacks cause asynchronous problems. Experience tells us that an embedded watermark is often severely degraded [21] when these attacks are encountered. Therefore, it is important to deal with such an attack so that damage can be minimized. It is noted that the order of wavelet coefficients is different before and after an attack and might vary significantly under attacks having the inherent asynchronous property. Consequently, in order to recover a "correct" watermark, the wavelet coefficients of an attacked watermarked image must be relocated to their original positions before watermark detection is executed. In the relocation operation, we propose to re-arrange the wavelet coefficients of an attacked watermarked image into the same order as those of its corresponding watermarked image. Recall that $H^m_{s,o}(x,y)$ and $H^a_{s,o}(x,y)$ are the modulated and the attacked wavelet coefficients, respectively. Let the modulated wavelet coefficients be sorted in an increasing order having coordinates $(C^m_X(i), C^m_Y(i))$, where $1 \le i \le ImageSize$. So, we have $H^m_{s,o}(C^m_X(i), C^m_Y(i)) < H^m_{s,o}(C^m_X(i+1), C^m_Y(i+1))$. Let the coordinates of attacked wavelet coefficients be stored as $(C^a_X(i), C^a_Y(i))$ in the order of row-major without sorting. Then, the re-arranged attacked wavelet coefficients are

$$\tilde{H}^a_{s,o}\left(C^m_X\left(C^a_X(i)\right), C^m_Y\left(C^a_Y(i)\right)\right) = H^a_{s,o}\left(C^a_X(i), C^a_Y(i)\right).$$

Generally speaking, by preserving the orders damage to the extracted watermark can always be reduced. In the experiments, one can find that the detector response measured after applying the relocation step is significantly improved.

