A Hypothesis Testing Approach to Semifragile Watermark-Based Authentication

Chuhong Fei, *Member, IEEE*, Raymond H. Kwong, *Senior Member, IEEE*, and Deepa Kundur, *Senior Member, IEEE*

***Abstract*—This paper studies the problem of achieving watermark semifragility in watermark-based authentication systems through a composite hypothesis testing approach. Embedding a semifragile watermark serves to distinguish legitimate distortions caused by signal-processing manipulations from illegitimate ones caused by malicious tampering. This leads us to consider authentication verification as a composite hypothesis testing problem with the watermark as side information. Based on the hypothesis testing model, we investigate effective embedding strategies to assist the watermark verifier to make correct decisions. Our results demonstrate that quantization-based watermarking is more appropriate than spread-spectrum-based methods to achieve the semifragility tradeoff between two error probabilities. This observation is confirmed by a case study of an additive Gaussian white noise channel with a Gaussian source using two figures of merit: 1) relative entropy of the two hypothesis distributions and 2) the receiver operating characteristic. Finally, we focus on common signal-processing distortions, such as JPEG compression and image filtering, and investigate the discrimination statistic and optimal decision regions to distinguish legitimate and illegitimate distortions. The results of this paper show that our approach provides insights for authentication watermarking and allows for better control of semifragility in specific applications.**

***Index Terms*—Digital watermarking, hypothesis testing, multimedia authentication, semifragile.**

I. Introduction

MANY watermark-based multimedia authentication systems have been proposed in the last few years for ensuring the integrity and origin of multimedia data such as images. These systems fall into two broad categories: 1) fragile and 2) semifragile. Fragile authentication watermarking systems [1]–[3] often detect any modifications to the marked signal in a similar way to traditional digital signatures. Semifragile systems [4]–[6], however, are designed to detect content-changing modifications, but tolerate certain kinds of content-preserving processing, such as high-quality compression. The primary advantage of employing semifragile watermarking over digital signature and fragile watermarking technologies is that there is greater potential in characterizing the tamper distortion. Though not used for authentication applications but copyright protection, it is worth mentioning another type of watermarking—robust watermarking—which is designed to detect or extract the embedded watermark itself even under modifications of the marked signal.

Many semifragile watermarking systems have been proposed in the literature [4]–[19]. One of the first approaches to semifragile watermarking, called telltale tamper proofing, was proposed by Kundur and Hatzinakos [4] to determine the extent of modification in the spatial and frequency domains of a signal by using a statistics-based tamper assessment function. Another influential semifragile system is the self-authentication-and-recovery image (SARI) method developed by Lin and Chang [5], [6] in which a semifragile signature is designed to survive JPEG compression up to a certain level. To distinguish JPEG compression from other malicious manipulations, two invariant properties of quantization are used. The first property shows that a prequantized coefficient can be exactly reconstructed after subsequent JPEG compression if the original quantization step is larger than the one used for JPEG compression; this property is used for watermark embedding to guarantee robustness up to a certain level of JPEG compression. The second property involves an invariant relationship between a pair of coefficients before and after JPEG compression, and is used to generate the signature. Although the SARI system works well under JPEG compression, its ad-hoc design using the unique properties of JPEG quantization limits its portability to different applications. Other previously proposed semifragile watermarking methods [10], [12], [14] are achieved by carefully "scaling" a robust watermark so that it is likely to be destroyed if the distortion exceeds a particular level. Lin *et al.* [12] propose a semifragile watermarking technique based on extending a simple spread-spectrum watermarking method with a modified detector. Yu *et al.* [10] use a mean-quantization-based fragile watermark to detect malicious tampering while tolerating some incidental distortions. Most recently, new semifragile watermarking schemes have been proposed by using random bios and nonuniform quantization [17], integer wavelet transform [18], and spatiotemporal chaos [19].

Overall, semifragile multimedia authentication systems are designed with two objectives: 1) to authenticate legitimate changes and 2) to detect content-changing modifications. Thus, possible distortions are often classified into two categories: 1) legitimate and 2) illegitimate changes. In general, modifications which do not alter the "perceptual" content of the multimedia signal are considered to be legitimate. These

Manuscript received October 28, 2008; revised January 07, 2009. First published March 04, 2009; current version published May 15, 2009. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ton Kalker.

C. Fei was with the Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON M5S 3G4, Canada. He is now with the A.U.G. Signals Ltd., Toronto, ON M5H 4E8, Canada (e-mail: fei@control.toronto.edu).

R. H. Kwong is with the Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON M5S 3G4, Canada (e-mail: kwong@control.toronto.edu).

D. Kundur is with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843-3128 USA (e-mail: deepa@ece.tamu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIFS.2009.2015039

typically include minor modifications, such as high-rate JPEG compression and image enhancement filtering. Severe modifications, such as low-rate compression, image blurring filtering, and malicious image object removal or substitution are typically considered illegitimate. When a marked signal undergoes legitimate distortion which does not alter the visual content of the data, the authentication system should indicate that the signal is authentic from the original sender. Conversely, when it undergoes illegitimate tampering, the distorted signal should be rejected as inauthentic. Therefore, a successful multimedia authentication system should be well designed such that it is robust to legitimate distortions but fragile to illegitimate distortions. Note that the legitimate and illegitimate transition region is application dependent and that the semifragile methodology should employ this information during the design phase.

Multimedia authentication watermarking faces two significant challenges. One challenge is that there is typically no clear distinction boundary between legitimate and illegitimate distortions. This intrinsic uncertainty makes semifragile authentication challenging and necessarily ad hoc in most applications. The other major difficulty is the fact that the original host is not available at the receiver side for authentication verification. Therefore, in some contexts, the original host serves as an interference in authentication. In practical applications, the original host generally has a much larger magnitude than the allowed legitimate channel distortions. The unavailability of the original host makes it hard to differentiate legitimate distortions from illegitimate ones. These challenges motivate us to investigate the semifragile nature of multimedia authentication by using the hypothesis testing approach.

Statistical hypothesis testing is the fundamental approach in signal detection theory [20]. It has been employed in robust watermarking to derive various optimal watermark detectors for certain attack models [21], [22] as well as in steganalysis for detecting hidden data [23]. Hypothesis testing is also an important approach for the information-theoretical analysis of message authentication. In [24], message authentication is interpreted as a hypothesis testing problem to analyze lower bounds for authentication attacks, in which the relative entropy is employed to evaluate the ability to differentiate two hypothesis distributions. In this paper, we take a similar approach of hypothesis testing in analyzing semifragile watermarking regarding to two classes of multimedia changes. In our formulation, authentication verification is modeled as a problem of differentiating two classes of changes (i.e., two composite hypotheses) and the role of watermark embedding is modeled as side information to assist the authentication verification procedure. Using the composite hypothesis testing approach, the best authentication verification strategy is derived and relative entropy is also employed in this paper to evaluate the authentication differentiation ability due to the watermark side information. This new composite hypothesis testing approach attempts to address the multimedia authentication challenges in a statistical sense, which enables us to reveal the nature of watermark embedding and provide insightful design for watermark-based authentication systems.

Within this design methodology using hypothesis testing, this paper also investigates a fundamental question as to the type of embedding that works well for authentication. Empirically, it has been found that the quantization-based method is better than the spread-spectrum method for authentication watermarking since many new proposed semifragile systems adopt the quantization-based embedding method [4], [6], [10], [25]. Through the hypothesis testing framework, our analytical results show that the quantization-based embedding method is better than the spread-spectrum method to achieve the tradeoff between two error probabilities. This paper complements our previous paper [26]. In [26], we derived a coding structure for watermark-based authentication and analyzed the security aspects of semifragile systems. In this work, we are more focused on semifragility requirements of authentication watermarking systems. Our goal is to determine the most effective way to embed a watermark for differentiating legitimate and illegitimate multimedia changes.

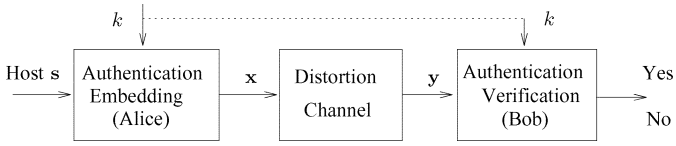


Fig. 1. General authentication watermarking model.

This paper is structured as follows: Section II formulates authentication watermarking as a semifragile hypothesis testing model and identifies the role of watermark embedding. In Section III, for tractability of the solution, a simple case of additive white Gaussian noise channels with a Gaussian source is considered to confirm the analytical result that quantized-based embedding is more effective in authentication watermarking. Section IV focuses on common image-processing distortions, such as JPEG compression and image filtering, and analyzes test statistics for authentication. Finally, conclusions are made in Section V.

II. Semifragile Hypothesis Testing Model

We now describe our hypothesis testing approach, and investigate how semifragility can be characterized by error probabilities arising in a hypothesis testing model. In the context of detection theory, our theoretical analysis attempts to address a fundamental question as to the type of embedding that is better for authentication.

A. Mathematical Model

We consider the general authentication watermarking system shown in Fig. 1 which contains three components: 1) an embedder (Alice), 2) a distortion channel, and 3) the corresponding watermark verifier (Bob). The n-dimensional multimedia signal $\mathbf{s} = (s_1, s_2, \ldots, s_n) \in \mathcal{S}^n \subseteq \mathbb{R}^n$ is the host signal, which is a block of data or transform coefficients from an image, video, audio, or other signal which Alice wants to authenticate. The host signal vector $\mathbf{s}$ takes values from signal space $\mathcal{S}^n$ where the alphabet $\mathcal{S}$ could be $\mathbb{Z}_{256} = \{0, 1, \cdots, 255\}$ for image pixels, or $\{x \in \mathbb{R} | -1024 \leq x \leq 1023\}$ for discrete cosine transform (DCT) coefficients of images. The watermark message k is secret information which is unique to each transmitter. In this model, we assume for simplicity a symmetric key scheme in which Alice's secret key is also available to Bob, but not to the opponent. Secret information shared between the sender and

receiver may include the traditional key in cryptographic techniques as well as the watermark in some semifragile authentication systems. For example, in the Yeung–Mintzer scheme [2], a watermark logo image is shared between the transmitter and the receiver; thus this is also part of the secret message. In the Wong scheme [3], the secret message is an encryption key as well as a binary watermark image. In these two examples, the secret information is long, but we can index it by using a lookup table. Thus, without loss of generality, we suppose k is an integer from an index set $\mathcal{K} = \{1, 2, \cdots, K\}$ where K is the total number of messages.

Alice wants to send the host signal $\mathbf{s}$ to Bob through a public channel. In order for Bob to be ensured that the signal did originate from Alice, Alice authenticates the host source $\mathbf{s}$ with the secret watermark k to produce an authenticated signal $\mathbf{x} \in \mathcal{S}^n$ without introducing perceptible visual distortion. The authentication embedding procedure is described as a function f which takes the host $\mathbf{s}$ and the watermark k as inputs to produce the composite signal $\mathbf{x}$

$$\mathbf{x} = f(\mathbf{s}, k). \tag{1}$$

The signal $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is often referred to as the authenticated signal or the watermarked signal. The authentication distortion or embedding distortion is defined as

$$d(\mathbf{s}, \mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} d(s_i, x_i) \tag{2}$$

where $d : \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}^+$ is a distortion measure, and d is also used to denote the distortion between two vectors. The embedding procedure should not introduce visual artifacts, so the embedding distortion must be below a given maximum allowed value. A common and analytically convenient distortion measure for real-valued signals is the squared difference, i.e., $d(x, y) = (x - y)^2$. Under this distortion measure, the embedding distortion is the mean square error (MSE) defined as $\mathrm{MSE} = (1/n) \sum_{i=1}^{n} (s_i - x_i)^2$ between the host and the composite signal. A commonly used measure of visual distortion for images is the peak signal-to-noise ratio (PSNR), defined as

$$\mathrm{PSNR} = 10 \log_{10} \frac{\max_i s_i^2}{\mathrm{MSE}} = 10 \log_{10} \frac{n \max_i s_i^2}{\sum_{i=1}^{n} (s_i - x_i)^2}. \tag{3}$$

Other practical but sophisticated visual-quality assessment measures for images include Watson's visual model [27], [28] and the structural similarity index [29] which take advantage of known characteristics of the human visual system (HVS).

