A Hypothesis Testing Approach for Achieving Semi-fragility in Multimedia Authentication

Chuhong Fei[a], Deepa Kundur[b], and Raymond Kwong[a]

[a]University of Toronto, 10 King's College Road, Toronto, ON Canada M5S 3G4;
[b]Texas A&M University, 3128 TAMU, College Station, TX USA 77843-3128

ABSTRACT

This paper studies the problem of achieving watermark semi-fragility in multimedia authentication through a composite hypothesis testing approach. The embedding of a semi-fragile 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 a priori information. Based on the hypothesis testing model, we investigate the best embedding strategy which assists the watermark verifier to make correct decisions. Our results show that the quantization-based watermarking method is more appropriate than the spread spectrum method to achieve the best tradeoff between two error probabilities. This observation is confirmed by a case study of additive Gaussian white noise channel with Gaussian source using two figures of merit: relative entropy of the two hypothesis distributions and the receiver operating characteristic. Finally, we focus on certain common signal processing distortions such as JPEG compression and image filtering, and investigate the best test statistic and optimal decision regions to distinguish legitimate and illegitimate distortions. The results of the paper show that our approach provides insights for authentication watermarking and allows better control of semi-fragility in specific applications.

Keywords: Digital Watermarking, Semi-fragile watermarking, Multimedia Authentication, Hypothesis Testing

1. INTRODUCTION

This work focuses on achieving semi-fragility in multimedia authentication using a hypothesis testing approach to verify the legitimacy of possible distortions. The goal of multimedia authentication is to authenticate the visual meaning of multimedia data to ensure its integrity. Thus, content authentication of a host signal has two main objectives: to alert a party to unacceptable distortions on the host and to authenticate the legitimate source. Possible distortions on a signal can be divided into two groups: legitimate and illegitimate distortions. When a signal undergoes a legitimate distortion which does not alter the visual content of the data, the system should indicate that the signal is authentic. 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 ones.

Many multimedia authentication systems have been proposed in the last few years which employ semi-fragile watermarks. A watermark is imperceptibly *embedded* in the multimedia signal to assist in verifying the integrity of the associated signal. The hidden watermark can be a hash value [1] or a set of coarser content features such as block histograms, or edge maps [2]. The primary advantage of employing semi-fragile watermarking over traditional digital signature authentication technology is that there is greater potential in characterizing the tamper distortion, and in designing a method which is robust to certain kinds of processing. In semi-fragile watermarking, the watermark must survive legitimate distortions, but be fully destroyed by illegitimate modifications applied to the signal. One of the first approaches to semi-fragile watermarking, called telltale tamper-proofing, was proposed by Kundur and Hatzinakos [3] to determine the extent of modification both in the spatial and frequency domains of a signal using a statistics-based tamper assessment function. Another influential semi-fragile system is the self-authentication-and-recovery image (SARI) method developed by Lin and Chang [4] in which a semi-fragile signature is embedded to survive JPEG compression up to a certain level.

E-mail: fei@control.toronto.edu, deepa@ee.tamu.edu, kwong@control.toronto.edu

Security, Steganography, and Watermarking of Multimedia Contents VIII, edited by Edward J. Delp III, Ping Wah Wong,
Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 6072, 607220, © 2006 SPIE-IS&T · 0277-786X/06/$15

There are two major challenges in multimedia authentication watermarking to distinguish legitimate distortions caused by incidental manipulations from those caused by illegitimate manipulations. One challenge is that there is typically no clear distinction boundary between authentic and inauthentic signals. In general, modifications which do not alter the content of the multimedia signal are considered to be legitimate. These include minor modifications such as high rate JPEG compression, image enhancement filtering, and even geometric distortions such as rotation, scaling and translation. Severe modifications such as low rate compression, image blurring filtering, and malicious image object removal or substitution are considered illegitimate. The other major difficulty to distinguish legitimate and illegitimate distortions is the fact that the original host is not available at the receiver side for verification. In practical applications, the original host generally has much larger magnitude than allowed legitimate channel distortions. The blindness of the original host in authentication verification makes it hard to differentiate legitimate distortions from illegitimate distortions. Most proposed schemes to date are either designed to achieve robustness to specific distortions (usually compression) using ad hoc measures, or by carefully tuning a robust watermarking scheme so that it is likely to be destroyed if the distortion exceeds a particular level [5]. Such schemes may attain desired robustness, but do not necessarily help provide fragility to illegitimate distortions. A well-designed semi-fragile system should simultaneously address the robustness and fragility objectives associated with legitimate and illegitimate distortions.

