Denoising and Copy Attacks Resilient Watermarking by Exploiting Prior Knowledge at Detector

Chun-Shien Lu*, **Hong-Yuan Mark Liao***, **and Martin Kutter****

*Institute of Information Science,
Academia Sinica, Taipei, Taiwan.
**AlpVision, Ch. des Combes 17A
1802 Corseaux, Switzerland.
E-mail: {lcs, liao}@iis.sinica.edu.tw; martin.kutter@kutter.ch

Abstract

Watermarking with both oblivious detection and high robustness capabilities is still a challenging problem. The existing methods are either robust or oblivious, but it is difficult to achieve both goals simultaneously. In this paper, we tackle the above-mentioned problem. Our basic design methodology is to exploit prior knowledge available at the detector side and then use it to design a "non-blind" embedder. We prove that the proposed scheme can resist two famous denoising-based attacks, which have successfully cracked many existing watermarking schemes. False negative and false positive analyses are conducted to verify the performance of our scheme. The experimental results show that the new method is indeed powerful.

keywords: Watermarking, Robustness, Oblivious detection, Shrinkage-based denoising, Watermark prediction, Attacks

1 Introduction

Watermarking [8, 18, 21] is a technique which conceals one or more watermarks in a medium. Embedded watermarks can be used to declare rightful ownership (robust watermarking), to authenticate credibility (fragile watermarking) or to carry useful information (captioning watermarking). Usually, a watermark itself can be a random signal, a meaningful message, or a company's logo. An effective watermarking scheme should satisfy a set of typical requirements, including transparency, robustness, oblivious (blind) detection, and so on. The main purpose of robust watermarking is to prevent hidden watermark(s) from being removed or destroyed so that ownership can be guaranteed. Watermarks can be detected with the help of the host media (called non-oblivious detection) or without access to the original media (called oblivious detection). Oblivious detection is practical but is still a challenge if high robustness is the major concern. Since the original source cannot be used in oblivious detection, the embedded watermark should be predicted from an attacked media. Under these circumstances, the predicted watermark values more or less deviate from their original ones. In other words, the degree of robustness will be affected. Therefore, robustness and oblivious detection are, in effect, in conflict with each other. However, if one can find a good watermark prediction scheme and then use it as part of the design methodology, then the degree of robustness degradation can be minimized. In this paper, we aim to tackle the above-mentioned problem using image watermarking as our domain.

Watermarking with oblivious (blind) detection [1, 10, 11, 22] has been extensively explored in recent years. Most of the existing methods detect watermarks by means of prediction, and this kind of strategy usually is not directly related to its hiding strategy. Therefore, robustness cannot be guaranteed. In [23], Voloshynovskiy *et al.* proposed a stochastic model to seriously address the watermark prediction problem. Since an oblivious approach usually detects watermarks by means of prediction, it is also possible that a pirate may successfully remove an embedded watermark by means of prediction. Voloshynovskiy *et al.* [24] called this kind of attack a "denoising and remodulation attack." In some situations, a predicted watermark may be maliciously added to another cover image that belongs to other people. This kind of attack aims to create the false positive problem. Kutter *et al.* [12] called this kind of attack a "copy attack." Since the aforementioned two attacks are very difficult to resist, any watermarking approach that claims to be *robust* may be cracked when either of the two attacks is encountered. Since a predicted watermark (for oblivious detection) may sacrifice robustness to some extent, we propose to design a watermarking system by taking both the embedding strategy and the detection strategy together into consideration. In other words, the characteristics of a predetermined detection model can be used as part of the criteria for designing a better watermarking system. In [4], Cox *et al.* proposed a new concept which views watermarking as communications with side information. This concept makes it possible to design a new watermarking method with better efficiency. In [17], Miller *et al.* adopted a similar concept [4] to design four different informed embedding strategies.

In this paper, we present a novel watermarking scheme which exploits the available information at the watermark detection (prediction) side. Based on the information obtained from the prediction side, we are able to use these prior information as part of the criteria for designing a better embedder. We shall take the shrinkage-based denoising model as our watermark prediction module [16] because it naturally leads to blind detection. Since the shrinkage-based denoising approach [6, 9] adopts a soft-thresholding strategy to "gradually" decrease the magnitude of selected coefficients, it is more linear and easy to control the behaviors of denoising. Since the knowledge at the detector side is used to design an embedder, we call it a "non-blind" embedder. In sum, the proposed system is composed of a non-blind embedder and a blind detector. We shall analyze the performance of our scheme when denoising-based attacks [12, 24] are encountered.

The remainder of this paper is organized as follows. In Sec. 2, oblivious watermark detection formulated as a denoising problem is described. In Sec. 3, the proposed scheme is described, and some performance analyses are discussed. Finally, experimental results are given in Sec. 4 and concluding remarks made in Sec. 5.

2 Formulating Oblivious Detection as Watermark Prediction by means of Shrinkage-based Denoising

In this paper, oblivious watermark detection is formulated as a watermark prediction problem. Under the assumption that a watermark hiding/attacking process is modeled as a noise adding process, we can separate an embedded watermark from an attacked image by using the shrinkage-based denoising technique. Under the circumstances, the separated noise can be regarded as an extracted watermark, which more or less deviates from its original shape due to the execution of denoising and the effects of attacks. Based on observation that the shrinkage operation tends to *gradually* decrease the magnitude of transformed coefficients, we propose to use shrinkage-based denoising to predict this noise (watermark). In the following, we will use the sparse code shrinkage (SCS) [9] strategy to model the watermark prediction process since it is a generalization of shrinkage-based image denoising methods. In Secs. 2.1 and 2.2, we shall describe in detail how to model the above-mentioned processes by means of Gaussian modeling. Next, we will discuss how to use the SCS strategy to solve the denoising problem in Sec. 2.3.

2.1 Gaussian Modeling of Coefficient Magnitude Update in the Hiding Process

Let $\mathbf{X}$ be an image, let A be a modulation operation in the hiding process, and let ψ be a wavelet function. Let the wavelet transformed image be $\mathbf{X}^{\psi}$ in the space-frequency domain. For wavelet-based watermarking, the result of modulating $\mathbf{X}^{\psi}$ by means of A is a watermarked image, $\mathbf{X}^{\psi} * A$, where $*$ is a convolution operation. $\mathbf{X}^{\psi} * A$ can be converted into another form:

$$\mathbf{X}^{\psi} * A = (\psi * \mathbf{X}) * A = \psi * (\mathbf{X} \times A^{s}) = \psi * (\mathbf{X}^{A^{s}}), \tag{1}$$

according to the associativity of convolution, where $\mathbf{X}^{A^{s}}$ is a modulated image in the spatial domain and A^{s} is the spatial version of A. The above equation indicates that the wavelet transform (using ψ) of the watermarked image $\mathbf{X}^{A^{s}}$ is equivalent to the modulation (using A) of a wavelet transformed image $\mathbf{X}^{\psi}$.

Now, suppose a watermark has been embedded into a host image in the wavelet domain. This means that the original image $\mathbf{X}$ is first wavelet transformed using ψ and then modulated using A. Under these circumstances, the modulated wavelet coefficients can be modeled as the original wavelet coefficients plus Gaussian noise added in the wavelet domain. That is,

$$w_{s,o}^{m}(x,y) = w_{s,o}(x,y) + n(i), \tag{2}$$

where $w_{s,o}(x,y)$ is the original wavelet coefficient, $w_{s,o}^{m}(x,y)$ is the modulated wavelet coefficient and $n(i)$ is the i'th element of the hidden Gaussian noise-like watermark $\mathbf{n}$. The relationship between (x,y) and i will be defined in Sec. 3. By means of Eq. (1), the Gaussian modeling can be similarly defined for $\mathbf{X}^{A^{s}}$ in the spatial domain. Let $x(j)$ be the pixel intensity of an original image $\mathbf{X}$ with size $N \times M$ at a position j $(1 \leq j \leq N \times M)$, and, let $N(j)$ be the noise value (which results from the hidden Gaussian-noise watermarks). The intensity of a noisy pixel $\tilde{x}(j)$ can be calculated by $\tilde{x}(j) = x(j) + N(j)$. Therefore, the watermarked image can be modeled as

$$\tilde{\mathbf{X}} = \mathbf{X} + \mathbf{N}, \tag{3}$$

where $\mathbf{N}$ is the noise sequence. $\tilde{\mathbf{X}}$ is equivalent to $\mathbf{X}^{A^{s}}$ in Eq. (1).