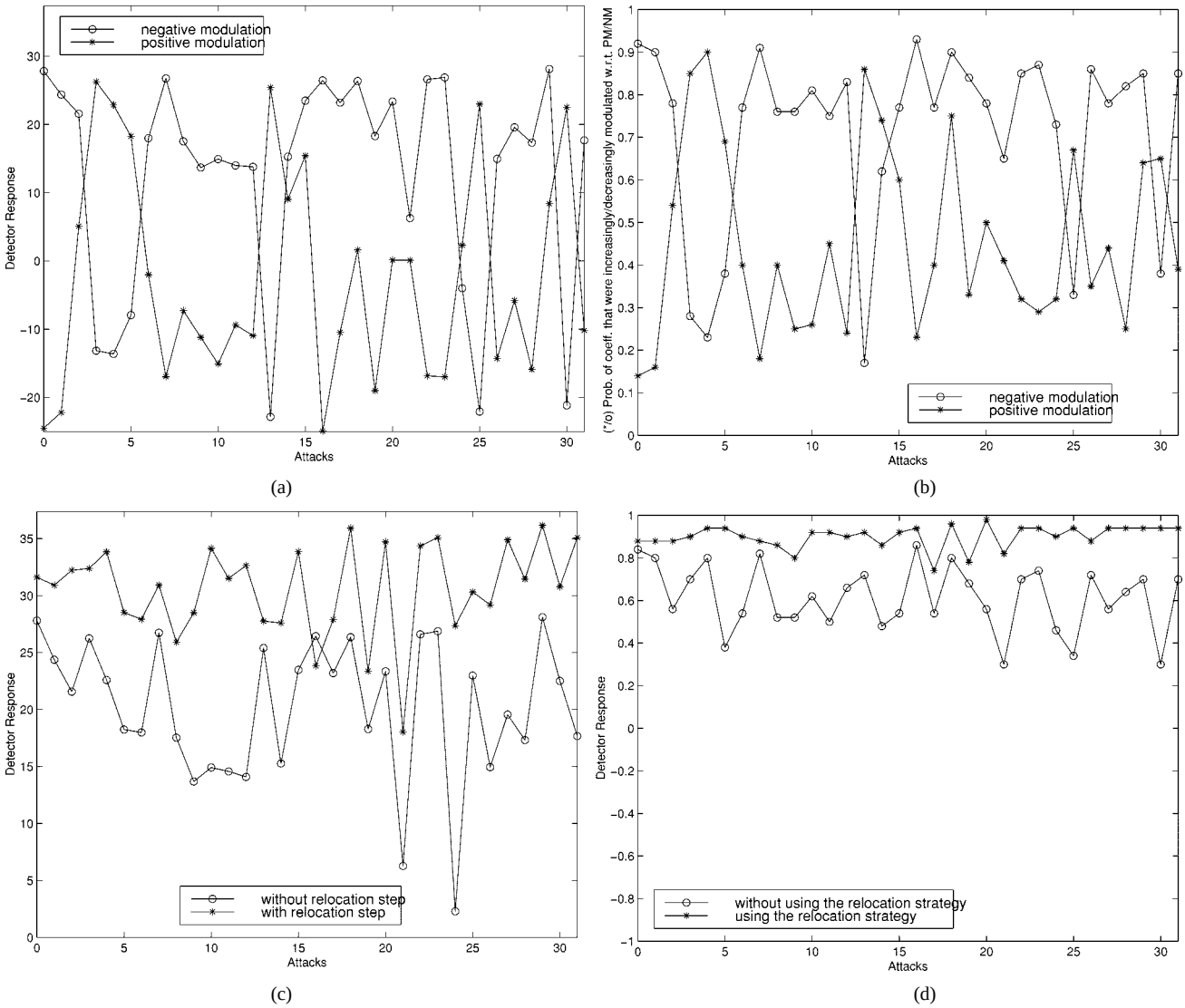


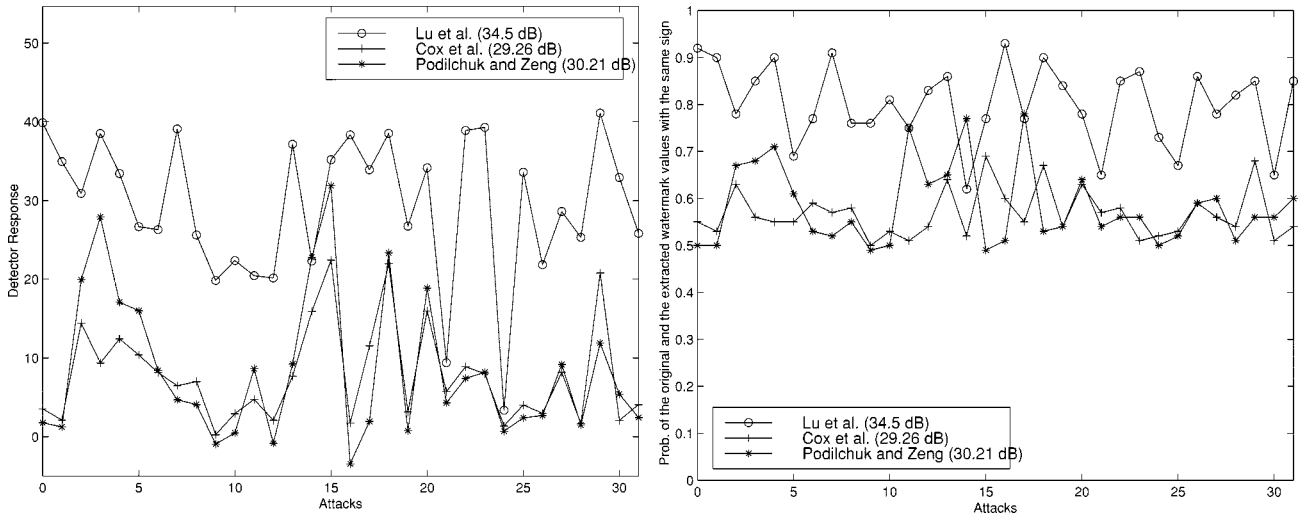
Fig. 6. Results obtained using cocktail watermarking (where the maximum detector response was 37.37 and 1 for noise-style and bipolar watermarks detection, respectively): (a) the obtained detector responses (without relocation step) under 32 attacks after noise-style watermark detection; (b) probabilities of coefficients that were increasingly/decreasingly modulated with respect to positive/negative modulation; (c) a comparison of the detector responses with/without use of the relocation step after noise-style watermark detection; (d) a comparison of the detector responses with/without use of the relocation step after bipolar watermark detection.

V. Experimental Results

A series of experiments was conducted to verify the effectiveness of the proposed method. The experimental results are reported in the following.

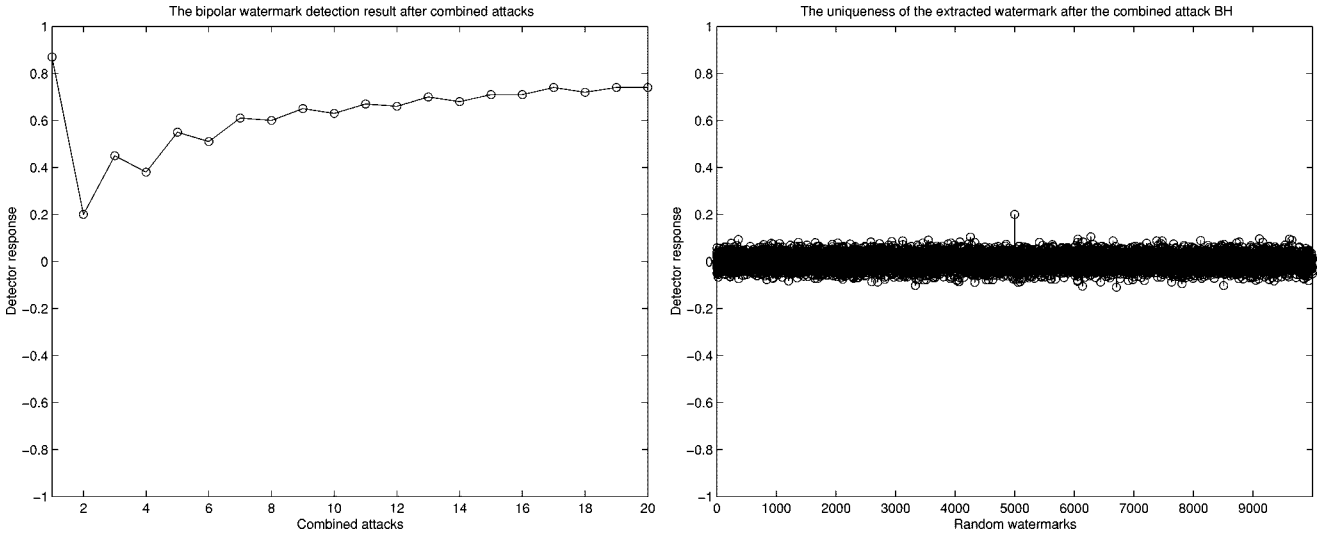
A. Bipolar Watermark versus Noise-Style Watermark