Our approach to the design of semi-fragile systems is to classify different types of common channel distortions as legitimate or illegitimate, and to embed a watermark effectively to distinguish legitimate distortions from illegitimate ones. This leads us naturally to consider the authentication verification procedure as a hypothesis testing problem with the watermark as *a priori* information. The receiver's best strategy is to identify legitimacy of the distortion channel using the a priori watermark information while the embedder's best embedding strategy is to assist the receiver to make correct decisions. The results of our hypothesis testing model show that the quantization-based watermarking method is better than the spread spectrum method to achieve the tradeoff between two error probabilities. Finally, we apply our hypothesis testing model to certain common signal processing distortions such as JPEG compression and image filtering, and determine the best test statistic and optimal decision regions to distinguish legitimate and illegitimate distortions.

The paper is organized as follows. We model authentication watermarking through a composite hypothesis testing model in Section 2. The best watermark embedding method to help the receiver to make correct decisions is discussed in Section 3. Section 4 provides a case study of additive Gaussian white noise channel with Gaussian source. Section 5 focuses on certain common signal processing distortions such as JPEG compression and image filtering, and derive the corresponding best statistics. Conclusions are drawn in Section 6.

2. SEMI-FRAGILE HYPOTHESIS TESTING MODEL

In this section, we describe our hypothesis testing approach, and show how semi-fragility can be characterized by error probabilities arising in a hypothesis testing model.

2.1. Authentication Watermarking Model

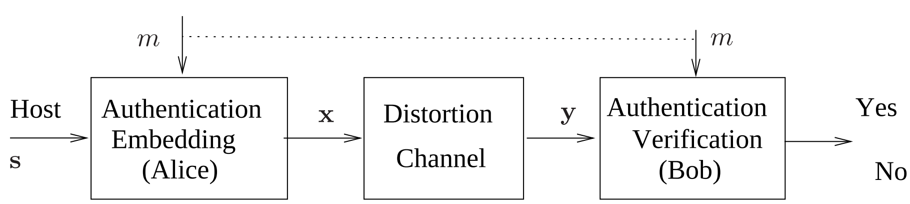


Figure 1. General authentication watermarking model.

We consider a general authentication watermarking system contains three components: an embedder (Alice), a distortion channel and the corresponding watermark verifier (Bob). In order for Bob to be assured that the signal did originate from Alice, Alice authenticates the host source $\mathbf{s}$ by embedding her watermark m to produce an authenticated signal $\mathbf{x}$ without introducing perceptible visual distortion. The watermarked signal is represented by $\mathbf{x} = f(\mathbf{s}, w)$ where $f(\cdot, \cdot)$ is the embedding function. The watermarked signal $\mathbf{x}$ is received by Bob through a public distortion channel. Knowing Alice's watermark m, the receiver tries to decide whether the received signal $\mathbf{y}$ is authentic or not by verifying the presence of Alice's watermark. A binary decision on authenticity is made based on the observed signal $\mathbf{y}$ and the sender's watermark m which the receiver knows prior to authentication verification.

2.2. Authentication Verification: A Composite Hypothesis Testing Approach

Possible distortion on the watermarked image is modelled as a distortion channel. We use $\mathbf{y} = \mathcal{P}(\mathbf{x})$ to denote a distortion channel, mapping the input $\mathbf{x}$ to the output $\mathbf{y}$. The behavior of the random channel $\mathcal{P}$ is characterized by its conditional probability density function (PDF) from channel input $\mathbf{x}$ to output $\mathbf{y}$. The random distortion channel includes deterministic modifications as a special case. Possible distortion channels are grouped into two categories: the legitimate channel set $\mathcal{L}_0$ and the illegitimate channel set $\mathcal{L}_1$. The legitimate channel set $\mathcal{L}_0$ may include minor modifications such as high rate JPEG compression, image enhancement filtering, and even unobtrusive rotation, scaling and translation 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.