At the receiver end, knowing Alice' secret key k, Bob attempts to verify whether the received signal $\mathbf{y} \in \mathcal{S}^n$ is an authentic copy from Alice. In authentication systems, authenticity implies that the integrity of the data communicated as well as the origin of the source are verifiable. In other words, if $\mathbf{y}$ is a legitimate copy of the data originating from Alice with watermark k, $\mathbf{y}$ should be accepted by Bob as authentic. In general, the verification procedure is described as a binary function $g(\mathbf{y}, k)$ with $\mathbf{y}$ and k as inputs. The received signal $\mathbf{y}$ is authentic from Alice if $g(\mathbf{y}, k) = 1$, and inauthentic if $g(\mathbf{y}, k) = 0$.

$$\mathbf{s} \xrightarrow{k} \mathbf{x} \xrightarrow[\mathcal{P} \in \mathcal{L}_1]{\mathcal{P} \in \mathcal{L}_0} \mathbf{y} \qquad \mathbf{s} \xrightarrow[k_1]{k_0} \mathbf{x} \xrightarrow{\mathcal{P}} \mathbf{y}$$

(a) (b)

Fig. 2. Distinction between hypothesis testing approaches in semifragile watermarking and robust watermarking. (a) Semifragile authentication watermarking. (b) Robust watermark detection.

B. Authentication Semifragility Verification: Composite Hypothesis Testing Approach

During the transmission in the public channel, the authenticated signal $\mathbf{x}$ may be altered by distortions, as depicted in Fig. 1. Possible modification of the watermarked signal is modelled as a distortion channel $\mathbf{y} = \mathcal{P}(\mathbf{x})$ where $\mathbf{x}$ is the channel input and $\mathbf{y}$ is the channel output. The behavior of a distortion channel $\mathcal{P}$ is characterized by its probability density function (PDF) of $\mathbf{y}$ conditioned on $\mathbf{x}$, denoted by $P(\mathbf{y}|\mathbf{x})$. The random distortion channel includes deterministic modifications as a special case.

We introduce two categories of channels: 1) the legitimate channel set $\mathcal{L}_0$ and 2) the illegitimate channel set $\mathcal{L}_1$. Depending on the application, $\mathcal{L}_0$ may include minor modifications, such as high-rate JPEG compression, image enhancement filtering, and other unobtrusive manipulations. The illegitimate channel set $\mathcal{L}_1$ may include severe modifications, such as low-rate compression, image blurring filtering, and malicious tampering, such as image object removal or substitution. With channel input data $\mathbf{x}$ originating from k, if a channel $\mathcal{P} \in \mathcal{L}_0$, the channel output $\mathbf{y}$ should be considered authentic, while the channel output $\mathbf{y}$ of an illegitimate channel $\mathcal{P} \in \mathcal{L}_1$ should be considered inauthentic.

At the receiver end, the receiver attempts to verify whether $\mathbf{y}$ is a legitimate copy of the data originating from the sender with watermark k. This is equivalent to determining whether $\mathcal{P} \in \mathcal{L}_0$ or $\mathcal{L}_1$ is based on the received signal $\mathbf{y}$ and the sender's watermark k. Semifragile watermark verification can therefore be viewed mathematically as a composite hypothesis testing problem to identify legitimacy of a channel. Two composite hypotheses are present: 1) the null hypothesis H_0 is $\mathcal{L}_0$ and 2) the alternative hypothesis H_1 is $\mathcal{L}_1$, defined as

$$H_0 : \quad \mathbf{y} = \mathcal{P}(\mathbf{x}) \text{ for } \mathcal{P} \in \mathcal{L}_0 \tag{4a}$$

$$H_1 : \quad \mathbf{y} = \mathcal{P}(\mathbf{x}) \text{ for } \mathcal{P} \in \mathcal{L}_1 \tag{4b}$$

that are tested by using the observation $\mathbf{y}$ and the sender's watermark k. Here, the watermark k represents the side information to the receiver and is used to help make an appropriate hypothesis testing decision. This side information helps decision making because it serves to reveal partial information about the signal to authenticate $\mathbf{x} = f(\mathbf{s}, k)$. However, the embedding function f influences the level of side information available at the verification end. This is, in part, due to the fact that the host signal $\mathbf{s}$ represents a form of interference for the channel differentiation.

From the model of (4), we can identify a fundamental difference between hypothesis approaches for semifragile and robust watermarking, which is illustrated in Fig. 2. The purpose of hypothesis testing in semifragile watermarking is to identify the type of the channel by leveraging the watermark characteristics to provide useful side information for detection. The purpose of

hypothesis testing in robust watermarking is to determine which watermark message k_0 or k_1 has been embedded in a host with some channel noise as interference [22]. Due to the diversity of the characteristics of legitimate and illegitimate changes and blind nature of the source signal in authentication watermarking, the hypothesis testing problem in Fig. 2(a) is, in some ways, more challenging than that of robust watermarking. That is also why the watermark k is needed in Fig. 2(a) to assist semifragile hypothesis testing.

1) Semifragility: Two Types of Errors: With respect to the hypothesis testing model, there are two types of authentication errors in semifragile watermarking [26], [30]. Type I error, often called false positive error, or false alarm, results when the distortion channel $\mathcal{P}$ is identified to be in $\mathcal{L}_1$ when it is actually in $\mathcal{L}_0$. This type of authentication error characterizes the robustness of the semifragile authentication system. Let A_n be the decision region in n-dimensional signal space, which the receiver uses to verify authenticity of the received signal $\mathbf{y}$. Type I error probability is given by

$$\alpha_n(\mathcal{P}) = P[\mathbf{y} \notin A_n | \mathcal{P} \in \mathcal{L}_0]. \tag{5}$$

Type II error, often called false negative error, or miss, occurs when $\mathbf{x}$ has been illegitimately tampered but the received signal $\mathbf{y}$ is incorrectly verified by the receiver as authentic. This type of authentication error characterizes the fragility of the semifragile system. Type II error probability is given by

$$\beta_n(\mathcal{P}) = P[\mathbf{y} \in A_n | \mathcal{P} \in \mathcal{L}_1]. \tag{6}$$

In general, two hypotheses $\mathcal{L}_0$ and $\mathcal{L}_1$ are composite, so there are two families of authentication error probabilities for a given decision region. These competing families of error probabilities are sensitive to the selection region A_n, resulting in a natural performance tradeoff. The most attractive decision region gives the best tradeoff between two families of authentication error probabilities under the Neyman–Pearson criterion [31], [32].

2) Common Approach to Composite Hypothesis Testing: The most commonly used approach for composite hypothesis testing is the generalized likelihood-ratio test (GLRT) [30]. In the GLRT approach, the most probable individual hypothesis is used to determine the likelihood of the composite hypothesis. The associated generalized likelihood ratio is defined as the ratio of the maximum value of the likelihood under H_0 to the maximum under H_1. That is

$$\mathrm{GLR} = \frac{\sup_{\mathcal{P}\in\mathcal{L}_0} f(\mathbf{y}|\mathcal{P}, H_0, k)}{\sup_{\mathcal{P}\in\mathcal{L}_1} f(\mathbf{y}|\mathcal{P}, H_1, k)} \tag{7}$$

where $f(\mathbf{y}|\mathcal{P}, H_i, k)$ is the likelihood of the received sequence $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ under the hypothesis H_i with the side information k. For ease of computation, we employ the following generalized log-likelihood ratio normalized by the dimension n:

$$\mathrm{GLLR} = \frac{1}{n}\log \sup_{\mathcal{P}\in\mathcal{L}_0} f(\mathbf{y}|\mathcal{P}, H_0, k) - \frac{1}{n}\log \sup_{\mathcal{P}\in\mathcal{L}_1} f(\mathbf{y}|\mathcal{P}, H_1, k). \tag{8}$$

Hypothesis H_0 is accepted if the aforementioned test statistic is greater than a given threshold T; otherwise, H_1 is accepted. The predefined threshold T is chosen to tradeoff two types of error probabilities.

The presented work utilizes the GLRT approach to develop a practical test statistic to verify the authenticity of a given signal. By introducing this composite hypothesis testing and GLRT framework to semifragile analysis, we are able to investigate the role of watermark embedding in enhancing the ability to differentiate legitimate from illegitimate distortions.

C. Authentication Embedding: Reducing Interference From the Host Signal

The authentication performance criterion and the comparison of watermark embedding strategies are now considered. The performance of two predominant classes of embedding methods is studied: spread-spectrum- and quantization-based approaches. Specifically, we provide an information-theoretic explanation of how quantization-based embedding helps the receiver distinguish the two channel classes. For semifragile authentication, a quantization-based scheme will be shown to allow the receiver to achieve the best tradeoff between Type I and II error probabilities.

Let $p_Y(y)$ and $q_Y(y)$ denote the probability density functions of the two most probable single hypotheses corresponding to legitimate and illegitimate channels, respectively, discussed in relation to (7). A well-known result in hypothesis testing provides a relationship between the error probabilities α and β and the relative entropy $D(p_Y\|q_Y) = \int_y p_Y(y)\log(p_Y(y)/q_Y(y))$ as follows: $\hat{d}(\alpha,\beta) \le D(p_Y|q_Y)$, where $\hat{d}(\alpha,\beta)$ is defined as $\hat{d}(\alpha,\beta) = \alpha\log(\alpha/(1-\beta)) + (1-\alpha)\log((1-\alpha)/\beta)$ [24] [33], [34]. In particular, for $\alpha = 0$, we have $\beta \ge 2^{-D(p_Y\|q_Y)}$. In other words, $D(p_Y\|q_Y)$ characterizes the error exponent of Type II error probabilities for $\alpha = 0$. Although it is mathematically not a true distance measure, the aforementioned relationship suggests using $D(p_Y\|q_Y)$ as a figure of merit to measure the receiver's capability to verify the authenticity of a received signal. Note that the counterpart $D(q_Y\|p_Y)$ is also a measure of figure of merit to evaluate the Type I error rate. Within the context of the semifragile watermarking problem depicted in Fig. 1, where S, K, X, and Y are random variables corresponding to the source, the watermark, the channel input and output, respectively; the associated figure of merit for semifragile verification is $D(p_{Y|K}\|q_{Y|K})$ since the watermark K is known to the receiver.

With the relative entropy $D(p_{Y|K}\|q_{Y|K})$ as the figure of merit, we are able to explain two extreme embedding methods that provide performance bounds for any watermark-based authentication. It is shown in the Appendix that with any method of watermark embedding

$$D(p_Y\|q_Y) \le D(p_{Y|K}\|q_{Y|K}) \le D(p_{Y|X}\|q_{Y|X}). \tag{9}$$

The lower bound $D(p_Y\|q_Y)$ corresponds to the scenario that the hypothesis testing decision is made only based on the received signal, without any help from the watermark. In this scenario, the channel input X itself serves as interference-to-channel distortion hypothesis testing since no information about X is available at the receiver through the watermark K. This lower bound result also justifies our intuition that the side information helps to alleviate interference from the host signal. The upper bound

$D(p_{Y|X}||q_{Y|X})$ corresponds to the nonblind authentication scenario that the host signal S is fully known to the receiver through the watermark K. With the known host signal, authentication errors are merely due to the fuzzy boundary between legitimate and illegitimate distortions and, thus, the receiver can make the best judgement on the legitimacy of a test signal. In practical watermark-based authentication systems, due to the limited length of the watermark, it is impossible to know fully about the host signal S through the watermark K, so this upper bound is not achievable. Nevertheless, the nonblind scenario provides us with an upper bound to evaluate any watermark-based systems.