2.2 Gaussian Modeling of Coefficient Magnitude Update in the Attacking Process

After watermark hiding, the watermarked image can be transmitted over the Internet and may be attacked by any process. At this time, the model of an attack is assumed to be the same as that of a modulation operation except that (i) the original image $\mathbf{X}$ in Eq. (1) is replaced by the watermarked image $\mathbf{X}^m$; (ii) A in Eq. (1) is now regarded as an attack rather than a modulation operation; (iii) $\tilde{\mathbf{X}}$ in Eq. (3) is an attacked image instead of a watermarked image; (iv) $\mathbf{N}$ in Eq. (3) is resultant noise which is contributed by the Gaussian noise-like watermark $\mathbf{n}$ and attacks. To simplify the analysis, we assume that $\mathbf{N}$ is still a Gaussian distribution with variance σ. The value of σ will be small/large when the imposed attack is weak/strong.

2.3 Sparse Code Shrinkage (SCS) Technique

After conducting Gaussian modeling of coefficient magnitude change with respect to an attacked image, the next step is to separate the host image $\mathbf{X}$ from the attacked image $\tilde{\mathbf{X}}$ by denoising $\mathbf{N}$. Using the denoising operation, the estimated host image $\bar{\mathbf{X}}$ can correctly approximate the original image, i.e., $\bar{\mathbf{X}} \approx \mathbf{X}$. In order to achieve the above mentioned goal, the ICA-based (Independent Component Analysis) sparse code shrinkage (SCS) technique [9] is employed to model the denoising problem. An SCS-based denoising algorithm includes the following steps: (i) model the noisy image $\tilde{\mathbf{X}}$ as a set of independent components; (ii) perform sparse code shrinkage on these components; (iii) invert the ICA representation.

The step by step procedure for an ICA-based SCS denoising algorithm is given in the following. First, one has to model the host image $\mathbf{X}$ using the independent component analysis process [3]. This process decides on the major components of the host image. On the other hand, we need to consider the noise part ($\mathbf{N}$) consisting of minor components, which can be shrunk (soft-thresholded) using an adaptive soft threshold during the ICA process. In an explicit format, the host image can be modeled as $\mathbf{X} = \mathbf{As}$, where $\mathbf{A}$ is a basis matrix and $\mathbf{s}$ is the vector of independent components (ICs). Analogous to traditional transformations, such as discrete Fourier transform or wavelet transform, $\mathbf{s}$ is composed of a set of selected transformed coefficients, and $\mathbf{A}$ is a synthesis filter. Therefore, ICA has the property that different independent components (ICs) are unlikely to be activated at the same time due to its sparse distributed nature (i.e., energy compaction). Therefore, the noisy image $\tilde{\mathbf{X}}$ can be denoted as

$$\tilde{\mathbf{X}} = \mathbf{X} + \mathbf{N} = \mathbf{As} + \mathbf{N}. \tag{4}$$

Suppose only the observed data $\tilde{\mathbf{X}}$ is given; the basis matrix ($\mathbf{A}$) and the ICs ($\mathbf{s}$) can be obtained by first finding a separating matrix $\mathbf{W}$ (with $\mathbf{W}^{-1} = \mathbf{A}$) via sparse coding [9]. Then, $\mathbf{s}$ can be determined by $\mathbf{s} = \mathbf{WX}$, where each component $\mathbf{s}_i = \mathbf{W}_i\mathbf{X}$. After sparse coding, the noisy image $\tilde{\mathbf{X}}$ can be transformed by means of $\mathbf{W}$, and a noisy independent component, $\mathbf{s}+\tilde{\mathbf{N}}$, can finally be derived as follows:

$$\mathbf{W}\tilde{\mathbf{X}} = \mathbf{WX} + \mathbf{WN} = \mathbf{WAs} + \mathbf{WN} = \mathbf{s} + \tilde{\mathbf{N}}. \tag{5}$$

In the second step, each noisy component, $\mathbf{s}_i + \tilde{\mathbf{N}}_i$, is shrunk by the denoising operation. When we use sparse code shrinkage to denoise $\mathbf{s} + \tilde{\mathbf{N}}$, we need to model the distribution of each component, $\mathbf{s}_i$, to see whether it satisfies the non-Gaussian requirement. One antecedent condition that image denoising by means of shrinkage can achieve is that each component $\mathbf{s}_i$ must be non-Gaussian so that it can be distinguished from normal Gaussian noise. Due to the energy compact representation of an ICA model, every independent component $\mathbf{s}_i$ is expected to exhibit sparse density. The second condition required for image denoising by shrinkage to function is that the variance of $\mathbf{N}$ must be assumed in advance [6]. After the sparse density of each $\mathbf{s}_i$ is modeled, their corresponding parameters can be generated to determine a suitable shrinkage function, $\mathbf{g}_i$ [9]. Then, one can shrink $\mathbf{s}_i + \tilde{\mathbf{N}}_i$ by means of $\mathbf{g}_i$ and then get the cleaned version of $\mathbf{s}$, which is represented as $\bar{\mathbf{s}}$, where

$$\bar{\mathbf{s}} = \mathbf{g}_i(\mathbf{s}_i + \tilde{\mathbf{N}}_i). \tag{6}$$

Generally speaking, the value of $\bar{\mathbf{s}}$ should be very close to $\mathbf{s}$. In the third step, the approximated host image $\bar{\mathbf{X}}$ can be derived by an inverse ICA transformation: $\bar{\mathbf{X}} = \mathbf{A}\bar{\mathbf{s}}$. After the estimated host image is determined, it can be used for blind detection.

Wavelet shrinkage [6] is a good alternative to SCS-based denoising [9] due to its capability of fast computation. In wavelet shrinkage, $\mathbf{W}$ and $\mathbf{A}$ form a pair of wavelet analysis and synthesis filters. In addition, the shrinkage function used in wavelet shrinkage is fixed and is independent of the distribution of independent components. Although the denoising results obtained by applying wavelet shrinkage-based denoising are worse than those obtained by applying SCS-based denoising, their function in watermark prediction is almost the same.

3 The Proposed Denoising-based Oblivious Watermarking Method

In this section, we will describe the proposed method and analyze its performance. In Sec. 3.1, we shall describe in detail how a robust embedder can be designed by exploiting the knowledge of shrinkage-based watermark prediction. The processes of watermark embedding and watermark detection will be described as well. In Sec. 3.2, performance analysis of the proposed scheme will be presented. In Sec. 3.3, the relationship between our scheme and Cox *et al.*'s new watermarking concept [4] will be examined.

3.1 The Proposed Approach: A Non-Blind Embedder

In this section, we shall describe in detail the proposed watermarking system. Let $k(i)$ be an element of a watermark $\mathbf{K}$, and let $k(i)$ be used to modulate a wavelet coefficient $w_{s,o}(x,y)$ as follows:

$$w^m_{s,o}(x,y) = w_{s,o}(x,y) + k(i). \tag{7}$$

After simple reorganization, we have

$$sign(k(i)) = sign(w^m_{s,o}(x,y) - w_{s,o}(x,y)), \tag{8}$$

where *sign* is an operator defined as

$$sign(t) \quad = \quad \begin{cases} +1, & \text{t} \geq 0; \\ -1, & \text{t} < 0. \end{cases} \tag{9}$$

In order to maintain transparency, the sign of $w^m_{s,o}(x,y)$ has to be the same as that of $w_{s,o}(x,y)$. That is, $sign(w^m_{s,o}(x,y)) = sign(w_{s,o}(x,y))$. On the other hand, the sign of an extracted watermark $\tilde{k}(i)$ can be derived by

$$sign(\tilde{k}(i)) = sign(w^a_{s,o}(x,y) - \tilde{w}_{s,o}(x,y)) = sign(w^a_{s,o}(x,y)), \tag{10}$$

which is, in essence, a denoising-based watermark detection process. As we have mentioned previously [15], the basic requirement for obtaining a higher correlation value between $k(i)$ and $\tilde{k}(i)$ is to get them to have the same sign. In order to achieve the above goal, $sign(w^a_{s,o}(x,y)) = sign(w^m_{s,o}(x,y) - w_{s,o}(x,y))$ must hold. However, both $w^a_{s,o}(x,y)$ and $w^m_{s,o}(x,y) - w_{s,o}(x,y)$ can be either positive or negative, which makes the correlation between $k(i)$ and $\tilde{k}(i)$ hard to predict. This situation indicates that a watermarking scheme which adopts a typical spread-spectrum hiding strategy together with a shrinkage-based prediction rule cannot guarantee robustness.