This experiment was intended to show that the detector responses obtained by embedding a bipolar watermark were superior to those obtained by embedding a noise-style watermark. Fig. 4(a) and (b) show a watermarked image and its brightness/contrast attacked version, respectively. Basically, the histogram of the watermarked image is significantly changed after the attack. Fig. 4(c) shows the noise-style watermark detection results against 1000 randomly generated watermarks. The two correct noise-style watermarks were located at the 400 (using the relocation strategy) and the 800 (without using the relocation strategy) positions, respectively. It is obvious that the detector responses of the two correct watermarks are indistinguishable among the 1000 detector responses. However, when a bipolar watermark was used, the resultant detector response corresponding to the correct watermark could be uniquely identified as shown in Fig. 4(d). This example illustrates that even when the signs of an extracted watermark are mostly kept the same as those of the original watermark, their correlation values calculated using (3) may be small. This is because the extracted noise-style watermark is dramatically altered such that the detector response is significantly reduced. An advantage of embedding a bipolar watermark instead of a noise-style watermark lies in its capability of tolerating combined attacks or repeated attacks. It is well known that when a noise-style watermark is embedded, the resultant detector response may drop significantly when a combined attack or a balanced attack is executed. As for a bipolar watermark, since its value is determined



(a) (b)

Fig. 7. A comparison between our method, Podilchuk and Zeng's method [31], and Cox *et al.*'s method [5]: (a) comparison in terms of detector responses with respect to 32 attacks (the normalized maximum detector response is 54.64); (b) comparison of the probabilities that the original and the extracted watermark values will have the same sign.



(a) (b)

Fig. 8. Combined attacks using blurring (B) and histogram equalization (H): (a) bipolar watermark detection results (without using the relocation technique) with respect to combined attacks; (b) the uniqueness of the extracted watermark obtained after combined attack BH among 10 000 random marks.

by the sign instead of the magnitude, its corresponding detector response will not be influenced by a balanced attack or a combined attack.

B. Complementary Effects of Cocktail Watermarking

As explained in the sequel, the performance of our cocktail watermarking was demonstrated by hiding both noise-style and bipolar watermarks. A tiger image of size 128×128, as shown in Fig. 5(a), was used in the tests. The length of a hidden watermark depends on the host image and the wavelet-based visual model. Here, its length was 1357. Using our modulation strategy, a total of 2714 wavelet coefficients needed to be modulated. The PSNR of the watermarked image [Fig. 5(b)] was 34.5 dB. We used 32 different attacks to test our cocktail watermarking scheme. The 32 attacked watermarked images are illustrated in Fig. 5. Among them, the attacked images [labeled (13) to (31)] were generated using PhotoShop while the others were obtained by applying common image processing techniques. The detector responses, $\rho^{\mathrm{CW}}_{N\,\mathrm{Re}}(\cdot,\cdot)$ (without employing the relocation step) with respect to the 32 attacks are plotted in Fig. 6(a). The two curves clearly demonstrate complementary effects. It is apparent that one watermark could be destroyed while the other one survived well. From the set of attacked watermarked images, it is not difficult to find that some attacks severely damaged the watermarked image, but that the embedded watermarks could still be extracted with high detector response. In addition, the probabilities, P^{a}_{PM} and P^{a}_{NM}, which correspond to the positive and the negative modulations (without employing the relocation step), are plotted in Fig. 6(b). It is obvious that the cocktail watermarking strategy enabled at least one watermark to have a high probability of survival under different kinds of attacks. Moreover, the detector responses yielded by $\rho^{\mathrm{CW}}_{N\,\mathrm{Re}}(\cdot,\cdot)$