For general authentication watermarking schemes, the watermark should be embedded so that the figure of merit $D(p_{Y|K}||q_{Y|K})$ is maximized over possible embedding functions. This optimization problem is very complex to solve since the embedding function should also satisfy an embedding distortion constraint. In this paper, we provide an intuitive explanation of how a good embedding function helps channel differentiation. Since the channel input X serves as a form of interference to distortion channel differentiation, the watermark should be embedded to reduce the degree of this interference. It is intuitively straightforward that the more random the signal X is, the more difficult it is to distinguish between two hypothesis probability functions $p_{Y|K}$ and $q_{Y|K}$ [34]. Therefore, one would like to reduce the uncertainty of X conditioned on K. In other words, the conditional entropy $H(X|K)$ should be minimized in order to reduce its interference to distortion channel differentiation.

In quantization-based embedding schemes, the watermark K is embedded in the source S by using a corresponding quantizer, and the embedded signal X is represented as $X = Q_K(S)$, where $Q_K(\cdot)$ is a watermark-related quantization operation. From information theory [34] $H(X|K) = H(Q_K(S)|K) < H(S|K) = H(S)$ since quantization is not a bijective function. Therefore, quantization-based embedding reduces interference from the blind host signal. In general, a larger quantization step will result in less entropy of $X|K$. However, a larger quantization step will also result in larger embedding distortion D. Therefore, there is a tradeoff in determining the quantization step. In practice, a multidimensional lattice quantizer will give less entropy of X than a uniform quantizer for given authentication distortion.

In standard spread-spectrum watermarking, a key-related watermark sequence $W(K)$ is added in the original host S, so the watermarked signal $X = S + W(K)$. Therefore, $H(X|K) = H(S)$. This is equivalent to the worst case that no watermark is embedded and used for verification since the watermark does not give any help to reduce the interference from the host signal. In some authentication schemes, the watermark sequence $W(K)$ is generated to be dependent on the source S for security reasons. However, the source-dependent sequence is often designed to be pseudorandom with respect to the source, so such dependence is not intended to reduce the randomness of the watermarked signal. The embedded watermark still suffers almost full interference from the host signal in channel differentiation.

Based on the aforementioned discussion, we can see that since the quantization-based method reduces the interference of the watermarked signal $\mathbf{x}$ to distortion channel differentiation, it is superior to spread-spectrum watermarking for achieving semifragility in multimedia authentication.

III. Analysis of AWGN Channels With a Gaussian Source

To support our assertion of the superiority of quantization-based embedding, in this section, we analyze a simple case of additive white Gaussian noise (AWGN) channels with a Gaussian source. The legitimacy of an AWGN distortion channel is specified as follows: an AWGN distortion channel is legitimate if its variance $\sigma^2 < a$ for a constant a, and illegitimate if $\sigma^2 > b$ for a constant $b \geq a$. It is also assumed that the host signal is Gaussian distributed with zero mean and variance σ_s^2. We use the GLRT to derive the optimal decision region for the nonblind scheme, the spread-spectrum method, and the quantization-index-modulation (QIM) scheme. These methods are assessed and compared by using the relative entropy between two hypothesis distributions as well as the receiver operating characteristic (ROC) curves.

A. Nonblind Authentication

We start with the ideal case where $\mathbf{x}$ is known since it gives a performance upper bound for watermark-based authentication. The composite hypothesis testing is to distinguish two sets of Gaussian noise as follows:

$$H_0: \quad \mathbf{y} = \mathbf{x} + \mathbf{z}(\sigma^2) \text{ for } \sigma^2 < a \tag{10a}$$

$$H_1: \quad \mathbf{y} = \mathbf{x} + \mathbf{z}(\sigma^2) \text{ for } \sigma^2 > b \tag{10b}$$

where $\mathbf{z}(\sigma^2)$ is a zero mean white Gaussian sequence with variance σ^2. Writing $\mathbf{z} = \mathbf{y} - \mathbf{x}$ with $\mathbf{x}$ known, the optimal decision region is derived by using the following generalized log-likelihood ratio test:

$$\text{GLLR} = \frac{1}{n} \log \sup_{\sigma^2 < a} f(\mathbf{z}, \sigma^2) - \frac{1}{n} \log \sup_{\sigma^2 > b} f(\mathbf{z}, \sigma^2) > T \tag{11}$$

for some threshold T. Here, the likelihood $f(\mathbf{z}, \sigma^2)$ for Gaussian noise is given by

$$\begin{aligned} f(\mathbf{z}, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{z_i^2}{2\sigma^2}\right) \\ &= \left(2\pi\sigma^2 \exp(\bar{\mathbf{z}^2}/\sigma^2)\right)^{-n/2} \end{aligned} \tag{12}$$

where $\bar{\mathbf{z}^2} = (1/n)||\mathbf{z}||^2$. For a fixed $\bar{\mathbf{z}^2}$, $2\pi\sigma^2 \exp(\bar{\mathbf{z}^2}/\sigma^2)$ is minimized at $\sigma^2 = \bar{\mathbf{z}^2}$. Now we have

$$\begin{aligned} &\frac{1}{n} \log \sup_{\sigma^2 < a} f(\mathbf{z}, \sigma^2) \\ &\quad = \begin{cases} -\frac{1}{2} \log(2\pi e \bar{\mathbf{z}^2}), & \text{if } \bar{\mathbf{z}^2} < a \\ -\frac{1}{2} \log\left(2\pi a \exp(\bar{\mathbf{z}^2}/a)\right), & \text{otherwise.} \end{cases} \end{aligned} \tag{13}$$

$$\begin{aligned} &\frac{1}{n} \log \sup_{\sigma^2 > b} f(\mathbf{z}, \sigma^2) \\ &\quad = \begin{cases} -\frac{1}{2} \log\left(2\pi b \exp(\bar{\mathbf{z}^2}/b)\right), & \text{if } \bar{\mathbf{z}^2} < b \\ -\frac{1}{2} \log(2\pi e \bar{\mathbf{z}^2}), & \text{otherwise.} \end{cases} \end{aligned} \tag{14}$$

Then, the generalized log-likelihood ratio is given by

$$\mathrm{GLLR} = \begin{cases} \frac{1}{2}\left(\frac{\bar{\mathbf{z}^2}}{b} - \log \bar{\mathbf{z}^2} + \log b - 1\right), & \text{if } \bar{\mathbf{z}^2} < a \\ \frac{1}{2}\left(\frac{\bar{\mathbf{z}^2}}{b} - \frac{\bar{\mathbf{z}^2}}{a} + \log b - \log a\right), & \text{if } a \le \bar{\mathbf{z}^2} \le b \\ \frac{1}{2}\left(\log \bar{\mathbf{z}^2} - \frac{\bar{\mathbf{z}^2}}{a} + 1 - \log a\right), & \text{otherwise} \end{cases} \tag{15}$$

which is a strictly decreasing function of $\bar{\mathbf{z}^2}$. Therefore, the optimal decision region A_n given by (11) can be simplified to

$$\bar{\mathbf{z}^2} = \frac{1}{n}\|\mathbf{z}\|^2 = \frac{1}{n}\|\mathbf{y}-\mathbf{x}\|^2 < r^2 \tag{16}$$

for some positive constant r^2, which is related to T, a, and b. This result is consistent with the fact that the mean square average statistic $\|\mathbf{z}\|^2/n$ is a sufficient statistic for the variance σ^2 of a Gaussian distribution.

Type I error probability $\alpha_n(\sigma^2)$ for a legitimate AWGN channel with variance $\sigma^2 < a$ and Type II error probability $\beta_n(\sigma^2)$ for an illegitimate channel with variance $\sigma^2 > b$ are, respectively, given by

$$\alpha_n(\sigma^2) = P[\mathbf{y} \in A_n^c] = P\left[\chi^2(n) > \frac{nr^2}{\sigma^2}\right] \text{ for } \sigma^2 < a \tag{17}$$

$$\beta_n(\sigma^2) = P[\mathbf{y} \in A_n] = P\left[\chi^2(n) < \frac{nr^2}{\sigma^2}\right] \text{ for } \sigma^2 > b \tag{18}$$

where $\chi^2(n)$ denotes chi-square distribution with degree n, and A_n^c is the complement set of A_n.

B. Spread-Spectrum Scheme

The embedding function for the spread-spectrum scheme is given by

$$\mathbf{x} = \mathbf{s} + \mathbf{w}(k) \tag{19}$$

where $\mathbf{w}(k)$ is an additive watermark signal related to the message k and independent of $\mathbf{s}$. The verification procedure for spread-spectrum embedding is the following hypothesis testing problem:

$$H_0: \quad \mathbf{y} = \mathbf{s} + \mathbf{w}(k) + \mathbf{z}(\sigma^2) \text{ for } \sigma^2 < a \tag{20a}$$

$$H_1: \quad \mathbf{y} = \mathbf{s} + \mathbf{w}(k) + \mathbf{z}(\sigma^2) \text{ for } \sigma^2 > b. \tag{20b}$$

The receiver knows the watermark k and, thus, the spread-spectrum signal $\mathbf{w}(k)$. However, the original signal $\mathbf{s}$ is not known to the receiver and, thus, serves as noise to the hypothesis testing of two channels. Using the generalized log-likelihood ratio test, the optimal decision region A_n is given by

$$\|\mathbf{y} - \mathbf{w}(k)\|^2 < nr^2 \tag{21}$$

for some positive constant r^2. Equation (21) gives the best authentication detector structure for spread-spectrum watermarking, which is a distance detector to the embedded watermark. In contrast, in robust watermarking with Gaussian noise, testing the existence of the embedded watermark is best performed by a correlation detector. This illustrates an important distinction between robust and semifragile authentication watermarking. Therefore, the best detector for robust watermarking is not good for authentication watermarking in terms of sem-fragility characterized by two types of error probabilities.

Given the decision region in (21), the Type I and II error probabilities are given by

$$\alpha_n(\sigma^2) = P[\mathbf{y} \in A_n^c] = P\left[\chi^2(n) > \frac{nr^2}{(\sigma_s^2 + \sigma^2)}\right] \text{ for } \sigma^2 < a, \tag{22}$$

$$\beta_n(\sigma^2) = P[\mathbf{y} \in A_n] = P\left[\chi^2(n) < \frac{nr^2}{(\sigma_s^2 + \sigma^2)}\right] \text{ for } \sigma^2 < b \tag{23}$$

respectively, where $\chi^2(n)$ denotes chi-square distribution with degree n, and A_n^c denotes the complement set of A_n. From the aforementioned results, we can see that the additive spread-spectrum signal $\mathbf{w}(k)$ does not help tradeoff the two error probabilities. One would essentially obtain the same results if no signal $\mathbf{w}(k)$ is embedded, thus confirming the intuitive explanation of the spread-spectrum method in Section II-C.

C. Quantization-Based Embedding

In quantization-based schemes, a watermark k is embedded by quantizing the host by $\mathbf{s}$ using a quantization function associated with the watermark. The embedding function is described as follows:

$$\mathbf{x} = Q(\mathbf{s}, k) \tag{24}$$

where $Q(\cdot, k)$ is the quantization function corresponding to the watermark k. Let $\mathcal{C}(k) = \{Q(\mathbf{s}, k) | \forall \mathbf{s} \in \mathcal{S}^n\}$ (i.e., the reconstruction point set of the quantizer). Then, $\mathbf{x}$ is the nearest neighbor of $\mathbf{s}$ in $\mathcal{C}(k)$ in order to reduce embedding distortion. The authenticated signal $\mathbf{x}$ is discretely distributed over the code set $\mathcal{C}(k)$. Its probability function $p(\mathbf{x}|k)$ for $\mathbf{x} \in \mathcal{C}(k)$ conditioned on watermark k can be derived from the distribution of the source $\mathbf{s}$ as follows: $p(\mathbf{x}|k) = P[Q(\mathbf{s}, k) = \mathbf{x}|k] = P[\mathbf{s} \in \mathcal{V}(\mathbf{x})] = \int_{\mathcal{V}(\mathbf{x})} f(\mathbf{s}) d\mathbf{s}$ where $\mathcal{V}(\mathbf{x})$ is the Voronoi region around $\mathbf{x}$ associated with $\mathcal{C}(k)$, and $f(\mathbf{s})$ is the PDF of the host $\mathbf{s}$.