From Eq. (10), we realize that the sign of an extracted watermark is dependent on the attacked wavelet coefficient due to the nature of shrinkage-based denoising. Therefore, if we can use the knowledge derived from the denoising-based prediction side, then we can design a suitable hiding strategy. In what follows, we shall discuss how to design a good hiding strategy. It is known that a pirate will not perceptually damage an image. Therefore, it is reasonable to assume that the signs of $w^m_{s,o}(x,y)$ and $w^a_{s,o}(x,y)$ are the same, i.e.,

$$sign(w^m_{s,o}(x,y)) = sign(w^a_{s,o}(x,y)). \tag{11}$$

By combining Eqs. (8), (10), and (11), we can design the watermark embedding strategy so as to satisfy $sign(w^m_{s,o}(x,y) - w_{s,o}(x,y)) = sign(w^m_{s,o}(x,y))$. That is, the watermark should be embedded in order to increase the magnitudes of the chosen coefficients such that

$$|w^m_{s,o}(x,y)| > |w_{s,o}(x,y)| \tag{12}$$

holds. This derived result is exactly the same as the effect of positive modulation of cocktail watermarking [15]. Therefore, in this paper, only one watermark will be embedded in an image using positive modulation. In order to satisfy the requirement of transparency and robustness, the highest-frequency bands and the lowest-frequency band in the wavelet domain will not be used to hide watermarks. The proposed watermarking method is described as follows.

In the watermark hiding process, suppose that $\mathbf{X}$ is an image of size $N \times M$, and that an S-scale wavelet transform is performed on $\mathbf{X}$. Let the wavelet coefficient to be modulated be $w_{s,o}(x,y)$, where $0 < s \leq S$. Since the highest-frequency subbands will not be watermarked, s must be larger than 1 (the finest scale). In addition, the lowest-frequency subband located at the S-scale is usually very small in size and is non-watermarked to preserve transparency. Therefore, it is not difficult to figure out that the length of a hidden watermark ($\mathbf{K}$) is about one-quarter of the original image size. Using positive modulation to hide a watermark $\mathbf{K}$, we obtain the modulated wavelet coefficient:

$$w^m_{s,o}(x,y) = \begin{cases} w_{s,o}(x,y) + J_{s,o}(x,y) \times \tilde{k}(bottom(i)) \times \alpha, & w_{s,o}(x,y) > J_{s,o}(x,y), \\ w_{s,o}(x,y) + J_{s,o}(x,y) \times \tilde{k}(top(i)) \times \alpha, & w_{s,o}(x,y) < -J_{s,o}(x,y), \end{cases} \tag{13}$$

where $J_{s,o}(.,.)$ (> 0) represents the JND values obtained from the visual model [25] and α ($0 < \alpha \leq 1$) is an image-dependent weight used to control the maximum possible modification that will lead to the least image quality degradation. Basically, α may be selected to satisfy perceptual transparency subjectively or to make the $PNSR$ of a watermarked image larger than a certain value objectively. We shall see later in this section that α (no matter what value it is) will not affect the detection of watermarks. $\tilde{k}(top(i))/\tilde{k}(bottom(i))$ refers to a watermark value, which is retrieved from the first/last i position (usually a negative/positive value) of the watermark sequence $\tilde{\mathbf{K}}$ sorted from $\mathbf{K}$ in decreasing order. Under these circumstances, $|w^m_{s,o}(x,y)| > |w_{s,o}(x,y)|$ is always guaranteed. The above sorted results will be recorded as

$$p(x,y) = j, \tag{14}$$

where j, denoting $top(i)$ or $bottom(i)$ of Eq. (13), is the index of the sorted watermark sequence $\tilde{\mathbf{K}}$. In the watermark detection process, an attacked image is first denoised by means of a shrinkage-based denoising process. After this process is finished, the original image can be estimated. Then, the estimated original image can be used to conduct blind watermark detection by retrieving the watermark elements $k^e(i)$ of $\mathbf{K}^e$, where

$$k^e(i) = \frac{w^a_{s,o}(x,y) - \bar{w}_{s,o}(x,y)}{J_{s,o}(x,y) \times \alpha}. \tag{15}$$

Finally, the normalized correlation value is calculated to measure the similarity between $\mathbf{K}$ and $\mathbf{K}^e$ by means of

$$\rho(\mathbf{K}, \mathbf{K}^e) = \frac{\sum_{i=1}^{||\mathbf{K}||} sign(k(i)) \times sign(k^e(i))}{||\mathbf{K}||}, \tag{16}$$

where $||\mathbf{K}||$ denotes the length of the watermark. In Eq. (16), it can be easily checked from $sign(\cdot)$ function and normalized correlation that α being image-dependent does not affect the watermark detection.

It is clear that the time bottleneck in the proposed system is in the sparse code calculation. Since efficiency is a major concern in watermark detection, we shall use wavelet transform to perform the shrinkage-based denoising task [6]. Since a secret key is required to generate a hidden watermark and this hidden watermark must be sorted in the embedding process, a sorted watermark instead of a secret key needs to be provided in the watermark detection process. This implies that the secret key (in fact, a secret sequence) is longer than those in conventional methods. In addition, we have to enforce each image to be associated with a secret sequence.

3.2 Performance Analysis

Some issues regarding performance evaluation of the proposed method are discussed in the following.

3.2.1 False Negative and False Positive Analysis

In our scheme, we use a threshold T to indicate the presence/absence of a watermark if a correlation value is larger/smaller than T. The error probabilities, composed of a false negative (miss detection) and a false positive (false alarm), will be used to evaluate our system. In our analysis, the distributions of the detection results with respect to attacked images (including watermarked images) and non-watermarked images are, respectively, approximated using Gaussian probability density functions (PDFs). In fact, the detection results of attacked images are represented using a normal Gaussian distribution while those of non-watermarked images are represented using a generalized Gaussian. The statistics of the above mentioned distributions can be estimated by means of experiments. Suppose the mean and the variance of the distribution of non-watermarked images and those of the attacked images are μ_n, σ_n^2 and μ_a, σ_a^2, respectively, with $\mu_n < \mu_a$. The intersection area of the two distributions is defined as the error probability, and the intersection point of the above two distributions is defined as the threshold T $(-1 \leq T \leq 1)$. Then, the false negative probability can be derived as follows:

$$\begin{aligned} p_{fn} &= \frac{\int_{-1}^{T} e^{\frac{(x-\mu_a)^2}{2\sigma_a^2}} dt}{\int_{-1}^{1} e^{\frac{(x-\mu_a)^2}{2\sigma_a^2}} dt} \\ &= \frac{erf(\frac{(1+\mu_a)}{\sqrt{2}\sigma_a}) + erf(\frac{\mu_a - T}{\sqrt{2}\sigma_a})}{erf(\frac{(1+\mu_a)}{\sqrt{2}\sigma_a}) + erf(\frac{1-\mu_a}{\sqrt{2}\sigma_a})}. \end{aligned} \tag{17}$$

Similarly, the false positive probability can be derived as

$$\begin{aligned} p_{fp} &= \frac{\int_{T}^{1} e^{\frac{(x-\mu_n)^2}{2\sigma_n^2}} dt}{\int_{-1}^{1} e^{\frac{(x-\mu_n)^2}{2\sigma_n^2}} dt} \\ &= \frac{erf(\frac{(1-\mu_n)}{\sqrt{2}\sigma_n}) - erf(\frac{T-\mu_n}{\sqrt{2}\sigma_n})}{erf(\frac{(1+\mu_n)}{\sqrt{2}\sigma_n}) + erf(\frac{1-\mu_n}{\sqrt{2}\sigma_n})}. \end{aligned} \tag{18}$$

False negative and false positive numerical results for different threshold values were obtained in our experiments.

3.2.2 Analysis of Denoising-based Prediction with Different Noise Variance

For sparse code shrinkage [9] or wavelet shrinkage [6], the variance of a noise distribution, σ (relevant to the denoising capability), should be determined in advance in order to separate the original image, $\mathbf{X}$, from its embedded noise, $\mathbf{N}$. It should be noted that the value of σ is hard to predict but definitely affects the final reconstruction result. Fortunately, the major concern here is not the original image. What we are concerned about is the detected correlation values. Therefore, it is sufficient if the watermark extracted from the estimated host image is highly correlated with the hidden watermark.