At the receiver end, knowing the sender's watermark m, a watermark verification process attempts to identify the type of distortion channel based on the received signal $\mathbf{y}$, i.e. whether $\mathcal{P} \in \mathcal{L}_0$ or $\mathcal{L}_1$. Semi-fragile watermarking verification can therefore be viewed mathematically as a hypothesis testing problem to identify legitimacy of a channel. We have two composite hypotheses: the null hypothesis H_0 is $\mathcal{L}_0$ and the alternative hypothesis H_1 is $\mathcal{L}_1$. The authentication verification process needs to test the following composite hypothesis problem,

$$H_0 : \mathbf{y} = \mathcal{P}(\mathbf{x}) \text{ for } \mathcal{P} \in \mathcal{L}_0 \tag{1a}$$

$$H_1 : \mathbf{y} = \mathcal{P}(\mathbf{x}) \text{ for } \mathcal{P} \in \mathcal{L}_1 \tag{1b}$$

based on the observation $\mathbf{y}$ and the sender's watermark m. The watermark m represents a priori information to the receiver, and is used to help make a correct hypothesis testing decision. This a priori information can help the decision-making process because it is embedded in the watermarked signal $\mathbf{x}$, so partial information of $\mathbf{x}$ is revealed to the receiver through the side information. From the hypothesis testing model, we can see the fundamental difference of robust watermarking and semi-fragile watermarking. The hypothesis testing problem in robust watermarking is to determine which watermark message has been embedded in a host with channel noise as an interference. The hypothesis testing in semi-fragile watermarking is to identify legitimacy of the channel with the host signal as an interference. The embedding of a watermark, which is a priori information to the receiver, is to alleviate the interference of the host signal to channel differentiation.

With respect to the hypothesis testing model, there are two types of authentication errors in semi-fragile watermarking. 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 semi-fragile authentication system. 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 wrongly verified by the receiver as authentic. This type of authentication error characterizes the fragility of the semi-fragile system. Let A_n be the decision region which the receiver uses to verify authenticity of the received signal $\mathbf{y}$. Type I error probability is given by $\alpha_n(\mathcal{P}) = P[\mathbf{y} \notin A_n | H_0]$ for a legitimate channel $\mathcal{P} \in \mathcal{L}_0$, and type II error probability is given by $\beta_n(\mathcal{P}) = P[\mathbf{y} \in A_n | H_1]$ for an illegitimate channel $\mathcal{P} \in \mathcal{L}_1$.

Another interesting measure of two error probabilities is their asymptotic behavior. Both families of error probabilities should decrease to zero as the length n increases since more observation data $\mathbf{y}$ are available to make a decision. To evaluate how fast both families of error probabilities decrease as the dimension n increases, the error exponents of both errors, assuming the limits exist, are defined as follows [6], $E_\alpha(\mathcal{P}) = \lim_{n\to\infty} -\frac{1}{n} \ln \alpha_n(\mathcal{P})$ for a legitimate channel $\mathcal{P} \in \mathcal{L}_0$, and $E_\beta(\mathcal{P}) = \lim_{n\to\infty} -\frac{1}{n} \ln \beta_n(\mathcal{P})$ for $\mathcal{P} \in \mathcal{L}_1$.

2.3. Generalized Likelihood Ratio Test

The best decision region should give the best trade-off between two families of error probabilities under the Neyman-Pearson criterion. A common approach to the composite hypothesis testing is the generalized likelihood ratio test (GLRT) [7]. In the GLRT approach, the most probable individual hypothesis is computed to represent the likelihood of the composite hypothesis. The generalized likelihood ratio is defined as the ratio of the maximum value of the likelihood under H_0 to the maximum under H_1. For easy computation, the generalized log-likelihood ratio is used in which the length n is normalized. That is,

$$GLLR = \frac{1}{n} \log \sup_{\mathcal{P} \in \mathcal{L}_0} f(\mathbf{y}|\mathcal{P}, H_0, m) - \frac{1}{n} \log \sup_{\mathcal{P} \in \mathcal{L}_1} f(\mathbf{y}|\mathcal{P}, H_1, m) \tag{2}$$

where $f(\mathbf{y}|\mathcal{P}, H_i, m)$ is the likelihood of the received sequence $\mathbf{y} = [y_1, y_2, \ldots, y_n]$ under the two hypotheses with a known watermark m. Hypothesis H_0 is accepted if the test statistic $GLLR$ is greater than a given threshold T; otherwise, H_1 is accepted.

3. THE BEST EMBEDDING METHOD

After analyzing the receiver's best strategy to identify legitimacy of the distortion channel, we then investigate the embedder's best embedding strategy which assists the receiver to make correct decisions. There are mainly two classes of watermark embedding methods: spread spectrum method and quantization-based method. In this section, we provide an information-theoretic explanation of how the quantization-based embedding method allows the receiver to achieve the best trade-off between Type I and II error probabilities.