The composite hypothesis testing problem for a QIM embedding scheme becomes the following:

$$H_0: \quad \mathbf{y} = \mathbf{x} + \mathbf{z}(\sigma^2) \text{ for } \sigma^2 < a \tag{25a}$$

$$H_1: \quad \mathbf{y} = \mathbf{x} + \mathbf{z}(\sigma^2) \text{ for } \sigma^2 > a \tag{25b}$$

where $\mathbf{x}$ is distributed over $\mathcal{C}(k)$ with the probability mass function $p(\mathbf{x}|k)$ derived in the above for some given watermark k. Let $f(\mathbf{z}, \sigma^2)$ be the PDF of the zero mean Gaussian sequence $\mathbf{z}$ with variance σ^2, as given in (12). The PDF of $\mathbf{y}$ is given by a convolution of $p(\mathbf{x}|k)$ and $f(\mathbf{z}, \sigma^2)$, which is $\sum_{\mathbf{x} \in \mathcal{C}(k)} p(\mathbf{x}|k) f(\mathbf{y} - \mathbf{x}, \sigma^2)$. The generalized log-likelihood ratio test is given by

$$\begin{aligned} \mathrm{GLLR} = & \frac{1}{n} \log \sup_{\sigma^2 < a} \sum_{\mathbf{x} \in \mathcal{C}(k)} p(\mathbf{x}|k) f(\mathbf{y} - \mathbf{x}, \sigma^2) \\ & - \frac{1}{n} \log \sup_{\sigma^2 > b} \sum_{\mathbf{x} \in \mathcal{C}(k)} p(\mathbf{x}|k) f(\mathbf{y} - \mathbf{x}, \sigma^2) \\ & > T \end{aligned} \tag{26}$$

for some constant T.

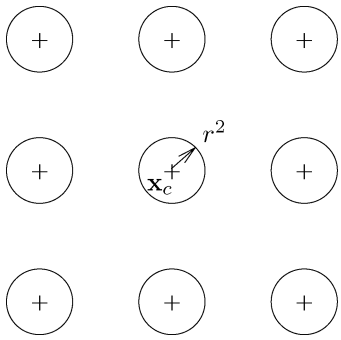


Fig. 3. Decision region for the QIM scheme contains blocks around the reconstruction set of the quantizer associated with the watermark k.

The GLLR depends on finding the value of σ^2, which maximizes the summation of a weighted likelihood, which is difficult to obtain explicitly. Since the term $f(\mathbf{y}-\mathbf{x},\sigma^2) = (2\pi\sigma^2\exp((1/n)\|\mathbf{y}-\mathbf{x}\|^2/\sigma^2))^{-n/2}$ is a decreasing function of $\|\mathbf{y}-\mathbf{x}\|$, the codeword closest to $\mathbf{y}$ has the smallest distance $(1/n)\|\mathbf{y}-\mathbf{x}\|$, thus it is the dominant term in the summation, especially for large n. Since the probability function of $p(\mathbf{x}|k)$ over $\mathcal{C}(k)$ is relatively flat between neighboring codewords, the effect of the $p(\mathbf{x}|k)$ term in the summation is not so significant, compared with the $f(\mathbf{y}-\mathbf{x},\sigma^2)$ term. Therefore, we use the dominant term of the closest codeword to approximate the test statistic.

Let $\mathbf{x}_c$ be the closest codeword in $\mathcal{C}(k)$ to the received signal $\mathbf{y}$. The generalized likelihood ratio test is simplified to the following equation by just using the dominant term of $\mathbf{x}_c$:

$$\frac{1}{n}\log\sup_{\sigma^2<a} p(\mathbf{x}_c|k)f(\mathbf{y}-\mathbf{x}_c,\sigma^2) - \frac{1}{n}\log\sup_{\sigma^2>b} p(\mathbf{x}_c|k)f(\mathbf{y}-\mathbf{x}_c,\sigma^2) > T. \quad (27)$$

Using a similar calculation in the nonblind authentication case, the decision region is obtained as follows:

$$\|\mathbf{y}-\mathbf{x}_c\| < nr^2 \quad (28)$$

for some positive constant r^2. For QIM schemes in which $\mathcal{C}(k)$ is a scalar dithered uniform quantizer, the closest codeword $\mathbf{x}_c$ to $\mathbf{y}$ is given by $\mathbf{x}_c = Q(\mathbf{y}-\mathbf{d}(k))+\mathbf{d}(k)$. The decision region for the QIM scheme is illustrated in Fig. 3. The decision region can be represented by $A_n = \mathcal{C}(k)+Q_n$ where Q_n is an n-dimensional sphere with radius r (i.e., $Q_n = \{\mathbf{z}\in\mathbb{R}^n | \|\mathbf{z}\|^2 < nr^2\}$).

Now we compute Type I and II error probabilities associated with the derived decision region in Fig. 3. The Type I error probability for a legitimate noise with variance $\sigma^2 < a$ is given by $\alpha_n(\sigma^2) = P[\mathbf{y}\notin A_n] = P[\mathbf{x}+\mathbf{z}(\sigma^2)\notin \mathcal{C}(k)+Q_n] = P[\mathbf{z}(\sigma^2)\notin (\mathcal{C}(k)-\mathbf{x})+Q_n]$, which is the probability that $\mathbf{z}(\sigma^2)$ is not in any of the blocks in Fig. 3. This error probability is less than the probability that $\mathbf{z}(\sigma^2)$ is just not in the block around $\mathbf{x}$. So we have

$$\alpha_n(\sigma^2) \le P\left[\mathbf{z}(\sigma^2)\notin Q_n\right] = P\left[\chi^2(n) > \frac{nr^2}{\sigma^2}\right]. \quad (29)$$

The Type II error probability for an illegitimate noise with variance $\sigma^2 > b$ is given by $\beta_n(\sigma^2) = P[\mathbf{y}\in A_n] = P[\mathbf{z}(\sigma^2)\in (\mathcal{C}(k)-\mathbf{x})+Q_n] = P[\mathbf{z}(\sigma^2)\in \bigcup_{\mathbf{x}_c\in\mathcal{C}(k)}(\mathbf{x}_c-\mathbf{x}+Q_n)]$, where $(\mathbf{x}_c-\mathbf{x})+Q_n$ is the decision block around $\mathbf{x}_c-\mathbf{x}$ for $\mathbf{x},\mathbf{x}_c\in\mathcal{C}(k)$ as shown in Fig. 3. In QIM schemes where $\mathcal{C}(k) = \Lambda+\mathbf{d}(k)$ for a base quantizer Λ and a dither vector $\mathbf{d}(k)$, $\mathcal{C}(k)-\mathbf{x} = \Lambda$, so the aforementioned Type II probability is represented by

$$\beta_n(\sigma^2) = P\left[\mathbf{z}(\sigma^2)\in\Lambda+Q_n\right] = \sum_{\boldsymbol{\lambda}\in\Lambda} P\left[\mathbf{z}(\sigma^2)\in\mathcal{V}_{\boldsymbol{\lambda}}(\Lambda)\cap(\boldsymbol{\lambda}+Q_n)\right] \quad (30)$$

where $\mathcal{V}_{\boldsymbol{\lambda}}(\Lambda)$ denotes the Voronoi region around $\boldsymbol{\lambda}$.

TABLE I
RELATIVE ENTROPY BETWEEN LEGITIMATE AND ILLEGITIMATE CHANNELS FOR NONBLIND (NB), SPREAD SPECTRUM, AND QIM SCHEMES

	$\sigma_1^2=64$			$\sigma_1^2=100$		
	NB	SS	QIM	NB	SS	QIM
$\sigma_0^2=4$	1.3237	0.0220	0.1547	1.6294	0.0474	0.1791
$\sigma_0^2=10$	0.7304	0.0175	0.0208	1.0118	0.0409	0.0433

D. Comparison Results

In this section, we compare three scenarios by computing the relative entropy between two hypothesis distributions and two families of error probabilities by using the generalized likelihood-ratio test. We assume a Gaussian host $\mathbf{s}$ with variance $\sigma_s^2 = 200$. In our simulation, we set $a = b = 36$. In other words, the AWGN channel is legitimate for $\sigma^2 < 36$ but illegitimate for $\sigma^2 > 36$.

First, we compute the relative entropy $D(p_{Y|K}\|q_{Y|K})$ between a legitimate channel $p_{Y|X}(y)$ and an illegitimate channel $q_{Y|X}(y)$ in the three embedding scenarios. Since there is a set of legitimate channels, we choose two representative legitimate channels with parameter $\sigma_0^2 = 4$ and $\sigma_0^2 = 10$. Similarly, we select two representative illegitimate channels with $\sigma_1^2 = 64$ and $\sigma_1^2 = 100$. In the nonblind (NB) scheme, since $p_{Y|X}$ and $q_{Y|X}$ are zero-mean Gaussian probability functions with variance σ_0^2 and σ_1^2, respectively, the relative entropy $D(p_{Y|X}\|q_{Y|X}) = 0.5((\sigma_0^2/\sigma_1^2)-1)\log_2 e - 0.5\log_2(\sigma_0^2/\sigma_1^2)$. In the spread-spectrum scheme, a spread-spectrum watermark signal $\mathbf{w}(k)$ of zero mean and variance 5.33 is added to the host signal. So the conditional distributions $p_{Y|K}$ and $q_{Y|K}$ are zero-mean Gaussian probability functions with variance $\sigma_s^2+\sigma_0^2$ and $\sigma_s^2+\sigma_1^2$, respectively, so $D(p_{Y|K}\|q_{Y|K}) = 0.5(((\sigma_s^2+\sigma_0^2)/(\sigma_s^2+\sigma_1^2))-1)\log_2 e - 0.5\log_2((\sigma_s^2+\sigma_0^2)/(\sigma_s^2+\sigma_1^2))$. In the QIM scheme, a scalar uniform quantizer with step size 8 is employed, which results in the same embedding distortion as the SS scheme. The relative entropy $D(p_{Y|K}\|q_{Y|K})$ is numerically computed by using its definition. The values of relative entropy for different legitimate and illegitimate channels are shown in Table I. Recall that the relative entropy characterizes the error exponent of error probabilities and so the error probabilities follow a decreasing trend of $2^{-nD(p_{Y|K}\|q_{Y|K})}$ for a signal of length n. The values of the relative entropy in the table show that the ideal nonblind scenario expectedly has the largest values, and the spread-spectrum scheme has the smallest values. The QIM scheme achieves greater relative entropy than the spread–spectrum scheme. These results show that the QIM

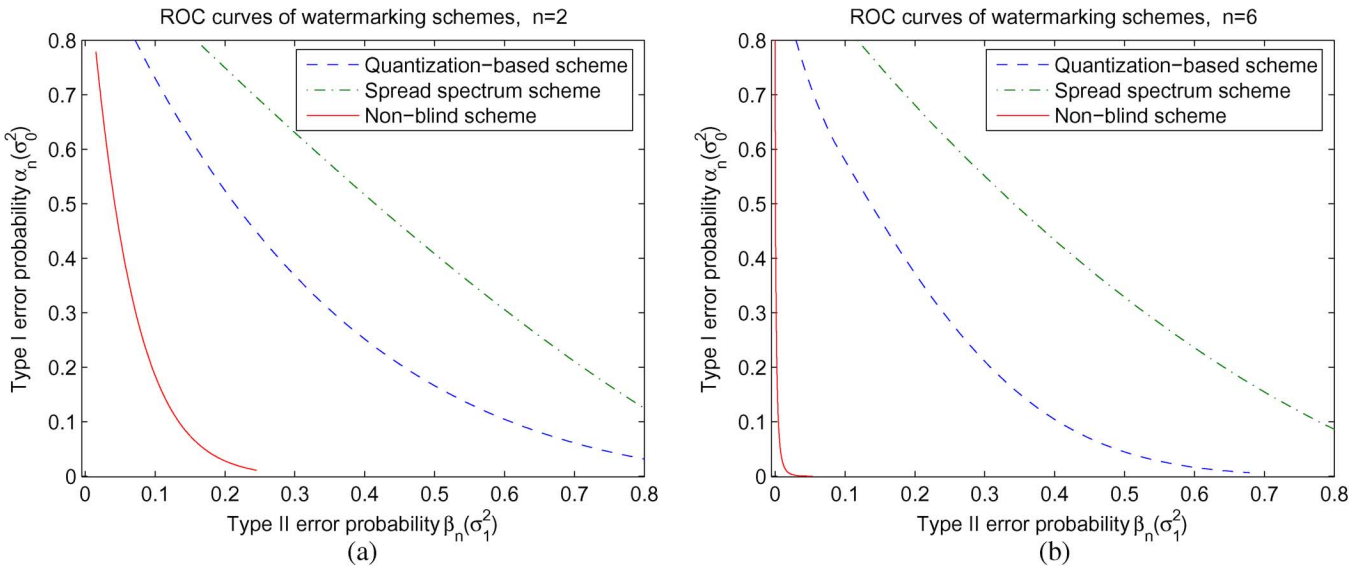


Fig. 4. ROC curves when $n = 2$ and $n = 6$.

scheme can achieve better tradeoff between the two types of error probabilities than the spread-spectrum scheme.