In the following, we shall evaluate the performance of the proposed system when noises with different variance (σ) values are used. Recall that $w_{s,o}(x,y)/w_{s,o}^m(x,y)$ is the original/modulated wavelet coefficient of $\mathbf{X}/\mathbf{X}^m$ at scale s, orientation o, and position (x,y). The attacked wavelet coefficient is denoted as $w_{s,o}^a(x,y)$ with respect to $\tilde{\mathbf{X}}$. After conducting sparse code shrinkage-based denoising on $\tilde{\mathbf{X}}$, the estimated original wavelet coefficient $\bar{w}_{s,o}(x,y)$ satisfies $|\bar{w}_{s,o}(x,y)| < |w_{s,o}^a(x,y)|$ because shrinkage (i.e., soft thresholding) is an operation which gradually reduces the magnitude of a coefficient. Now, we quantitatively analyze the relationship between the SCS-based denoising process and the positive modulation process (Eq. (13)) as follows. According to the function of positive modulation, we know that $|w^m_{s,o}(x,y)| > |w_{s,o}(x,y)|$. When attacks are encountered, we may have three possible situations: (P1) $|w_{s,o}(x,y)| < |w^m_{s,o}(x,y)| < |w^a_{s,o}(x,y)|$; (P2) $|w_{s,o}(x,y)| < |w^a_{s,o}(x,y)| < |w^m_{s,o}(x,y)|$; and (P3) $|w^a_{s,o}(x,y)| < |w_{s,o}(x,y)| < |w^m_{s,o}(x,y)|$. To simplify the analysis, we assume that Eq. (11) holds. If Eq. (11) does not hold, then either (i) $w^m_{s,o}(x,y)$ is small or (ii) the behavior caused by attacks is extremely different from that caused by the embedding process and is, thus, difficult to predict. With this basic assumption, the extracted watermark value $\tilde{k}(i)$ derived from $w^a_{s,o}(x,y) - \bar{w}_{s,o}(x,y)$ satisfies

$$sign(\tilde{k}(i)) = sign(w^a_{s,o}(x,y) - \tilde{w}_{s,o}(x,y)) = sign(w^a_{s,o}(x,y)). \tag{19}$$

Similarly, the hidden watermark value $k(i)$ satisfies

$$sign(k(i)) = sign(w^m_{s,o}(x,y) - w_{s,o}(x,y)). \tag{20}$$

Under situation (P1) and after applying sparse code shrinkage with different values of σ, we can get

$$|\bar{w}_{s,o}(x,y)| < |w^m_{s,o}(x,y)| < |w^a_{s,o}(x,y)| \tag{21}$$

when σ is large or

$$|w^m_{s,o}(x,y)| < |\bar{w}_{s,o}(x,y)| < |w^a_{s,o}(x,y)| \tag{22}$$

when σ is small. From Eq. (21) and Eq. (22), we know that the extracted watermark will have the same sign as the hidden watermark. It is intuitive that preservation of the same sign between the value of a hidden watermark and that of an extracted watermark will be beneficial for deriving a higher correlation value. Under the conditions that (P2) is valid and that sparse code shrinkage has been executed, we can get

$$|\bar{w}_{s,o}(x,y)| < |w^a_{s,o}(x,y)| < |w^m_{s,o}(x,y)| \tag{23}$$

whether σ is small or large. Again, Eq. (23) tends to help increase the correlation value, which is the same as in the case of (P1). Similarly, if the situation is (P3) and sparse code shrinkage has been executed, then we have

$$|\bar{w}_{s,o}(x,y)| < |w^a_{s,o}(x,y)| < |w^m_{s,o}(x,y)| \tag{24}$$

whether σ is small or large. Once again, Eq. (24) will help increase the correlation value, which is the same as in the cases of (P1) and (P2). From the above analysis, we find that different σ values will not affect the correlation value significantly because the polarity of the value of an extracted watermark can always be kept the same as that of the original watermark.

3.2.3 Resistance to Denoising Attacks [12, 24]

From the above analysis, we know that the predicted watermark is indeed very similar to the hidden one. Recently, Voloshynovskiy *et al.* [24] have presented a "denoising and perceptual remodulation attack" which is created by first predicting the hidden watermark using some denoising techniques and then removing the predicted watermark from a watermarked image by means of perceptual remodulation. Kutter *et al.* [12] also used denoising techniques to estimate a watermark. In contrast to Voloshynovskiy *et al.*'s work [24], Kutter *et al.* [12] added the estimated watermark into a non-watermarked image to create a false alarm situation. This kind of attack is a so-called "copy attack" and can be used to challenge the concept of watermarking. From the above two works, we know that a watermark can be predicted by means of denoising and then used to create either a miss detection [24] or false alarm [12] situation. One may ask: "Does *successful* prediction of a watermark also imply that watermark removal can be done successfully?" Our answer is *NO*. We will explain why such an attack cannot successfully destroy a watermark embedded using our method.

Resistance to the Denoising and Perceptual Remodulation Attack

First, we will examine "denoising and remodulation attacks" [24]. Let $\mathbf{X}^a$ be an attacked image which is obtained by applying a denoising operation to a watermarked image $\mathbf{X}^m$. Suppose the denoising operation is a technique such as low-pass filtering, median filtering, Wiener filtering, or shrinkage-based denoising [6, 9]. After applying the denoising operation, we will have either $|w^a_{s,o}(x,y)| < |w^m_{s,o}(x,y)|$ or $|w^a_{s,o}(x,y)| \geq |w^m_{s,o}(x,y)|$. In fact, most coefficients will be *gradually* reduced in magnitude during denoising except when some non-shrinkage-based denoising techniques (like low-pass filtering) are used. Therefore, $|w^a_{s,o}(x,y)| < |w^m_{s,o}(x,y)|$ holds for most coefficients. In our scheme, a hidden watermark is detected in an attacked image $\mathbf{X}^a$ by means of a shrinkage-based denoising operation. Therefore, the coefficients of the estimated original image $\tilde{\mathbf{X}}$ and those of the attacked image $\mathbf{X}^a$ should satisfy the following inequality:

$$|\tilde{w}_{s,o}(x,y)| < |w^a_{s,o}(x,y)|. \tag{25}$$

From the above analysis, we have

$$|\tilde{w}_{s,o}(x,y)| < |w^a_{s,o}(x,y)| < |w^m_{s,o}(x,y)|. \tag{26}$$

In the proposed scheme, positive modulation is applied to the original image. Therefore, we can obtain that $sign(w^m_{s,o}(x,y) - w_{s,o}(x,y))$ is equal to $sign(w^a_{s,o}(x,y) - \tilde{w}_{s,o}(x,y))$. This means that the overall correlation value will increase. From the above analysis, we conclude that a watermark embedded by our scheme is hard to remove using a shrinkage-based denoising algorithm.

Resistance to the Copy Attack

Next, we will examine the effect caused by the "copy attack" [12] on our scheme. Let $w^1_{s,o}(x,y)$ be the wavelet coefficient of an image $\mathbf{X}^1$ belonging to us, and let $w^2_{s,o}(x,y)$ be the wavelet coefficient of an image $\mathbf{X}^2$ belonging to someone else. Let the modulated, attacked, and denoised versions of $w^1_{s,o}(x,y)$ be denoted as $w^{1m}_{s,o}(x,y)$, $w^{1a}_{s,o}(x,y)$, and $\tilde{w}^1_{s,o}(x,y)$, respectively. Furthermore, let the hidden watermark be denoted as $\mathbf{n}^1$. Suppose a denoising technique such as Wiener filtering [13] or sparse code shrinkage [9] is applied to $\mathbf{X}^1$; the predicted watermark $\tilde{\mathbf{n}}^1$ will have the following value:

$$\tilde{k}^1(i) = w^{1m}_{s,o}(x,y) - \tilde{w}^1_{s,o}(x,y), \tag{27}$$

where $1 \leq i \leq ||\mathbf{K}||$. The predicted watermark value $\tilde{k}^1(i)$ is then added to the non-watermarked image $\mathbf{X}^2$ as

$$w^{2a}_{s,o}(x,y) = w^2_{s,o}(x,y) + \tilde{k}^1(i) \tag{28}$$

to create a counterfeit image $\mathbf{X}^{2a}$ with the wavelet coefficients $w^{2a}_{s,o}(x,y)$. Under these circumstances, we can check to see if a watermark retrieved from the counterfeit image is similar to the hidden one, i.e., $\mathbf{n}^1$. Using the proposed method, the watermark-free counterfeit image can be estimated by $\tilde{w}^2_{s,o}(x,y)$, where $|\tilde{w}^2_{s,o}(x,y)| < |w^{2a}_{s,o}(x,y)|$. As a consequence, the value of the predicted watermark $\tilde{\mathbf{n}}^2$ which can be calculated from $\mathbf{X}^{2a}$ is

$$\tilde{k}^2(i) = w^{2a}_{s,o}(x,y) - \tilde{w}^2_{s,o}(x,y) = w^2_{s,o}(x,y) + \tilde{k}^1(i) - \tilde{w}^2_{s,o}(x,y). \tag{29}$$