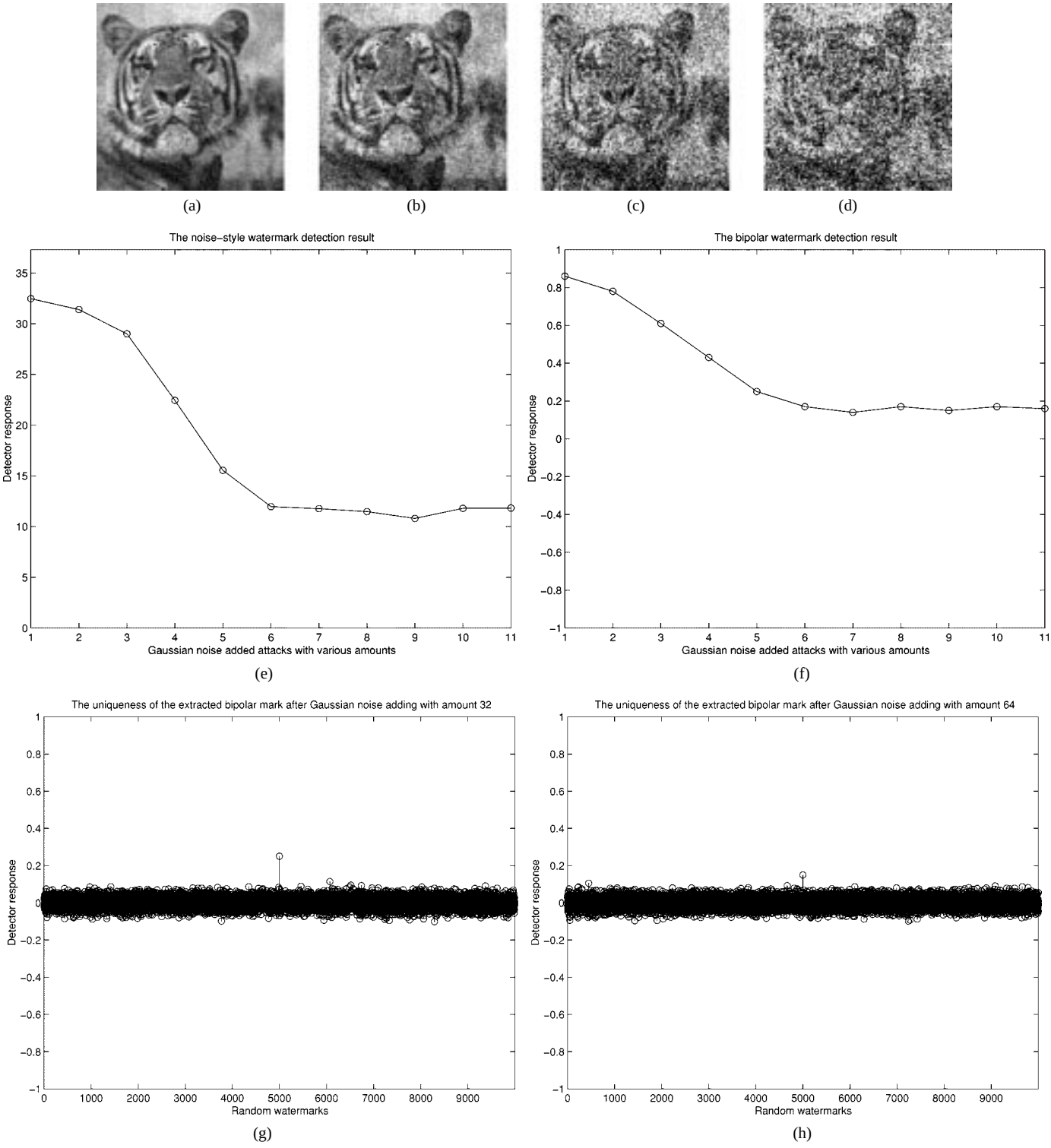


Fig. 9. Cocktail watermarking (without using the relocation technique) used against balanced attacks (Gaussian noise adding in amounts of 2, 4, 8, 16, 32, 48, 64, 80, 96, 112, 128): (a)–(d) Gaussian noise added watermarked images with amounts 16, 32, 64, and 96, respectively; (e) noise-style watermark detection; (f) bipolar watermark detection; (g) and (h) uniqueness verification of bipolar watermarks extracted under Gaussian noise added in amounts of 32 and 64, respectively, among 10 000 random marks.

and $\rho_{\mathrm{Re}}^{\mathrm{CW}}(\cdot,\cdot)$ were also compared to identify the significance of relocation. Fig. 6(c) shows two sets of detector responses, one for detection with relocation and the other for detection without relocation. From Fig. 6(c), one can see that the asynchronous phenomena caused by attacks were compensated by the relocation strategy. On the other hand, the result of detecting the bipolar watermark is shown in Fig. 6(d) for comparison. Again, almost all the detector responses were well above a certain threshold except for some detection results.