We also compute two families of error probabilities by using the generalized likelihood-ratio test associated with a decision region. A common approach in the assessment of hypothesis testing is the ROC. The ROC is a curve of Type I error probability versus Type II probability as the threshold for decision region varies. In our composite hypothesis testing model, we have two families of error probabilities. To obtain an ROC curve, we again choose a representative parameter from each parameter set. Let $\sigma_0^2 = 4$ be the representative parameter from the legitimate set, and $\sigma_1^2 = 64$ be the one from the illegitimate set. An ROC curve is obtained as $\alpha_n(\sigma_0^2)$ versus $\beta_n(\sigma_1^2)$ as the threshold parameter T or r^2 varies. The ROC curve moves from left to right as T or r^2 increases. Fig. 4 shows the ROC curves of three different schemes for sequence length 2 and 6. We can see from both figures, that the spread-spectrum scheme is the worst scheme, and the ideal nonblind scheme is the best with the QIM in between. When n is increased from 2 to 6, all schemes achieve lower error probabilities (i.e., better semifragility) since more channel outputs are observed in the hypothesis testing. Moreover, with larger n, the improvement of the QIM scheme over the spread-spectrum method also becomes larger. Such improvement will be more significant in a typical authentication system where the number of total pixels n in an image is at least thousands. These simulation results confirm our analytical observation that the quantization-based embedding method outperforms the spread-spectrum method in the ability to distinguish the legitimacy of a distortion channel.

IV. Common Image-Processing Distortions

Our hypothesis testing approach on the analysis of AWGN noise with Gaussian source confirms our intuition that quantization-based watermarking provides the better tradeoff in semifragility. In this section, we analyze certain signal-processing distortions and show how to distinguish effectively between minor and severe changes in quantization-based schemes. Malicious tampering, such as image object removal or substitution, always results in changes of large amplitude; thus, the tampered signal is out of the detection region with high probability. Therefore, we only focus on common signal-processing attacks in this paper.

In quantization-based schemes, a watermark is embedded by quantizing the host. The structure of the quantizer should provide a compromise among semifragility, embedding distortion, and security [35]. For authentication verification, given a test signal, the closest codeword $\mathbf{x}_c$ in the quantizer set corresponding to the watermark is found, and the quantization error $\mathbf{y} - \mathbf{x}_c$ is used to estimate the legitimacy of channel distortion. The test statistic based on the quantization error plays an important role in determining the degree of distortion in order to distinguish minor and severe incidental changes. Based on the hypothesis testing model, we examine the relevant test statistic for specific types of distortions: JPEG compression and filtering.

In general, authentication watermarking, the legitimate set $\mathcal{L}_0$, and the illegitimate set $\mathcal{L}_1$ may include many types of common distortions. For example, $\mathcal{L}_0$ may include high-rate JPEG compression, image enhancement filtering, and other unobtrusive manipulations while $\mathcal{L}_1$ may include low-rate compression, image blurring filtering, and malicious tampering. For such composite specifications, it is complicated to derive the best test statistic for all types of distortions even in the nonblind case because a decision criterion best for one type of distortion may not be optimal for others. Based on the GLRT, the most probable individual type of distortion should be considered. In this section, we therefore focus on only individual types of distortions and derive the test statistic to distinguish minor and severe changes. These single distortion situations reveal the design insight of the problem and, thus, provide general guidelines for general composite distortions.

A. JPEG Compression

JPEG compression is one of the most common incidental modifications due to its widespread use. Thus, many watermarking systems have been proposed to be semifragile to specific degrees of JPEG compression [4], [5], [10], [25]. They consistently utilize a common property of the uniform scalar quantizer that the quantization error due to JPEG compression is bounded in the range of $[-(\Delta/2),(\Delta/2)]$, where Δ is the quantization step for compression quantization in the DCT domain. However, in these systems, fragility against illegitimate JPEG compression is not investigated. The Type II authentication error probability cannot be eliminated since illegitimate quantization may still result in small quantization error in the detection region. In this section, we view quantization noise as a uniformly distributed signal, analyze both error probabilities, and derive the test statistic to identify the legitimacy of the JPEG compression distortion channel.

1) Composite Hypothesis Testing Model: JPEG compression is essentially a quantization operation on image coefficients in the DCT domain. Given the watermarked image coefficients $\mathbf{x}$ in the DCT domain, the quantized signal $\mathbf{y}=Q_\Delta(\mathbf{x})$ where the step size Δ is related to compression quality and the rate of the compressed signal. The quantization error is defined as $\mathbf{z}(\Delta)=\mathbf{y}-\mathbf{x}=Q_\Delta(\mathbf{x})-\mathbf{x}$. For a JPEG compression attack, the high quality factor down to certain level is regarded as legitimate but the low quality factor is illegitimate. The composite hypothesis testing problem for JPEG compression is described as follows:

$$H_0: \quad \mathbf{y}=\mathbf{x}+\mathbf{z}(\Delta) \text{ for } \Delta < a \tag{31a}$$
$$H_1: \quad \mathbf{y}=\mathbf{x}+\mathbf{z}(\Delta) \text{ for } \Delta > b. \tag{31b}$$

Often, the channel input $\mathbf{x}$ is watermarked by quantizing the host $\mathbf{s}$ by using a dithered uniform quantizer $\mathbf{x}=Q(\mathbf{s}-\mathbf{d}(k))+\mathbf{d}(k)$ with the dither value $\mathbf{d}(k)$ randomly chosen from a uniform distribution. The distribution of $\mathbf{x}$ is thus continuous so that the quantization error $\mathbf{z}$ due to subsequent quantization can still be assumed to be uniformly distributed in the range $[-(\Delta/2),(\Delta/2)]$ for high-rate quantization.

We first consider the ideal nonblind scenario that the watermarked signal $\mathbf{x}$ is known, and make a decision based on the quantization error $\mathbf{z}=\mathbf{y}-\mathbf{x}$. The decision region is obtained if

$$\text{GLLR}=\frac{1}{n}\log\sup_{\Delta<a} f_Z(\mathbf{z},\Delta)-\frac{1}{n}\log\sup_{\Delta>b} f_Z(\mathbf{z},\Delta)>T \tag{32}$$

for some threshold T, where $f_Z(\mathbf{z},\Delta)$ is the likelihood of uniformly distributed quantization error, given by

$$f_Z(\mathbf{z},\Delta)=\begin{cases}\left(\frac{1}{\Delta}\right)^n, & \text{if } \max_i|z_i|\le\frac{\Delta}{2}\\ 0, & \text{if } \max_i|z_i|>\frac{\Delta}{2}\end{cases}. \tag{33}$$

We have

$$\frac{1}{n}\log\sup_{\Delta<a} f_Z(\mathbf{z},\Delta) = \begin{cases}-\log\left(2\max_i|z_i|\right), & \text{if } \max_i|z_i|\le a/2\\ -\infty, & \text{if } \max_i|z_i|>a/2\end{cases} \tag{34}$$

$$\frac{1}{n}\log\sup_{\Delta>b} f_Z(\mathbf{z},\Delta) = \begin{cases}-\log b, & \text{if } \max_i|z_i|\le b/2\\ -\log\left(2\max_i|z_i|\right), & \text{if } \max_i|z_i|>b/2\end{cases}. \tag{35}$$

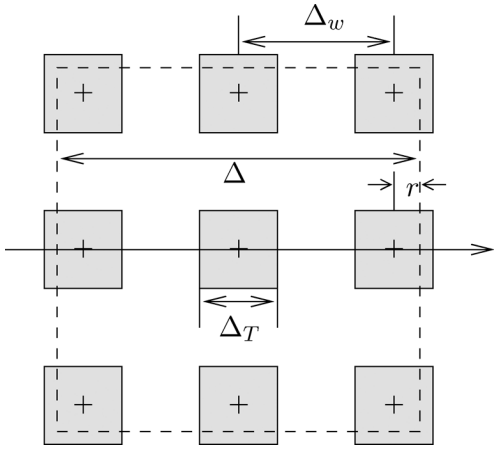


Fig. 5. Shaded region is the decision region A_n for the QIM scheme under uniformly distributed noise with the embedding step size Δ_w. The dashed line represents a uniform distribution $\mathbf{z}$ in the range of $[-(\Delta/2),(\Delta/2)]$. The probability $P[\mathbf{z}\in A_n]$ is the area ratio of the decision region within the dashed line to the entire area within the dashed line.

So (32) can be simplified to

$$\max_i|z_i|<\frac{1}{2}\Delta_T \tag{36}$$

where $\Delta_T=\min\{a,be^{-T}\}$. With the aforementioned optimal decision region, the Type I and Type II error probabilities are, respectively, given by

$$\alpha_n(\Delta)=P\left[\mathbf{z}\in A_n^c\right]=\begin{cases}0, & \text{if } \Delta\le\Delta_T\\ 1-\left(\frac{\Delta_T}{\Delta}\right)^n, & \text{otherwise}\end{cases} \tag{37}$$

for $\Delta<a$, and

$$\beta_n(\Delta)=P[\mathbf{z}\in A_n]=\left(\frac{\Delta_T}{\Delta}\right)^n \text{ for } \Delta>b. \tag{38}$$