Due to the gradual change caused by shrinkage-based denoising, we can guarantee that

$$sign(\tilde{k}^2(i)) = sign(w^{2a}_{s,o}(x,y)).$$

Because $\tilde{k}^1(i)$ cannot significantly affect $w^2_{s,o}(x,y)$ from the viewpoint of transparency, we are assured that

$$sign(\tilde{k}^2(i)) = sign(w^2_{s,o}(x,y)). \tag{30}$$

On the other hand, since the hidden watermark is designed to have the same sign as its corresponding wavelet coefficient $w^1_{s,o}(x,y)$, we have

$$sign(k^1(i)) = sign(w^1_{s,o}(x,y)). \tag{31}$$

From Eqs. (30) and (31), we find, in summary, that $sign(k^1(i) \times \tilde{k}^2(i)) = sign(w^1_{s,o}(x,y) \times w^2_{s,o}(x,y))$. The above conclusion indicates that the correlation value between $k^1(i)$ and $\tilde{k}^2(i)$ is directly related by the signs of their corresponding wavelet coefficients. Because the property of a non-watermarked image is random in nature, it can be expected that the correlation value between the retrieved watermark $\tilde{\mathbf{n}}^2$ and the hidden one $\mathbf{n}^1$ will be close to zero. This means that the proposed denoising-based oblivious watermarking method (positive modulation incorporated with shrinkage-based watermark prediction) is able to resist a "copy attack" [12].

3.3 Relationship with the Concept of Watermarking as Communications with Side Information

In [4], Cox *et al.* proposed a new concept which views watermarking as communications with side information. In their scheme, the embedded signal $\mathbf{S}$, which is composed of an extracted signal $\mathbf{V}$ and a watermark $\mathbf{K}$, is perceptually similar to the extracted signal to achieve fidelity and is highly correlated with the hidden watermark $\mathbf{K}$ to achieve robustness. In general, $\mathbf{S}$ can be obtained as a combination of $\mathbf{V}$ and $\mathbf{K}$ by a mixing function f, i.e.,

$$\mathbf{S} = f(\mathbf{V}, \mathbf{K}). \tag{32}$$

A sub-optimal way of computing $\mathbf{S}$ is defined as

$$\mathbf{S} = \mathbf{V} + \omega \cdot \mathbf{K}, \tag{33}$$

where ω is a weight. Recently, four different embedding strategies (including the above one) have been proposed as "informed embedders" [17]. Their performance was compared with that of blind embedding and it was found that informed embedding is better. If our watermarking scheme is interpreted as communications with side information, then we can derive the following result:

$$sign(\mathbf{S}) = sign(\mathbf{V}) = sign(\mathbf{K}). \tag{34}$$

This is because our scheme attempts to keep the signs of watermark values unchanged. In this paper, robustness can be guaranteed if the signs of watermark values remain unchanged after attacks, i.e., $sign(\mathbf{S}) = sign(\mathbf{K})$ holds. Under the assumption that an attacked image will not be perceptually different from the original one (Eq. (11)), $sign(\mathbf{S}) = sign(\mathbf{V})$ should hold. Based on the above, Eq. (34) can be derived. Therefore, the hiding strategy should be designed so as to satisfy $sign(\mathbf{V}) = sign(\mathbf{K})$. This design has been realized by means of positive modulation, as expressed in Eq. (13).

4 Experimental Results

Five standard images of size 256×256 were used as the host images in our experiments. Using our watermarking scheme, we set the length of a hidden watermark as 16128. After watermarking was applied, the PSNR values of the five watermarked image were between 41 and 42 dB, and no perceptual distortion could be observed. 21 commonly used attacks were used to test the robustness of our method. These attacks included (1) blurring; (2) median filtering; (3) Wiener filtering; (4) rescaling; (5) histogram equalization; (6) sharpening; (7) and (8) Gaussian noise addition with different variance values; (9) and (10) uniform noise addition with different variance values; (11) mosaic effects; (12) texturizing; (13) shading; (14) and (15) *JPEG* compression with quality factors of 10% and 5%; (16) and (17) *SPIHT* compression with ratios of $16:1$ and $32:1$; (18) StirMark [20]; (19) dithering; (20) wavelet shrinkage-based denoising [6]; (21) sparse code shrinkage-based denoising [9]. Therefore, there were in total 110 attacked images (including five watermarked images). Among them, the original and the watermarked Barbara images are, respectively, shown in Fig. 1(a) and Fig. 1(b). The two Barbara images which were attacked, respectively, by means of Gaussian noise adding and shading are shown in Figs. 1(c) and (d). Three watermark prediction techniques, wavelet shrinkage-based denoising [6], sparse code shrinkage-based denoising [9], and Wiener filtering [13], were compared in terms of robustness. The comparison results based on the Barbara image are shown in Fig. 1(e). From Fig. 1(e), it can be found that the results obtained by applying Wiener filtering was the worst since prediction (denoising) in this case is not consistent with our modulation operation. We also found that none of the three denoising techniques could correctly predict the hidden watermark from an attacked image with the shading effect (13-th attack). The reason why the shading effect attack could succeed was that the signs of most of the chosen coefficients changed. As a result, the predicted watermark values had signs which were different from their original ones. As we have noted with respect to Eq. (11), these sign changes violate our basic assumption and, thus, degrade the correlation value. Fig. 1(f) shows the result of the uniqueness test when the famous StirMark attack was applied.

In the second group of experiments, we applied $SPIHT$ compression with different compression ratios to see how the correlation value was affected. Fig. 2 shows a curve which reflects the change of the detector response under different compression ratios. It is apparent that when the ratio was small, its corresponding detector response was large. When the compression ratio reached $128:1$, the corresponding detector response dropped to 0.2.

In the third group of experiments, we obtain false positive and false negative numerical results. In Fig. 3(a), we compare the detection results obtained by applying 21 attacks to the five host images.

We find that the five curves are very consistent. A Gaussian distribution was used to approximate these 110 detection results, as shown in Fig. 3(b). In addition, the detection results obtained from 90 non-watermarked images were also approximated by means of another Gaussian distribution, as shown on the left hand side of Fig. 3(c). The distribution shown on the right hand side of Fig. 3(c) was a redrawn version of Fig. 3(b). It is clear that the distribution formed by the 90 non-watermarked images was a sharp peak clustered around a detection value close to zero. On the other hand, the distribution formed by the 5 watermarked images and 105 attacked images was an obtuse curve centered at a detection value close to 0.5. According to the results of our experiments, the mean and standard deviation of the distribution formed by the 90 watermarked but non-attacked images were 0.94 and 0.04, respectively. On the other hand, the mean and standard deviation of the distribution formed by the 90 non-watermarked images were 0.00 and 0.01, respectively. Based on Eqs. (17) and (18), a threshold could be easily determined to obtain that both the false negative and the false positive were negligibly small under a non-attack situation. However, when attacks were imposed, the mean and standard deviation of the distribution formed by the 110 attacked images were 0.54 and 0.24, respectively. Under these circumstances, both false negative and false positive were expected to increase no matter what the threshold T was. In Table 1, we show the false negative and the false positive probabilities corresponding to different threshold values.