The cocktail watermarking scheme was also compared with the methods proposed by Cox *et al.* [5] and Podilchuk and Zeng (IA-W) [31] under the same set of attacks. In order to make a

fair comparison, the parameters used by Cox *et al.* [5] were adopted. The PSNR of their watermarked image was 29.26 dB. Podilchuk and Zeng's method was image-adaptive and required no extra parameter. The PSNR of their watermarked image was 30.21 dB. In our cocktail watermarking scheme and Podilchuk and Zeng's approach, three-level wavelet transform was adopted for decomposing the tiger image. Among the three watermarked images generated, respectively, by Cox *et al.*'s method, Podilchuk and Zeng's method, and our method, our watermarked image had the highest PSNR. In other words, our watermark was the weakest in terms of strength. In order to make the comparison fair, the relocation step which would have made our approach even better was not used. Because the maximum detector responses generated by an attack-free watermarked image with respect to the three compared schemes were different, a normalization step was performed so that their maximum correlation values would be the same. A comparison of the detector responses with respect to the 32 attacks for the above three methods is shown in Fig. 7(a). In addition, the comparisons of the probability P^a mentioned in (10) is displayed in Fig. 7(b). It is observed that our complementary modulations quite consistently had higher probabilities than did random modulations [5], [31] (except for the 14th attack) even though our watermark's strength was the weakest. Recall that as we have discussed in Section II-C, greater strength is beneficial for achieving a higher detector response. From the experimental results described above, it is obvious that our scheme outperforms the other two.

C. Cocktail Watermarking under Combined Attacks

In this section, we will discuss a series of experiments conducted to show how a combined attack would influence a cocktail watermarked image. It has been found that blurring (B) and histogram equalization (H) are two types of attacks which have extremely different effect on a watermarked image. That is, the blurring operation tends to decrease the magnitudes of most of the wavelet coefficients. Histogram equalization, on the other hand, tends to increase the magnitudes of most of the wavelet coefficients. The purpose of this experiment was to check whether this kind of combination is able to remove the watermark of a cocktail watermarked image. Twenty combined attacks, including B(1st attack), BH(2nd attack), BHB(3rd attack), BHBH,. . ., BHBHBHBHBHBHBHBHBHBH(20-th attack), were used. Fig. 8(a) shows the curve of the bipolar watermark detector responses against 20 combined attacks with various lengths. It is not difficult to find that the results turned out to be good when combined attacks with different lengths were applied. In other words, a longer combined attack does not really mean to destroy our cocktail watermarks more seriously. In order to show the capability of watermark detection in uniqueness verification under a combined attack, we drew 10000 random marks (including the correct one) to correlate the watermark extracted after the combined attack *BH*. Fig. 8(a) shows that the detector response under the *BH* attack was the worst. Fig. 8(b) shows that the detector response corresponding to the correct mark was a small peak among the 10000 random marks. In other words, our cocktail watermarking is still robust under a combined attack.

TABLE I
FALSE NEGATIVE ANALYSIS OF COCKTAIL WATERMARKING

Threshold (T)	*Probability (p_1)*				
	0.5	0.6	0.61	0.62	0.65
0.10	1	9.1×10^{-9}	5.6×10^{-12}	5.7×10^{-11}	2.9×10^{-27}
0.12	1	1.6×10^{-6}	6.2×10^{-9}	7.9×10^{-12}	1.1×10^{-23}
0.15	1	1.0×10^{-3}	1.9×10^{-5}	1.3×10^{-7}	1.4×10^{-16}

TABLE II
FALSE POSITIVE ANALYSIS OF COCKTAIL WATERMARKING

Threshold (T)	*Probability (p_1)*
	0.5
0.10	2.2×10^{-4}
0.12	1.1×10^{-5}
0.15	8.0×10^{-8}