The above error probabilities are just those for the nonblind case where the channel input $\mathbf{x}$ is known. For quantization-based schemes, $\mathbf{x}$ has to be estimated from the closest quantized signal. Since its decision region contains all decision blocks around quantization points, the error probabilities in other blocks should also be counted. Suppose the embedding scheme is a QIM scheme by using scalar quantizers of step size Δ_w, and $\Delta_w>\Delta_T$. The decision region for the QIM scheme under uniformly distributed noise is illustrated in Fig. 5. From the figure, the Type I and Type II error probabilities for the QIM scheme are, respectively, given by

$$\alpha_n^{QB}(\Delta)=P\left[\mathbf{z}\in A_n^c\right]=1-P[\mathbf{z}\in A_n] \text{ for } \Delta<a \tag{39}$$
$$\beta_n^{QB}(\Delta)=P[\mathbf{z}\in A_n] \text{ for } \Delta>b \tag{40}$$

where the probability $P[\mathbf{z}\in A_n]$ is the area ratio of the decision region within the range of $[-(\Delta/2),(\Delta/2)]$ to the entire area within the range, given by

$$P[\mathbf{z}\in A_n]=\left(\frac{2(l\Delta_T+r)}{\Delta}\right)^n \tag{41}$$

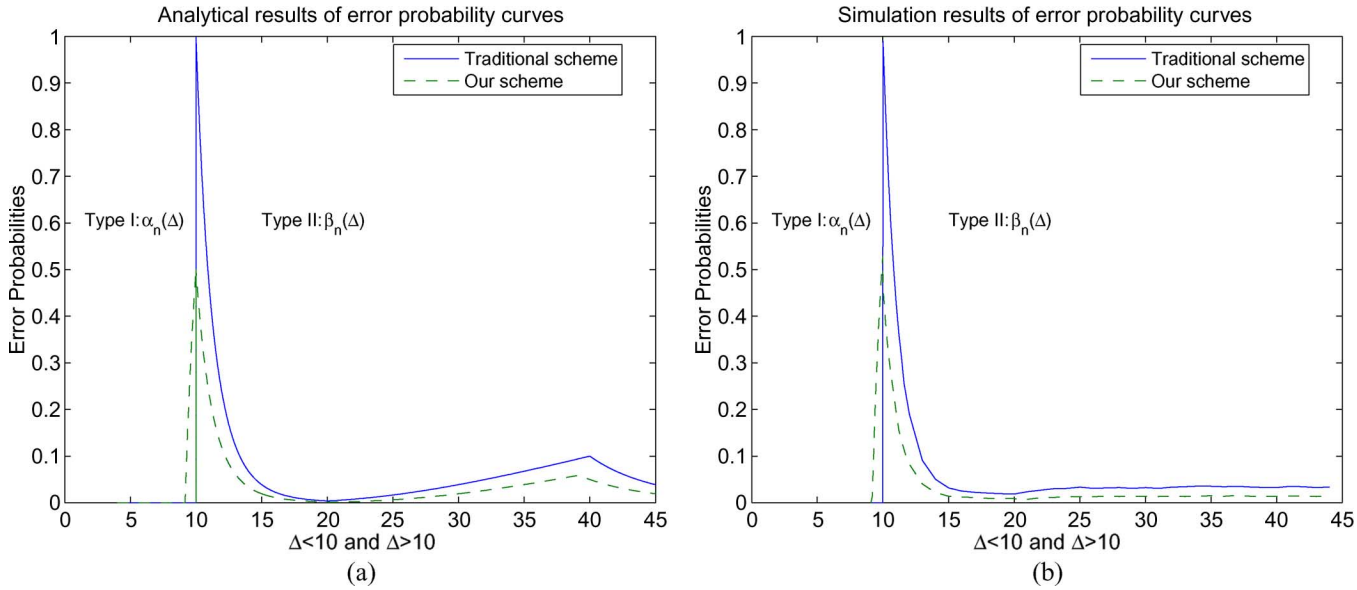


Fig. 6. Error probability curves associated with quantization noise for different choices of decision sets when $n = 8$. (a) Analytical results. (b) Simulation results.

where $l = [\Delta/2\Delta_w]$ (i.e., the nearest integer of the ratio $\Delta/2\Delta_w$), and r is related to the rounding error $e = (\Delta/2) - l\Delta_w$ as follows:

$$r = \begin{cases} -\frac{\Delta_T}{2}, & \text{if } e < -\frac{\Delta_T}{2} \\ e, & \text{if } -\frac{\Delta_T}{2} \le e \le \frac{\Delta_T}{2} \\ \frac{\Delta_T}{2}, & \text{if } e > \frac{\Delta_T}{2}. \end{cases} \tag{42}$$

A detailed derivation of (42) can be found in [36].

2) Comparison Results: Existing semifragile systems for JPEG compression employ a robust watermarking scheme. Since all legitimate quantization steps $\Delta < a$, the quantization error is bounded within $[-(a/2), (a/2)]$. This is equivalent to using the decision threshold $\Delta_T = a$. For this setting, the quantization error due to legitimate quantization falls in the decision region, so Type I error probabilities are all zero. However, Type II error probabilities are not taken into consideration. From the equations of Type I and Type II error probabilities that were previously shown, an optimal choice of Δ_T should balance two error probabilities. We choose Δ_T so that the maximal values of both Type I and II error probabilities are equal. Since Type I error probability achieves its maximum at $\Delta = a$ and Type II error probability achieves its maximum at $\Delta = b$, we have $\Delta_T = ab/\sqrt[n]{a^n + b^n}$.

We compare the traditional scheme with our scheme by setting $n = 8$, $a = b = 10$, and $\Delta_w = 15$. The error curves are shown in Fig. 6(a). The error curve when $\Delta < 10$ describes the robustness property and the curve when $\Delta > 10$ describes the fragility property. We can see that the traditional scheme does not have any Type I error probability, so the robustness objective is fully achieved. However, for all $\Delta > 10$, the scheme has greater Type II error probability than our scheme, so the fragility objective is worse. We see from the figure that our scheme has smaller average overall error probability.

In practice, quantization is a deterministic process, and the quantization error is also dependent on the watermarked signal. Therefore, we also simulate the semifragile scheme to quantization distortion in the channel. We simulate the semifragile QIM scheme on 8×8 blocks of image Lenna. In the simulation, medium DCT coefficients from frequency band 5 to 40 in zigzag order of 64 bands are used for embedding and the same quantization step size $\Delta_w = 15$ is used for all coefficients as in Fig. 6(a). Type I and II error rates are measured by the percentage of estimated quantization noise greater than the decision threshold. Fig. 6(b) shows the error-rate curves associated with quantization distortion as the quantization step Δ varies. We see results similar to those in Fig. 6(a). The traditional scheme does not have any Type I error probability, so it is better than our scheme in terms of robustness, but worse in terms of fragility when the quantization becomes illegitimate. Our scheme can achieve smaller average overall error probability associated with fragility which is more important than robustness requirements in semifragile watermarking.

B. Image Filtering

One common objective of image filtering is to remove noise from an image, while still keeping good visual quality. Real natural images have energy concentrated in low frequencies in contrast to noise that can often occur at higher frequencies. Therefore, image filtering is often low pass in order to remove those components of noise without interfering with the signal component. For image applications, there are three common categories of filters: 1) linear filters, including neighbor averaging and Gaussian filters; 2) rank-value filters, such as median filters; and 3) adaptive filters.

It is better to analyze images in the frequency domain to investigate the effects of filtering distortions. Here, we focus on linear filters since they provide a tractable channel representation. The effects of nonlinear or adaptive filters can be approximated by linear filters. In the frequency domain, filtering is regarded as a product operation as $Y(U,V) = X(U,V)H(U,V)$, where $X(U,V)$, $Y(U,V)$, and $H(U,V)$ are the host image, the filtered image, and the filter, respectively. Then, the filtering

model can be represented by an additive model by taking a logarithm on the magnitude of the frequency as follows:

$$\log|Y(U,V)| = \log|X(U,V)| + \log|H(U,V)|. \tag{43}$$

The additive term $\log|H(U,V)|$, representing the effect of a filtering distortion on the host signal, is small for smooth filtering and large for severe filtering. We then can apply our approach to semifragile watermarking to detect the degree or legitimacy of filtering distortions. The idea here is to apply quantization-based watermarking in the signal $\log|X(U,V)|$ at the frequency band (U,V) so that the severity of the filtering can be measured from the quantization error of the test signal $\log|Y(U,V)|$.

Whether filtering distortion is legitimate or not depends on the marked image $X(U,V)$ and the frequency band (U,V). We do not attempt to produce a legitimate or illegitimate answer. Rather, we provide a test statistic to measure the severity degree of filtering distortion and let the receiver judge the legitimacy from the test statistic value. To better illustrate our idea of quantization-based watermarking, we use a Gaussian filter as an example and control its degree of degradation using a single parameter.

1) Gaussian Filters: The Gaussian filter is a linear filter whose low-pass filter curve is a Gaussian function with a single degradation parameter. Gaussian filters have advantages of the absence of ringing artifacts and noise leakage since there are no sidelobes in the spatial and frequency domains [37].

A Gaussian filter is given by its impulse response

$$h(m,n) = \frac{1}{2\pi\sigma^2} e^{-(m^2+n^2)/2\sigma^2} \tag{44}$$

where σ^2 is a parameter which determines the degree of degradation. Its frequency spectrum is approximated by [37]

$$H(U,V) \approx e^{-2\pi^2\sigma^2(U^2+V^2)} \text{ for } |U|,|V| < 1/2, \tag{45}$$

which is also a Gaussian function. The parameter σ^2 controls the shapes of spatial and frequency responses of Gaussian filters. When σ^2 is small, the filter has flat frequency response, thus removing high-frequency components but preserving most of the low- and medium-frequency components. When σ^2 becomes large, the filter removes medium and high frequencies or even low frequencies, which may blur the edges. Therefore, Gaussian filters are legitimate when $\sigma^2 < a$, but illegitimate when $\sigma^2 > b$, where a, b are two positive constant and $b \geq a$. Given the received image $Y(U,V)$ represented in the frequency domain and the known embedded watermark in $X(U,V)$, the composite hypothesis testing problem of Gaussian filtering is described as follows:

$$H_0: \quad \log|Y(U,V)| = \log|X(U,V)| - 2\pi^2\sigma^2(U^2+V^2) \quad \text{for } \sigma^2 < a \tag{46a}$$

$$H_1: \quad \log|Y(U,V)| = \log|X(U,V)| - 2\pi^2\sigma^2(U^2+V^2) \quad \text{for } \sigma^2 > b. \tag{46b}$$

2) Proposed Quantization-Based Scheme: We apply a QIM scheme to embed a watermark by quantizing the signal $\log|X(U,V)|$. In the frequency band (U,V), the dithered quantizer set $C(k)$ is designed to be that of $X(U,V)$, satisfying $\log|X(U,V)| = \Delta(U,V)(i + d(k))$ for some integer i and a given dither value $d(k)$, where $\Delta(U,V)$ is the quantization step size in frequency (U,V). We set $\Delta(U,V) > 2\pi^2 a(U^2+V^2)$ to allow correct recovery of $\log|X(U,V)|$ under legitimate filtering.

Assume that the host image has coefficients $S(U,V)$ in the frequency domain. The authentication embedder is to find an appropriate codeword $X(U,V)$ in the code set. The closest codeword $X(U,V)$ is obtained by

$$\log|X(U,V)| = \left[\frac{\log|S(U,V)|}{\Delta(U,V)} - d(k)\right]\Delta(U,V) + d(k)\Delta(U,V) \tag{47}$$

where $[\cdot]$ denotes rounding to the nearest integer. The watermark is only embedded in the magnitude of the coefficients and the phase is kept unchanged.

At the receiver side, given a received image $Y(U,V)$ represented in the frequency domain, the watermarked signal $\log|X(U,V)|$ is first recovered from $\log|Y(U,V)|$ as follows:

$$\log\left|\hat{X}(U,V)\right| = \left\lceil\frac{\log|Y(U,V)|}{\Delta(U,V)} - d(k)\right\rceil\Delta(U,V) + d(k)\Delta(U,V) \tag{48}$$

where $\lceil x \rceil$ denotes the ceiling function, which gives the smallest integer $\geq x$. The equation guarantees $\log|\hat{X}(U,V)| \geq \log|Y(U,V)|$. We then can estimate σ^2 from $\log|Y(U,V)|$ and the recovered $\log|\hat{X}(U,V)|$ as follows:

$$\hat{\sigma}^2(U,V) = \frac{\log\left|\hat{X}(U,V)\right| - \log|Y(U,V)|}{2\pi^2(U^2+V^2)}. \tag{49}$$

From individual estimates $\hat{\sigma}^2(U,V)$ in all available frequencies, we then estimate the overall degree of degradation $\hat{\sigma}^2$. A natural choice is the weighted average over all frequency bands, i.e.,

$$\hat{\sigma}^2 = \frac{\sum_{U,V} T(U,V)\hat{\sigma}^2(U,V)}{\sum_{U,V} T(U,V)} \tag{50}$$

where $T(U,V)$ is the weight function. Since natural images have energy concentrated in low frequencies, we should trust more of the individual estimates in low-frequency bands. Generally, a larger weight should be given in low frequencies than in high frequencies. In our experiment, we set $T(U,V) = |Y(U,V)|^2$ since the image magnitude $|Y(U,V)|$ is larger in low frequencies than in high frequencies.