Finally, we conducted experiments to see how a “denoising and remodulation attack” [24] and a “copy attack” [12] would affect the proposed scheme. First, the hidden watermark was predicted from the watermarked image shown in Fig. 1(b) using Wiener filtering [13]. The predicted watermark was shown in Fig. 4(a). As for the “denoising and remodulation attack,” the predicted watermark was subtracted from the watermarked image to which it belonged (Fig. 1(b)). Since our objective was to demonstrate how to remove the predicted watermark, the transparency issue was not a major concern. Therefore, the predicted watermark was directly subtracted from Fig. 1(b), and the de-watermarked image is shown in Fig. 4(b). In addition, the predicted watermark was also triplicated and then subtracted to yield a de-watermarked image, as shown in Fig. 4(c). As expected, Fig. 4(c) is less transparent than Fig. 4(b). However, the detection results obtained from Figs. 4(b) and (c) show that the hidden watermark still survived with a high correlation value (≈ 0.79). This implies that the proposed scheme is insensitive to the weight added to the predicted watermark which is to be removed. This phenomenon clearly indicates that our scheme is able to preserve the signs of the watermark values. In addition, the predicted watermarks with different weights were added to the non-watermarked “Lenna” image, as shown in Figs. 4(d) and (e), to examine the effect caused by a "copy attack." Similarly, the detection results reveal that no watermark was detected when our scheme was applied (the detection values were close to zero). That is, the false positive problem did not occur.

5 Conclusion

In this paper, a novel watermarking approach, called the "non-blind" embedder, has been applied by exploiting the available information of denoising-based watermark prediction. We have found that the information obtained using shrinkage-based denoising (soft-thresholding) techniques is easy to control, and, that denoising itself is, in fact, a solution for oblivious watermark detection. The knowledge at the detector side can then be utilized to design a "non-blind" embedder, which is extremely different from the common blind embedder. On the other hand, the predicted watermark can be purposely used to remove a hidden watermark or to confuse judgement about legal ownership. Therefore, we have conducted analysis to confirm that our method indeed can resist the "denoising and remodulation attack" and the "copy attack." The performance of our scheme, composed of a non-blind embedder and a blind detector, has also been analyzed regarding false negative and false positive probabilities.

At the present, it still is not possible for a watermarking scheme to resist all attacks because attackers are always smarter and one step ahead. Therefore, our first future work will focus on the problem of geometric attack resistance [14, 19], which has not been treated in this paper. In addition, we simply use the watermark with a 1-bit payload to indicate its presence or absence. Longer payload [2] should be implemented in order to provide richer information about the owner's affiliation. Finally, the public-key detection [7] and ownership deadlock problems [5] should also be studied in order to obtain a complete, practical watermarking system.

References

[1] M. Barni, F. Bartolini, V. Cappellini, and A. Piva, "Copyright Protection of Digital Images by Embedded Unperceivable Marks", *Image and Vision Computing*, Vol. 16, pp. 897-906, 1998.

[2] S. Baudry, P. Nguyen, and H. Maitre, "Channel Coding in Video Watermarking: Use of Soft Decoding to Improve the Watermark Retrieval", *Proc. IEEE Int. Conf. on Image Processing*, Vancouver, Canada, Vol. III, pp. 21-24, 2000.

[3] P. Comon, "Independent Component Analysis – a new concept", *Signal Processing*, Vol. 36, pp. 287-314, 1994.

[4] I. J. Cox, M. L. Miller, and A. McKellips, "Watermarking as Communications with Side Information", *Proc. of the IEEE*, Vol. 87, No. 7, pp. 1127-1141, 1999.

[5] S. Craver, N. Memon, B.-L. Yeo, and M. M. Yeung, "Resolving Rightful Ownerships with Invisible Watermarking Techniques: Limitations, Attacks, and Implications", *IEEE Journal on Selected Areas in Communications*, Vol. 16, pp. 573-586, 1998.

[6] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian and D. Picard, "Wavelet Shrinkage: Asymptopia?", *J. Royal Stat. Soc. ser B*, Vol. 57, pp. 301-337, 1995.

[7] T. Furon and F. P. Duhamel, "Robustness of An Asymmetric Watermarking Method", *Proc. IEEE Int. Conf. on Image Processing*, Vancouver, Canada, Vol. III, pp. 21-24, 2000.

[8] F. Hartung and M. Kutter, "Multimedia Watermarking Techniques", *Proceedings of the IEEE*, Vol. 87, pp. 1079-1107, 1999.

[9] A. Hyvarinen, "Sparse Code Shrinkage: Denoising of nongaussian data by maximum likelihood estimation", *Neural Computation*, Vol. 11, pp. 1739-1768, 1999.

[10] D. Kundur and D. Hatzinakos, "Digital Watermarking for TellTale Tamper Proofing and Authentication", *Procceedings of the IEEE*, Vol. 87, pp. 1167-1180, 1999.

[11] M. Kutter, F. Jordan, and F. Bossen, "Digital Signature of Color Images using Amplitude Modulation", *Journal of Electronic Imaging*, Vol. 7, pp. 326-332, 1998.

[12] M. Kutter, S. Voloshynovskiy, and A. Herrigel, "The Watermark Copy Attack", *Proc. SPIE: Security and Watermarking of Multimedia Contents II*, Vol. 3971, San Jose, CA, USA, 2000.

[13] J. S. Lee, “Digital Image Enhancement and Noise Filtering by Use of Local Statistics”, *IEEE Trans. on Pattern Anal. and Machine Intell.*, Vol. 2, No. 2, pp. 165-168, 1980.

[14] C. Y. Lin, M. Wu, J. A. Bloom, I. J. Cox, M. L. Miller, and Lui, “Rotation, Scale, and Translation Resilient Public Watermarking for Images”, *Security and Watermarking of Multimedia Contents II, Proceedings of SPIE*, Vol. 3971, pp. 90-98, 2000.

[15] C. S. Lu, S. K. Huang, C. J. Sze, and H. Y. Mark Liao, “Cocktail Watermarking for Digital Image Protection”, *IEEE Trans. on Multimedia*, Vol. 2, No. 4, pp. 209-224, 2000.

[16] C. S. Lu and H. Y. Mark Liao, “Oblivious Cocktail Watermarking by Sparse Code Shrinkage: A Regional- and Global-based Approach”, *Proc. 7th IEEE Int. Conf. on Image Processing: special session on second generation watermarking methods*, Vancouver, Canada, Vol. III, pp. 13-16, 2000.

[17] M. L. Miller, I. J. Cox, and J. A. Bloom, “Informed Embedder: Exploiting Image and Detector Information During Watermark Insertion”, *Proc. 7th IEEE Int. Conf. on Image Processing: special session on second generation watermarking methods*, Vancouver, Canada, Vol. III, pp. 1-4, 2000.

[18] F. Mintzer and G. W. Braudaway, “If one watermark is good, are more better?”, *Proc. IEEE Int. Conf. on Acoustic, Speech, and Signal Processing*, pp. 2067-2070, 1999.

[19] S. Pereira and T. Pun, “Fast Robust Template Matching for Affine Resistant Image Watermarks”, *3rd Int. Workshop on Information Hiding*, LNCS 1768, pp. 199-210, Sept. 29-Oct. 1, 1999.

[20] F. Petitcolas, R. J. Anderson, and M. G. Kuhn, “Attacks on Copyright Marking Systems”, *Second Workshop on Information Hiding*, USA, pp. 218-238, 1998.

[21] F. Petitcolas, R. J. Anderson, and M. G. Kuhn, “Information Hiding: A Survey”, *Proc. of the IEEE*, Vol. 87, pp. 1062-1078, 1999.

[22] P. C. Su, C.-C. Jay Kuo, and H. J. Wang, “Blind Digital Watermarking for Cartoon and Map Images”, SPIE International Symposium Electronic Imaging, 1999.

[23] S. Voloshynovskiy, A. Herrigel, N. Baumgartner, and T. Pun, “A Stochastic Approach to Content Adaptive Digital Image Watermarking”, *3rd Int. Workshop on Information Hiding*, Dresden, Germany, LNCS 1768, pp. 211-236, Sept. 29-Oct. 1, 1999.

[24] S. Voloshynovskiy, S. Pereira, A. Herrigel, N. Baumgartner, and T. Pun, “Generalized Watermarking Attack Based on Watermark Estimation and Perceptual Remodulation”, *Proc. SPIE: Security and Watermarking of Multimedia Contents II*, Vol. 3971, San Jose, CA, USA, 2000.

[25] 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.