D. Cocktail Watermarking under Balanced Attacks with Various Strength

In this section, we shall discuss a series of experiments conducted to show whether the resultant detector responses would drop dramatically when balanced attacks with various strengths were applied. In this series of experiments, the relocation strategy was not used. Balanced attacks, such as Gaussian noise addition, are apt to force the intensity of image pixels to be bounded within a close approximation. Under these circumstances, the intensity of image pixels is just as likely to increase as decrease. Fig. 9(a)–(d) show four Gaussian noise added watermarked images (with noise amount 16, 32, 64, and 96, respectively). It is observed that the watermarked images were severely degraded when the amount of added noise increased. Fig. 9(e) shows the curve of the detector responses after noise-type watermark detection. It is noted that when the amount of added noise increased, the detector response dropped significantly at first but tended to stabilize when the amount was increased to 64. It is not difficult to find that the stabilized curve stayed at a height of 12, but we cannot simply use this result to judge the existence of a hidden watermark. As a consequence, the bipolar watermarks extracted under Gaussian noise addition with amounts of 32 and 64, respectively, were chosen to verify the uniqueness as shown in Fig. 9(g) and (h). From Fig. 9(g) and (h), we can clearly see a peak in Fig. 9(g) while the peak shown in Fig. 9(h) is not so clear. The best way to solve this problem is to seek the compromise between the false positive probability and the false negative probability discussed in Section IV-B. Tables I and II listed some estimated results for the purpose of determining an appropriate threshold. Table I shows some values of the false negative analysis. p_1 indicates the probability that the hidden watermark values and their corresponding extracted watermark values have the same sign. From Table I, it is obvious that p_1 is lower bounded by 0.5 when our cocktail watermarking scheme was applied. In the experiments described in Section V-B, the lowest detector response received among the 32 attacks was 0.3 [Fig. 6(d)], but its corresponding p_1 value was 0.65. As to the combined attacks and the balanced attacks discussed in Section V-C and this section, the lowest detector responses received were both 0.2 (under the constraint that the attacked image was not severely degraded.) Their corresponding p_1 values were both 0.6. In sum, the p_1 values are greater than or equal to 0.6 in most cases. From Table I, we can

see that the false negative probability corresponding to $p_1 = 0.6$ and threshold $T = 0.12$ was 1.6^{-6}. That means, the miss detection rate was 0.000 16%. When T was maintained at 0.12 and the p_1 value was slightly increased to 0.61, the miss detection rate was lowered down to 6.2^{-7}% which was extremely negligible. As to the false positive probabilities listed in Table II, p_1 was consistently maintained at the value of 0.5 due to the characteristic of randomness. Under the circumstances, when T was set to 0.12, the corresponding false positive probability (false alarm) was 1.1×10^{-5}, which was negligibly small. Tables I and II also listed the false negative and the false positive probabilities when T was set to 0.1 and 0.15, respectively. However, we found that when T was equal to 0.12, the trade-off between the false negative probability and the false positive probability was the best.

VI. Conclusion

A cocktail watermarking scheme for digital image protection has been developed in this work. The proposed scheme has two features: 1) embedding two complementary watermarks makes it difficult for attackers to destroy both of them; 2) statistical analysis has provided a lower bound for our cocktail watermarking. Experimental results have demonstrated that our watermarking scheme is quite robust while still satisfying typical watermarking requirements.

Another important feature of the proposed cocktail watermarking technique is that it can be applied to other types of media such as audio [24] or video. We have also improved this nonoblivious cocktail watermarking scheme to become an oblivious one [25] while preserving equivalent robustness. In addition to the robustness issue of watermarking addressed in this paper, the rightful ownership deadlock problem [6], [39], the capacity problem [18], [34] and the public-key detection problem [11] will be important issues for future research. Except for the above requirements of robust watermarking, the need of multiple bits as a payload [37] containing information about the owner or seller of a given image in a copyright protection system is also required.

References

[1] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, "Image coding using wavelet transform," *IEEE Trans. Image Processing*, vol. 1, pp. 205–220, 1992.

[2] M. Barni, F. Bartolini, V. Cappellini, and A. Piva, "Copyright protection of digital images by embedded unperceivable marks," *Image Vis. Comput.*, vol. 16, pp. 897–906, 1998.

[3] W. Bender, D. Gruhl, N. Morimoto, and A. Lu, "Techniques for data hiding," *IBM Syst. J.*, vol. 25, pp. 313–335, 1996.

[4] G. W. Braudaway, "Results of attacks on a claimed robust digital image watermark," presented at the Proc. SPIE: Optical Security and Counterfeit Deterrence Techniques II, vol. 3314, San Jose, CA, 1998.

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

[6] S. Craver, N. Memon, B.-L. Yeo, and M. M. Yeung, "Resolving rightful ownerships with invisible watermarking techniques: Limitations, attacks, and implications," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 573–586, 1998.

[7] G. Csurka, F. Deguillaume, J. J. K. Ruanaidh, and T. Pun, "A Bayesian approach to affine transformation resistant image and video watermarking," in *3rd Int. Workshop on Information Hiding*, Dresden, Germany, 1999, pp. 270–285.

[8] J. F. Delaigle, C. De Vleeschouwer, and B. Macq, "Watermarking algorithms based on a human visual model," *Signal Process.*, vol. 66, pp. 319–336, 1998.

[9] J. Fridrich, "Applications of data hiding in digital images," presented at the Tutorial for the ISPACS Conference, 1998.

[10] ——, "Combining low-frequency and spread spectrum watermarking," presented at the SPIE Int. Symp. on Optical Science, Engineering, and Instrumentation, 1998.

[11] T. Furon and F. P. Duhamel, "Robustness of an asymmetric watermarking method," in *Proc. IEEE Int. Conf. on Image Processing*, vol. III, Vancouver, Canada, 2000, pp. 21–24.