Such estimation of the overall degree of degradation can be extended to general filtering operations which may not have a nice closed-form expression of their frequency response. We can again average the estimated distortions over all frequency bands as follows:

$$\hat{\sigma}^2 = \frac{\sum_{U,V} T(U,V)\left(\log\left|\hat{X}(U,V)\right| - \log|Y(U,V)|\right)}{\sum_{U,V} T(U,V)} \tag{51}$$

for a certain weight function $T(U,V)$.

Finally, legitimacy decision is made from the estimated overall degree of degradation $\hat{\sigma}^2$. The effect of a filtering operation on visual quality heavily depends on the original image. Whether filtering distortion is legitimate or not is a subjective decision. Therefore, we just report the estimated overall degree of degradation to the receiver, and leave it to the receiver for judgement on whether the filtering distortion is legitimate or not.

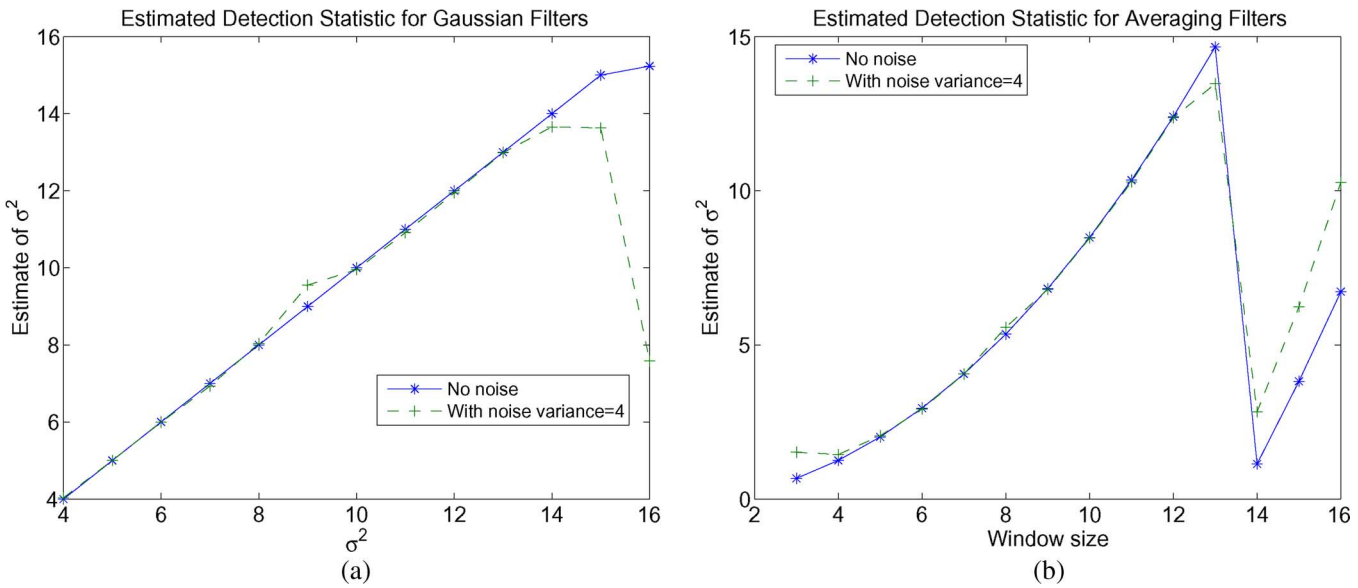


Fig. 7. Estimate of σ^2. The dashed line represents the results when additional Gaussian noise of variance 4 is added to the filtered image. (a) For Gaussian filters. (b) For averaging filters.

3) Simulation Results: In this section, we simulate our proposed semifragile algorithm for image filtering on the test image Lenna. Given the host image of size $M \times N$, we first transform it into the frequency domain by using discrete Fourier transform (DFT). Since the coefficients in frequency domain are complex, we only focus on their magnitude for embedding. Also, since the Fourier representation of a real signal obeys conjugate symmetry, we can use, at most, $MN/4$ of the coefficients in the frequency domain for watermark embedding. In our simulation, if the entire frequency coefficients are employed for authentication embedding, the watermark embedding results in visible artifacts since the quantization step $\Delta(U,V) > 2\pi^2 a(U^2+V^2)$ in all frequency bands. Therefore, we secretly select part of the frequency coefficients for authentication embedding. We choose three frequency segments for embedding: 16 coefficients in low frequency, 16 in medium frequency, and 8 in high frequency for each dimension of the image. In total, 40×40 coefficients are selected for watermark embedding.

In the simulation, (50) is used to estimate the overall degree of degradation from different frequency bands. We set the embedding quantization step for watermark embedding $\Delta(U,V) = 32\pi^2(U^2+V^2)$. These parameters result in allowable embedding distortion, described by a peak-signal-to-nose ratio $(\mathrm{PSNR}) = 45.48$ dB. In our simulation, to test the accuracy of the estimation, we also apply additive Gaussian noise to the filtered image, so the received image $y(m,n) = x(m,n) * h(m,n) + z(m,n)$, where $z(m,n)$ is the AWGN. Fig. 7(a) shows the simulation results of estimating the parameter σ^2 from the noisy filtered image $y(m,n)$ using Gaussian filters of various degrees σ^2. We can see that the estimation is very accurate for Gaussian filtering even if additional Gaussian noise of variance 4 is added. When the additive noise increases, the estimation becomes less accurate. When the actual σ^2 of the Gaussian filtering occurring on the watermarked image is close to 16, the estimate value of σ^2 drops to 8. This phenomenon takes places because the estimate of σ^2 is computed from the recovered watermarked image. When σ^2 is close to 16, the additive noise may push the watermarked signal into the next quantization segment. Therefore, the resulting individual estimate of σ^2 lies between 0 and 16, so the overall estimate $\hat{\sigma}^2$ approaches the average of 8.

In our experiments, we also apply averaging filters to the proposed system which was designed based on Gaussian filters in order to demonstrate its ability in differentiating minor and severe changes for general filtering operations. Fig. 7(b) shows simulation results of the estimated degree of degradation from (50) when averaging filters with various window sizes are applied to the watermarked image. We can see from the figure that the degree of degradation increases as window size increases. Analytic derivation by using the frequency response of the averaging filter on the right-hand side of (49) also shows that these two parameters are closely related. In the figure, when the window size exceeds 12, the estimated value drops to 0, and then increases again. Again, this is because the watermarking step size Δ is exceeded, so the quantized value jumps from Δ down to 0, then increases again. The simulation shows that although our scheme is designed based on Gaussian filters, the estimated degree of degradation is also a test statistic to detect the legitimacy of general filters according to filtering effects in different frequency bands. Authentication decisions are then made based on significance of the test statistic.

C. Remarks on Reducing Embedding Distortion

In the proposed quantization-based approach for compression and filtering, the embedding step size is chosen to be the maximum allowed step size for legitimate processing so that legitimate and illegitimate channel distortions can be effectively distinguished. Such embedding introduces as much distortion as legitimate processing can introduce, resulting in possible visible artifacts in the host. To reduce embedding distortion, we only employ a part of the coefficients for embedding instead of the entire host in our simulation for image filtering. By partial embedding, security on the remaining coefficients only relies on secret selection; thus, it could be weak against active attacks. Thus, such partial authentication embedding trades off between embedding distortion and system security requirements.

Another way to reduce embedding distortion is to use the distortion compensation technique proposed to achieve greater embedding channel capacity in communications with side information [38], [39]. In the distortion-compensated QIM embedding, quantization is performed on the $\alpha\mathbf{s}$ domain where α is a weighting compensation factor $0 < \alpha \leq 1$ and $\alpha = 1$ in the case of no compensation. The embedding function is described as

$$\mathbf{x} = \mathbf{s} - \mathbf{u} = (1-\alpha)\mathbf{s} + Q(\alpha\mathbf{s}, k) \tag{52}$$

where $\mathbf{u} = \alpha\mathbf{s} - Q(\alpha\mathbf{s}, k)$ is the distortion due to standard QIM embedding on $\alpha\mathbf{s}$. Our previous paper [26] has analyzed the distortion compensation technique to reduce embedding distortion. We find that the distortion compensation technique results in equivalent channel noise $\mathbf{z}_{eq} = (1-\alpha)\mathbf{u} + \alpha\mathbf{z}$ where $\mathbf{z}$ is the channel noise which needs to decide whether it is legitimate or not. By taking an appropriate value of $\alpha \leq (2D/(D+\sigma_z^2))$ where $D = E\{\mathbf{u}^2\}$ is the embedding distortion and $\sigma_z^2 = E\{\mathbf{z}^2\}$ is the variance of the channel noise, the variance of the equivalent channel noise $\sigma_{z_{eq}}^2$ is smaller than the embedding distortion D, so $\mathbf{z}_{eq}$ can be recovered from the quantization embedding on $\alpha\mathbf{s}$. However, with the self-noise term $(1-\alpha)\mathbf{u}$, it becomes more difficult to make a correct decision as to whether the channel noise $\mathbf{z}$ is legitimate or not, based on the recovered $\mathbf{z}_{eq}$. Therefore, the distortion compensation technique results in more Type I and II error probabilities than the no-compensation case. This is a tradeoff between embedding distortion and semifragility (characterized by two error probabilities).

V. Conclusion

This paper studies watermark embedding to achieve semifragile multimedia authentication through a composite hypothesis testing approach. Our results show that the quantization-based embedding method outperforms spread spectrum in the tradeoff between algorithm robustness and fragility. Based on the hypothesis testing model, we also analyze certain common image-processing distortions, such as JPEG compression and filtering, and demonstrate how our approach can distinguish effectively minor changes from severe ones in quantization-based authentication watermarking. The results in this paper show that the hypothesis testing model provides insights for authentication watermarking and allows better control of robustness and fragility in specific applications.

Appendix
Proof of (9) in Section II-C

Let K, X, and Y be the random variables corresponding to the watermark key, the channel input, and output, respectively. We use the notations p and q to denote different PDFs of certain variables under different channels $p_{Y|X}$ and $q_{Y|X}$, respectively. For example, $p_{Y|K}$ and $q_{Y|K}$ represent the conditional PDFs of Y with K under channels $p_{Y|X}$ and $q_{Y|X}$, respectively. Similarly, p_Y and q_Y are two marginal PDFs of Y under channels $p_{Y|X}$ and $q_{Y|X}$, respectively.

Theorem 1: Assume the random variables K, X, and Y form a Markov chain [34] in an authentication model as follows:

$$K \xrightarrow{f_{X|K}} X \underset{q_{Y|X}}{\overset{p_{Y|X}}{\rightarrow}} Y$$

where $f_{X|K}$ is the conditional PDF of X with K due to the embedding of K in a source, and $p_{Y|X}$ and $q_{Y|X}$ are two-channel PDFs to be differentiated. Then, the following inequalities hold:

$$D(p_Y \| q_Y) \leq D(p_{Y|K} \| q_{Y|K}) \leq D(p_{Y|X} \| q_{Y|X}). \tag{53}$$

Proof: We prove the first inequality. By the chain of relative entropy [34], we have two expansions

$$D(p_{Y,K} \| q_{Y,K}) = D(p_Y \| q_Y) + D(p_{K|Y} \| q_{K|Y}) \tag{54}$$

$$= D(p_K \| q_K) + D(p_{Y|K} \| q_{Y|K}). \tag{55}$$

Since p_K and q_K are equal under different channels, $D(p_K \| q_K) = 0$. Also from nonnegativity of $D(p_{K|Y} \| q_{K|Y})$, we have $D(p_Y \| q_Y) \leq D(p_{Y|K} \| q_{Y|K})$.