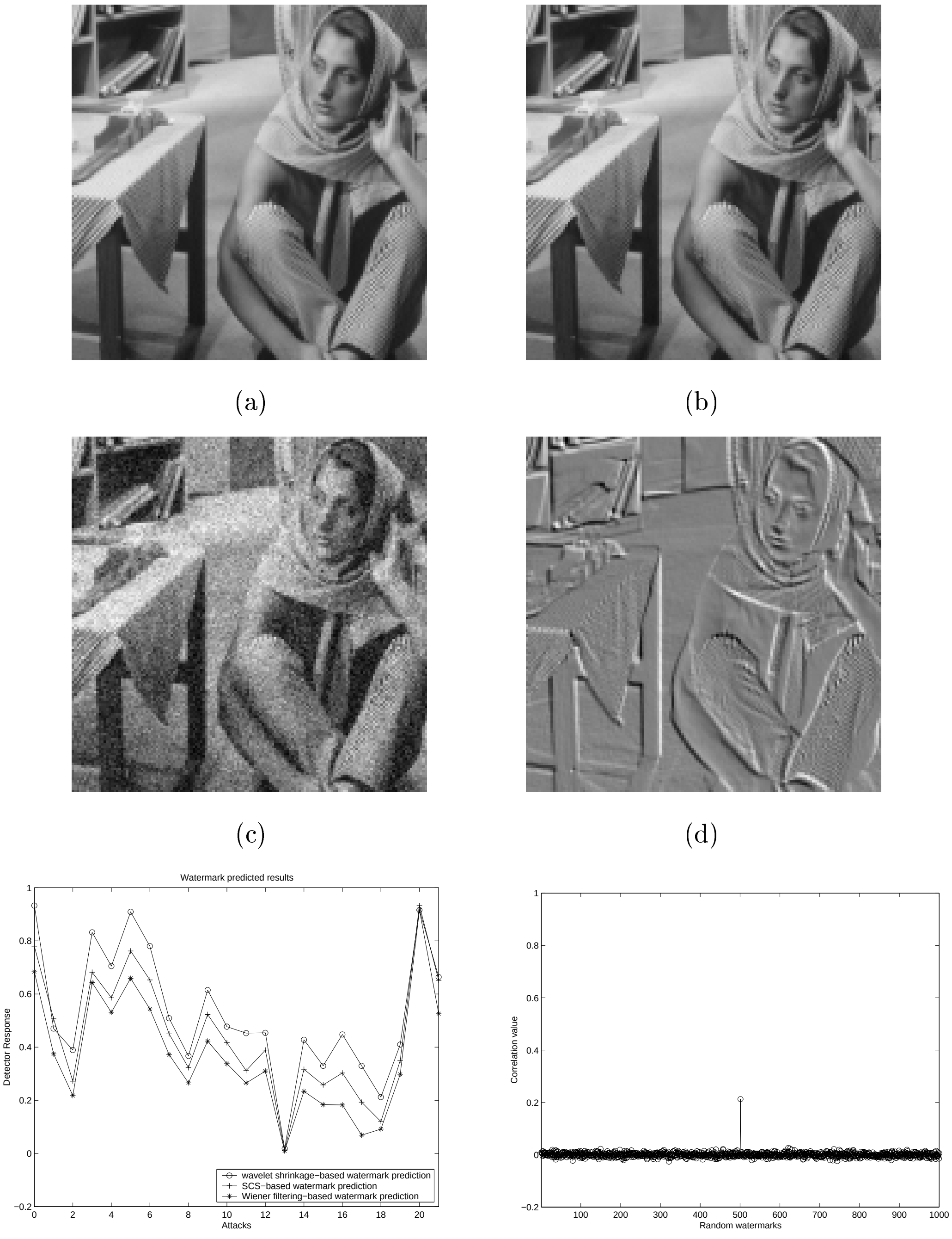
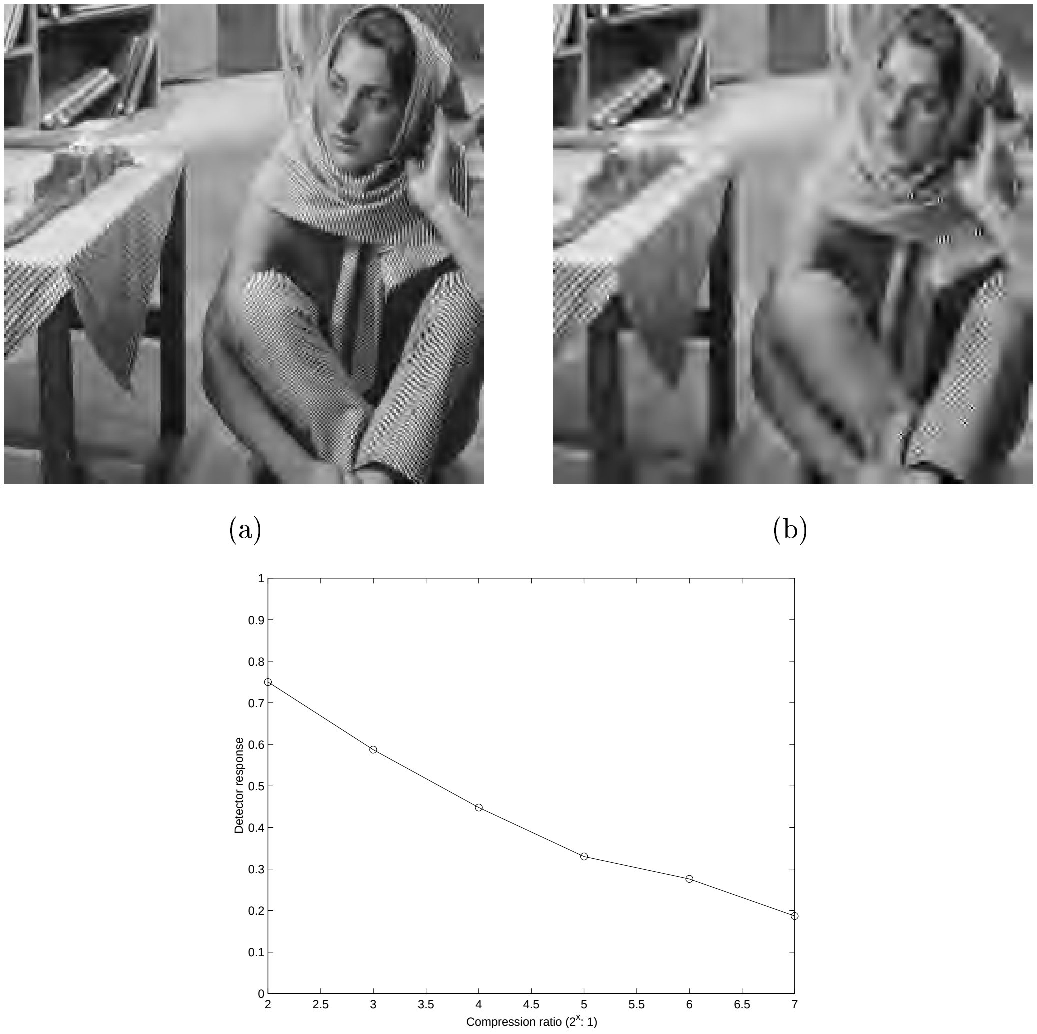

(a)
(b)
(c)
(d)
(e)
(f)

Figure 1: Robustness test of the proposed scheme (a non-blinder embedder and a blind detector): (a) host image; (b) watermarked image; (c) Gaussian noise added image; (d) attacked image with the shading effect; (e) comparison of detected watermarks, respectively, predicted using wavelet shrinkage, SCS, and Wiener filtering. The first response was obtained without applying any attack, and the remaining results were obtained by applying the 21 attacks described above (0-th attack denotes attack-free); (f) unique watermark test for the StirMark attack.