[12] F. Hartung and B. Girod, "Watermarking of uncompressed and compressed video," *Signal Process.*, vol. 66, pp. 283–302, 1998.

[13] F. Hartung, J. K. Su, and B. Girod, "Spread spectrum watermarking: Malicious attacks and counterattacks," *Proc. SPIE: Security and Watermarking of Multimedia Contents*, vol. 3657, 1999.

[14] F. Hartung and M. Kutter, "Multimedia watermarking techniques," *Proc. IEEE: Special Issue on Protection of Multimedia Content*, vol. 87, pp. 1079–1107, 1999.

[15] C. T. Hsu and J. L. Wu, "Multiresolution watermarking for digital images," *IEEE Trans. Circuits Syst. II*, vol. 45, pp. 1097–1101, 1998.

[16] E. Koch and J. Zhao, "Toward robust and hidden image copyright labeling," presented at the Nonlinear Signal and Image Processing Workshop, Greece, 1995.

[17] D. Kundur and D. Hatzinakos, "Digital watermarking using multiresolution wavelet decomposition," *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 5, pp. 2969–2972, 1998.

[18] D. Kundur, "Energy allocation for high-capacity watermarking in the presence of compression," in *Proc. IEEE Int. Conf. Image Processing*, vol. I, Vancouver, BC, Canada, 2000, pp. 423–426.

[19] M. Kutter, F. Jordan, and F. Bossen, "Digital signature of color images using amplitude modulation," *J. Electron. Imag.*, vol. 7, pp. 326–332, 1998.

[20] C.-Y. Lin, M. Wu, J. A. Bloom, M. L. Miller, I. Cox, and Y. M. Lui, "Rotation, scale, and translation resilient public watermarking for images," presented at the SPIE Security and Watermarking of Multimedia Contents II, EI'00, San Jose, CA, 2000.

[21] C. S. Lu, S. K. Huang, C. J. Sze, and H. Y. M. Liao, *Multimedia Image and Video Processing, A New Watermarking Technique for Multimedia Protection*, L. Guan, S. Y. Kung, and J. Larsen, Eds. Boca Raton, FL: CRC, 2000.

[22] C. S. Lu, H. Y. Mark Liao, S. K. Huang, and C. J. Sze, "Cocktail watermarking on images," in *Proc. 3rd Int. Workshop on Information Hiding*, Dresden, Germany, Sept. 29–Oct. 1, 1999, pp. 333–347. LNCS 1768, http://smart.iis.sinica.edu.tw/~lcs.

[23] ——, "Highly robust image watermarking using complementary modulations," in *Proc. 2nd International Information Security Workshop*, Malaysia, Nov. 6–7, 1999, pp. 136–153. LNCS 1729.

[24] C. S. Lu, H. Y. Mark Liao, and L. H. Chen, "Multipurpose audio watermarking," in *Proc. 15th Int. Conf. on Pattern Recognition*, vol. III, Barcelona, Spain, 2000, pp. 286–289.

[25] C. S. Lu and H. Y. Mark Liao, "Oblivious cocktail watermarking by sparse code shrinkage: A regional- and global-based scheme," in *Proc. IEEE Int. Conf. on Image Processing: Special Session on Second Generation Watermarking Methods*, vol. III, Vancouver, BC, Canada, 2000, pp. 13–16.

[26] S. Pereira and T. Pun, "Robust template matching for affine resistant image watermarks," *IEEE Trans. Image Processing*, vol. 9, pp. 1123–1129, June 2000.

[27] H. A. Peterson, "DCT basis function visibility threshold in RGB space," *SID Int. Symp. Dig. Tech. Papers,*, pp. 677–680, 1992.

[28] StirMark 2.3 Watermark Robustness Testing Software (1998). [Online]. Available: http://www.cl.cam.ac.uk/~fapp2/watermarking/stirmark/

[29] F. Petitcolas, R. J. Anderson, and M. G. Kuhn, "Attacks on copyright marking systems," in *Second Workshop on Information Hiding*, 1998, pp. 218–238.

[30] ——, "Information hiding: A survey," *Proc. IEEE: Special Issue on Protection of Multimedia Content*, vol. 87, pp. 1062–1078, 1999.

[31] C. I. Podilchuk and W. Zeng, "Image-adaptive watermarking using visual models," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 525–539, 1998.

[32] J. J. K. Ruanaidh and T. Pun, "Rotation, scale, and translation invariant spread spectrum digital image watermarking," *Signal Process.*, vol. 66, pp. 303–318, 1998.

[33] A. Said and W. A. Pearlman, "A new, fast, and efficient image codec based on set partitioning in hierarchical trees," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 6, pp. 243–250, 1996.