Similarly, we can prove the second inequality by using the chain of relative entropy

$$D(p_{Y,X|K} \| q_{Y,X|K}) = D(p_{Y|K} \| q_{Y|K}) + D(p_{X|Y,K} \| q_{X|Y,K}) \tag{56}$$

$$= D(p_{X|K} \| q_{X|K}) + D(p_{Y|X,K} \| q_{Y|X,K}). \tag{57}$$

Since the embedding mapping $K \rightarrow X$ is the same under two distortion channels from $X \rightarrow Y$, both $p_{X|K}$ and $q_{X|K}$ are equal to $f_{X|K}$, hence $D(p_{X|K} \| q_{X|K}) = 0$. From the nonnegativity of $D(p_{X|Y,K} \| q_{X|Y,K})$, we have

$$D(p_{Y|K} \| q_{Y|K}) \leq D(p_{Y|X,K} \| q_{Y|X,K}). \tag{58}$$

Since $K \rightarrow X \rightarrow Y$ forms a Markov chain, $p_{Y|X,K} = p_{Y|X}$ and $q_{Y|X,K} = q_{Y|X}$. Therefore, the second inequality $D(p_{Y|K} \| q_{Y|K}) \leq D(p_{Y|X} \| q_{Y|X})$ holds.

References

[1] S. Walton, "Image authentication for a slippery new age," *Dr. Dobb's J.* vol. 20, no. 4, pp. 18–26, Apr. 1995. [Online]. Available: http://www.ddj.com/documents/s=992/ddj9504a/.

[2] M. M. Yeung and F. Mintzer, "An invisible watermarking technique for image verification," presented at the IEEE Int. Conf. Image Processing Santa Barbara, CA, Oct. 1997.

[3] P. W. Wong, "A public key watermark for image verification and authentication," in *Proc. IEEE Int. Conf. Image Processing*, May 1998, vol. I, pp. 455–459.

[4] D. Kundur and D. Hatzinakos, "Digital watermarking for telltale tamper-proofing and authentication," *Proc. IEEE*, vol. 87, no. 7, pp. 1167–1180, Jul. 1999.

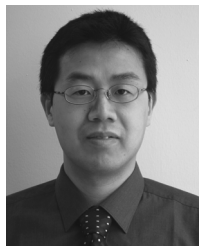
[5] C.-Y. Lin and S.-F. Chang, "Semi-fragile watermarking for authenticating JPEG visual content," in *Proc. SPIE: Security and Watermarking of Multimedia Content II*, Jan. 2000, pp. 140–151.

[6] C.-Y. Lin and S.-F. Chang, "A robust image authentication method distinguishing JPEG compression from malicious manipulation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 2, pp. 153–168, Feb. 2001.

[7] J. Fridrich, "Combining low frequency and spread spectrum watermarking," in *Proc. SPIE*, 1998, vol. 3456, pp. 2–12.

[8] J. Fridrich, "A hybrid watermark for tamper detection in digitlal images," in *Proc. ISSPA Conf.*, Brisbane, Australia, 1999, pp. 301–304.

[9] G.-J. Yu, C.-S. Lu, H.-Y. M. Liao, and J.-P. Sheu, "Mean quantization blind watermarking for image authentication," in *Proc. IEEE Int. Conf. Image Processing*, 2000, pp. 706–709.
[10] G.-J. Yu, C.-S. Lu, and H.-Y. M. Liao, "Mean quantization blind watermarking for image authentication," *Opt. Eng.*, vol. 40, no. 7, pp. 1396–1408, 2001.
[11] F. Alturki and R. Mersereau, "Secure fragile digital watermarking technique for image authentication," in *Proc. IEEE Int. Conf. Image Processing*, 2001, pp. 1031–1034.
[12] E. T. Lin, C. I. Podilchuk, and E. J. Delp, "Detection of image alterations using semi-fragile watermarks," in *Proc. SPIE*, 2000, vol. 3971, pp. 152–163.
[13] J. J. Eggers and B. Girod, "Blind watermarking applied to image authentication," presented at the ICASSP, Salt Lake City, UT, May 2001.
[14] C. Rey and J.-L. Dugelay, "Blind detection of malicious alterations on still images using robust watermarks," in *Inst. Elect. Eng. Seminar: Secure Images and Image Authentications*, 2000.
[15] Q. Sun, S.-F. Chang, M. Kurato, and M. Suto, "A new semi-fragile image authentication framework combining ECC and PKI," presented at the Special Session on Multimedia Watermarking, Phoenix, AZ, 2002.
[16] Q. Sun and S.-F. Chang, "Semi-fragile image authentication using generic wavelet domain features and ECC," presented at the IEEE Int. Conf. Image Processing, Rochester, NY, 2002.
[17] K. Maeno, Q. Sun, S.-F. Chang, and M. Suto, "New semi-fragile image authentication watermarking techniques using random bias and nonuniform quantization," *IEEE Trans. Multimedia*, vol. 8, no. 1, pp. 32–45, Feb. 2006.
[18] D. Zou, Y. Q. Shi, Z. Ni, and W. Su, "A semi-fragile lossless digital watermarking scheme based on integer wavelet transform," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 10, pp. 1294–1300, Oct. 2006.
[19] Z. Peng and W. Liu, "Color image authentication based on spatiotemporal chaos and SVD," *Chaos, Solitons Fractals*, vol. 36, no. 4, pp. 946–951, May 2008.
[20] H. L. V. Trees, *Detection, Estimation and Modulation Theory, Part I*. New York: Wiley, 2001.
[21] I. J. Cox, M. L. Miller, and J. A. Bloom, *Digital Watermarking*. San Mateo, CA: Morgan Kaufmann, 2002.
[22] P. Moulin and R. Koetter, "Data-hiding codes," *Proc. IEEE*, vol. 93, no. 12, pp. 2083–2126, Dec. 2005.
[23] O. Dabeer, K. Sullivan, U. Madhow, S. Chandrasekaran, and B. Manjunath, "Detection of hiding in the least significant bit," *IEEE Trans. Signal Proces. Supplement Secure Media I*, vol. 52, no. 10, pp. 3046–3058, Oct. 2004.
[24] U. M. Maurer, "Authentication theory and hypothesis testing," *IEEE Trans. Inf. Theory*, vol. 46, no. 4, pp. 1350–1356, Jul. 2000.
[25] M. Wu and B. Liu, "Watermarking for image authentication," in *Proc. IEEE Int. Conf. Image Processing*, Chicago, IL, 1998, vol. II, pp. 437–441.
[26] C. Fei, D. Kundur, and R. H. Kwong, "Analysis and design of secure watermark-based authentication systems," *IEEE Trans. Inf. Forensics Security*, vol. 1, no. 1, pp. 43–55, Mar. 2006.
[27] A. B. Watson, "DCT quantization matrices optimized for individual images," in *Proc. SPIE Human Vision, Visual Processing, and Digital Display IV*, 1993, pp. 202–216.
[28] A. B. Watson, G. Y. Yang, J. A. Solomon, and J. Villasenor, "Visual thresholds for wavelet quantization error," in *Proc. SPIE Human Vision, Visual Processing, and Digital Display IV*, 1996, pp. 381–392.
[29] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structual similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[30] S. M. Kay, *Fundamentals of Statistical Signal Processing: Detection Theory*. Upper Saddle River, NJ: Prentice-Hall, 1993.
[31] J. A. Rice, *Mathematical Statistics and Data Analysis*, 2nd ed. Belmont, CA: Duxbury, 1995.
[32] D. D. Wackerly, W. Mendenhall, III, and R. L. Scheaffer, *Mathematical Statistics With Applications*, 6th ed. Boston, MA: Duxbury, 2002.
[33] R. E. Blahut, *Principles and Practice of Information Theory*. Reading, MA: Addison-Wesley, 1987.
[34] T. M. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[35] C. Fei, D. Kundur, and R. H. Kwong, "Achieving computational and unconditional security in authentication watermarking: Analysis, insights, and algorithms," in *Proc. SPIE: Security, Steganography, and Watermarking Multimedia Content VII*, San Jose, CA, Jan. 2005, vol. 5681, pp. 697–708.
[36] C. Fei, "Analysis and design of watermark-based multimedia authentication systems," Ph.D. dissertation, Univ. Toronto, Toronto, ON, Canada, 2006.
[37] A. C. Bovik and S. T. Acton, "Basic linear filtering with application to image enhancement," in *Handbook of Image and Video Processing*, A. C. Bovik, Ed. London, U.K.: Academic Press, 2000, ch. 3.1, pp. 71–79.
[38] B. Chen and G. W. Wornell, "Quantization index modulation: A class of provably good methods for digital watermarking and information embedding," *IEEE Trans. Inf. Theory*, vol. 47, no. 4, pp. 1423–1443, May 2001.
[39] R. Zamir, S. Shamai, and U. Erez, "Nested linear/lattice codes for structured multiterminal binning," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1250–1275, Jun. 2002.



Chuhong Fei (S'04–M'06) was born in Zhejiang, China. He received the B.E. and M.E. degrees from Xi'an Jiaotong University, Xi'an, China, in 1994 and 1997, respectively, and the M.A.Sc. and Ph.D. degrees in electrical and computer engineering from the University of Toronto, Toronto, ON, Canada, in 2001 and 2006, respectively.

Currently, he is a Research Scientist with A.U.G. Signals Ltd., Toronto. His research interests include multimedia security, data hiding, multimedia signal processing, radar image processing, and remote sensing.

Raymond H. Kwong (M'75–SM'08) was born in Hong Kong, China, in 1949. He received the S.B., S.M., and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge, in 1971, 1972, and 1975, respectively.

From 1975 to 1977, he was a visiting Assistant Professor of Electrical Engineering at McGill University, Montreal, QC, Canada, and a Research Associate at the Centre de Recherches Mathematiques, Universite de Montreal, Montreal. Since 1977, he has been with the Edward S. Rogers Sr. Department of Electrical and Computer Engineering at the University of Toronto, Toronto, ON, Canada, where he is currently a Professor. His current research interests include estimation and stochastic control, adaptive signal processing and control, fault diagnosis and fault-tolerant control, discrete event systems, and multimedia security.

Deepa Kundur (S'93–M'99–SM'03) received the B.A.Sc., M.A.Sc., and Ph.D. degrees in electrical and computer engineering from the University of Toronto, Toronto, ON, Canada, in 1993, 1995, and 1999, respectively.

She joined the Department of Electrical Engineering at Texas A&M University, College Station, in 2003, where she is a member of the Wireless Communications Laboratory and is Associate Professor. Before joining Texas A&M, she was an Assistant Professor with the Edward S. Rogers Sr. Department of Electrical and Computer Engineering at the University of Toronto where she was the Bell Canada Junior Chair-holder in Multimedia and an Associate Member of the Nortel Institute for Telecommunications. Her research interests include protection of scalar and broadband sensor networks, multimedia security, and computer forensics.

Dr. Kundur is an elected member of the IEEE Information Forensics and Security Technical Committee, Vice-Chair of the Security Interest Group of the IEEE Multimedia Communications Technical Committee, and on the editorial boards of IEEE COMMUNICATION LETTERS, IEEE TRANSACTIONS ON MULTIMEDIA, and *EURASIP Journal on Information Security*. She was recently a General Chair of the 2007 ACM Workshop on Multimedia and Security and a Guest Editor of the 2007 *EURASIP Journal on Advances in Signal Processing Special Issue on Visual Sensor Networks*. She was a guest editor of the 2004 Proceedings of the IEEE Special Issue on Enabling Security Technologies for Digital Rights Management and the recipient of the 2005 Tenneco Meritorious Teaching Award, the 2006 Association of the Former Students College Level Teaching Award, and the 2007 Outstanding Professor Award in the ECE Department at Texas A&M University.