(a)

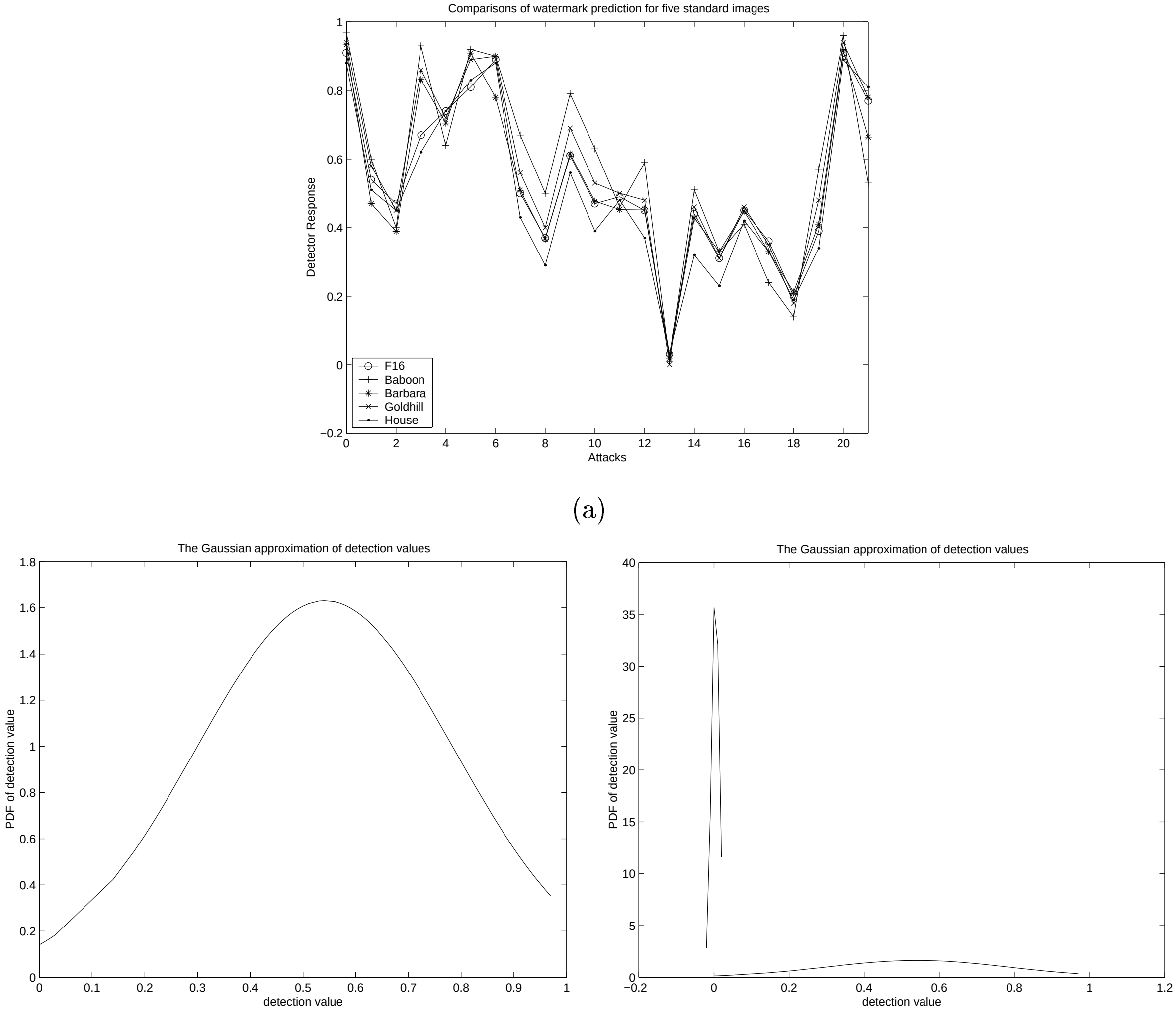
(b)

(c)

Figure 2: The proposed scheme under the $SPIHT$ compression attack: (a) the $SPIHT$ compressed image with a ratio of $16 : 1$; (b) the $SPIHT$ compressed image with a ratio of $64 : 1$; (c) the detection results obtained under $SPIHT$ compression at different ratios ranging from $2^2 : 1 \sim 2^7 : 1$.

Table 1: **Probabilities of false negative (p_{fn}) and false positive (p_{fp}) corresponding to different thresholds (T).**

Threshold (T)	p_{fn}	p_{fp}
0.0200	1.73×10^{-2}	2.25×10^{-2}
0.0225	1.77×10^{-2}	1.20×10^{-2}
0.0250	2.09×10^{-2}	6.10×10^{-3}

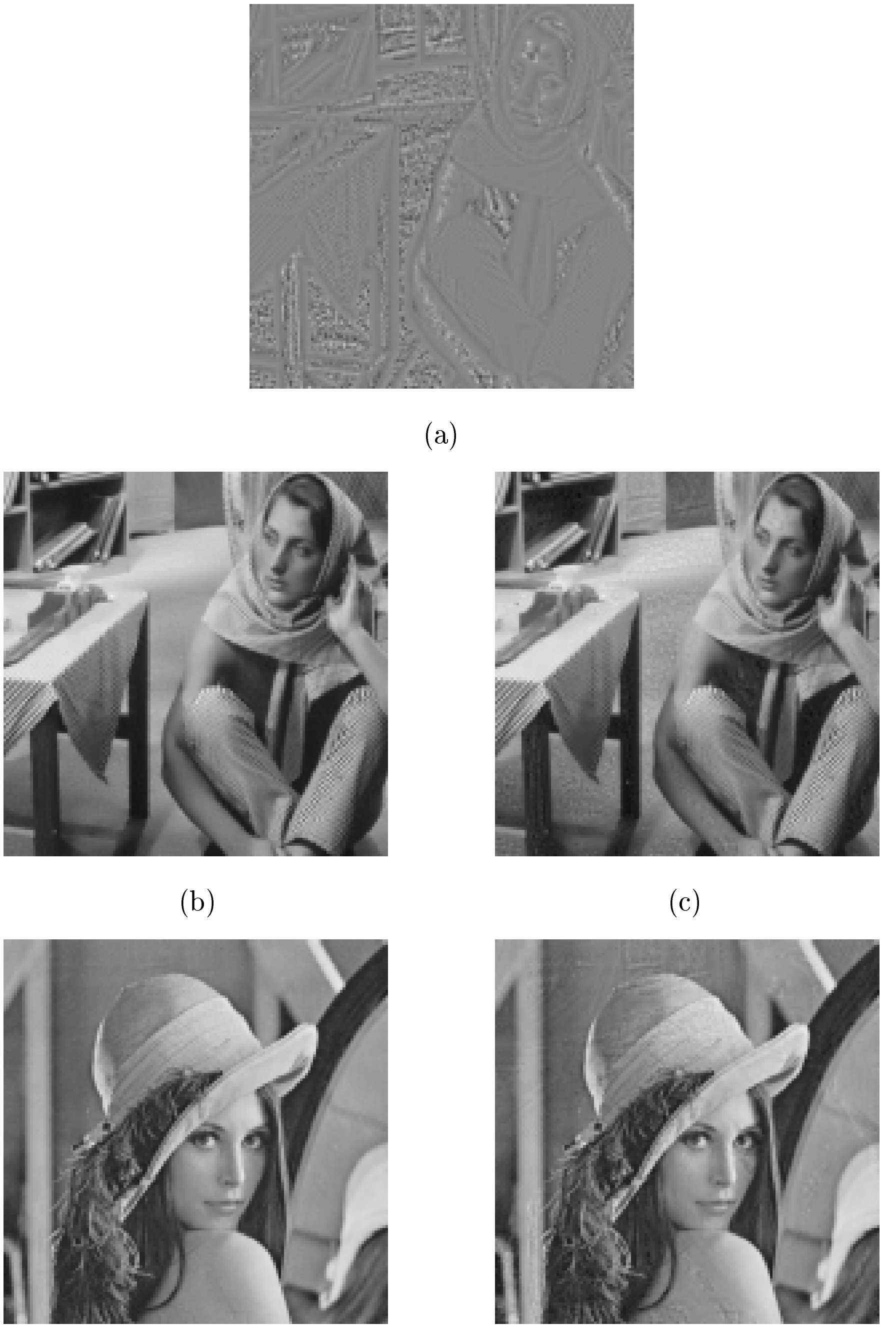


(a)

(b)

(c)

Figure 3: Analysis of false positive and false negative: (a) comparisons of the detection results for five images under 21 attacks; (b) the distribution of 110 watermarked/attacked images; (c) the distribution on the right is the rescaled version of (b) but the one on the left is the distribution formed by 90 non-watermarked images.



(a)

(b)

(c)

(d)

(e)

Figure 4: The effects of the "denoising and remodulation attack" [24] and the "copy attack" [12]: (a) the predicted watermark of Fig. 1(b) using the adaptive Wiener filter [13]; (b)$\sim$(c) the watermarked images with the predicted watermark (a) removed using different weights; (d)$\sim$(e) the predicted watermark (a) added into a non-watermarked image using different weights.