[34] S. D. Servetto, C. I. Podilchuk, and K. Ramchandran, "Capacity issues in digital image watermarking," presented at the 5th IEEE Conf. Image Processing, 1998.
[35] M. D. Swanson, M. Kobayashi, and A. H. Tewfik, "Multimedia data-embedding and watermarking technologies," *Proc. IEEE*, vol. 86, pp. 1064–1087, 1998.
[36] P. C. Su, C.-C. Jay Kuo, and H. J. Wang, "Blind digital watermarking for cartoon and map images," presented at the SPIE Int. Symp. Electronic Imaging, 1999.
[37] S. Voloshynvskiy, S. Pereira, A. Herrigel, N. Baumgartner, and T. Pun, "A generalized watermark attack based on stochastic watermark estimation and perceptual remodulation," presented at the IST/SPIE's 12th Annu. Symp., Electronic Imaging 2000: Security and Watermarking of Multimedia Content II, vol. 3971, San Jose, CA, 2000.
[38] A. B. Watson, G. Y. Yang, J. A. Solomon, and J. Villasenor, "Visibility of wavelet quantization noise," *IEEE Trans. Image Processing*, vol. 6, pp. 1164–1175, 1997.
[39] W. Zeng and B. Liu, "A statistical watermark detection technique without using original images for resolving rightful ownerships of digital images," *IEEE Trans. Image Processing*, vol. 8, pp. 1534–1548, Nov. 1999.
[40] J. Zhao and E. Koch, "A general digital watermarking model," *Comput. Graph.*, vol. 22, pp. 397–403, 1998.

Chun-Shien Lu (M'99) was born in Tainan, Taiwan, R.O.C., on December 5, 1967. He received the Ph.D. degree in electrical engineering from National Cheng-Kung University, Tainan, in 1998. His thesis is about wavelet-based 2-D/3-D texture analysis.

From September 1994 to June 1998, he was also a research assistant in the Institute of Information Science, Academia Sinica, Taipei, Taiwan. Since October 1998, he has been a Postdoctoral Fellow in the same institute. His current research interests include multimedia digital watermarking/data hiding, image processing, digital-signature data authentication, and intelligent signal processing.

Dr. Lu was the recipient of the excellent paper award of the Image Processing and Pattern Recognition Society of Taiwan in 2000 for his work on digital watermarking.

Shi-Kun Huang received the B.S., M.S., and Ph.D. degrees in computer science and information engineering from National Chiao Tung University, Hsinchu, Taiwan, R.O.C., in 1996.

He joined the Institute of Information Science, Academia Sinica, Tainan, Taiwan, as an Assistant Research Fellow in 1997. His research interests include information security, electronic rights management, and programming language. He currently takes part in projects related to information assurance in a collaborative programming environment. he is also active in organizing CERT (computer emergency response team) and system security auditing service in Taiwan.

Chwen-Jye Sze was born in Taipei, Taiwan, R.O.C., on September 26, 1966. He received the B.S. degree in electrical engineering from the Chinese Culture University, Taipei, in 1992, the M.S. degree in electrical engineering from National Chung-Cheng University, Chiayi, Taiwan, in 1994 and the Ph.D. degree in computer science and information engineering from National Central University, Chung-Li, in 1998.

Since October 1998, he has been a Postdoctoral Fellow in the Institute of Information Science, Academia Sinica, Taipei. His research interests are in image processing, video processing, and bioinformatics.

Hong-Yuan Mark Liao (S'87–M'88) received the B.S. degree in physics from National Tsing-Hua University, Hsinchu, Taiwan, R.O.C., in 1981, and the M.S. and Ph.D. degrees in electrical engineering from Northwestern University, Evanston, IL, in 1985 and 1990, respectively.

He was a Research Associate in the Computer Vision and Image Processing Laboratory, Northwestern University, during 1990–1991. In July 1991, he joined Institute of Information Science, Academia Sinica, Taipei, Taiwan, as an Assistant Research Fellow. He was promoted to Associate Research Fellow and then Research Fellow in 1995 and 1998, respectively. From August 1997 to July 2000, he served as the Deputy Director of the institute. His current research interests include multimedia signal processing, wavelet-based image analysis, content-based multimedia retrieval, and multimedia protection. He is on the editorial boards of the *International Journal of Visual Communication and Image Representation*; the *Acta Automatica Sinica*; and the *Journal of Information Science and Engineering*.

Dr. Liao was the recipient of the Young Investigators' award of Academia Sinica in 1998; the excellent paper award of the Image Processing and Pattern Recognition society of Taiwan in 1998 and 2000; and the paper award of the above society in 1996 and 1999. He served as the program chair of the International Symposium on Multimedia Information Processing (ISMIP'1997) and will serve as the program co-chair of the second IEEE Pacific-Rim Conference on Multimedia. He also served on the program committees of several international and local conferences. He is on the Editorial Board of the IEEE TRANSACTIONS ON MULTIMEDIA. He is a member of the IEEE Computer Society.