From the composite hypothesis testing model, we know that the watermarked signal is an interference to channel legitimacy identification, and the watermark m is a priori information related to the watermarked signal. Semi-fragile authentication generally involves two composite hypotheses, but we can view them as two single hypotheses based on the idea of the GLRT approach. The most probable individual hypothesis is computed to represent the likelihood of the composite hypothesis. Let $p_Y(y)$ and $q_Y(y)$ denote two probability distributions of two most probable single hypotheses corresponding to legitimate and illegitimate channels, respectively. A well-known result in hypothesis testing provides a relationship between the error probabilities α and β and the relative entropy $D(p_Y||q_Y)$. The Type I and Type II error probabilities satisfy the following inequality: $d(\alpha, \beta) \leq D(p_Y|q_Y)$ where the function $d(\alpha, \beta)$ is defined by $d(\alpha, \beta) = \alpha \log \frac{\alpha}{1-\beta} + (1-\alpha) \log \frac{1-\alpha}{\beta}$ [8] . In particular, for $\alpha = 0$, we have $\beta \geq 2^{-D(p_Y||q_Y)}$. In other words, $D(p_Y||q_Y)$ characterizes the error exponent of Type II error probabilities for $\alpha = 0$. The relative entropy $D(p_Y||q_Y)$ is a measure of performance bound of the hypothesis testing to differentiate two hypotheses with emphasis on Type II error probability. Therefore, we use the relative entropy $D(p_Y||q_Y)$ as a figure of merit to measure the receiver's capability to identify authenticity of a received signal. Although relative entropy is not a true distance metric because it is not symmetric and does not satisfy the triangle inequality, it can be useful to think of it as a distance. In the authentication watermarking model depicted in Fig. 1, let S, M, X and Y be the random variables corresponding to the source, the watermark, the channel input and output, respectively. The figure of merit of the hypothesis testing for semi-fragile authentication verification is the conditional relative entropy $D(p_{Y|M}||q_{Y|M})$ since the watermark M is known at the authentication verification process.

There are two extreme scenarios that provide performance bounds for watermark-based authentication. In one extreme scenario that the host signal is fully known to the receiver, the receiver can make the best judgement on the legitimacy of a test signal. Non-blind authentication is not practical for most authentication applications as the original signal is not always available. We consider this ideal scenario because it gives a upper bound of performance for authentication watermarking applications. Errors in non-blind authentication are merely due to the fuzzy boundary between legitimate and illegitimate distortions and one cannot eliminate such errors through watermarking process. When the channel input X is known, the figure of merit associated with two hypotheses is $D(p_{Y|X}||q_{Y|X})$ where X is the channel input random variable. The other extreme scenario is that no watermark information is used for authentication verification. Hypothesis test decision is made only based on the received signal, without any help of the a priori watermark. This scenario corresponds to the worst case of semi-fragility verification which provides a lower bound of performance for authentication watermarking applications. In this scenario, the channel input X itself serves as an interference to channel distortion hypothesis testing since no information about X is available at the receiver through the a priori watermark. The figure of merit associated with two channel distributions is $D(p_Y||q_Y)$ which is only based on the distribution of the channel output Y. From information theory, it can be shown that for any watermark embedding, $D(p_{Y|X}||q_{Y|X}) \geq D(p_{Y|M}||q_{Y|M}) \geq D(p_Y||q_Y)$. This result confirms our intuition about how a priori information helps to alleviate interference from the host signal.

In general authentication watermarking schemes, the watermark should be embedded such that $D(p_{Y|M}||q_{Y|M})$ is minimized over possible embedding functions. This optimization problem is very complex to solve since the embedding function should also satisfy an embedding distortion constraint. Here, we give an intuitive explanation of how a good embedding function helps channel differentiation. Since the channel input X serves as an interference to channel distortion differentiation, the watermark should be embedded to reduce the degree of the interference of X to help the receiver identify the distortion channel correctly. The more random the signal X is, the more difficult to distinguish between two hypothesis distributions $p_{Y|M}$ and $q_{Y|M}$. Therefore, one would like to reduce the uncertainty of X conditioned on M. In other words, the conditional entropy $H(X|M)$ should be minimized in order to reduce its interference to channel differentiation. From information theory, we have $H(X|M) = H(M|X) + H(X) - H(M)$. To minimize $H(X|M)$, we therefore should do the following

- minimize $H(M|X)$; To achieve this, the watermark M should be uniquely determined for a given watermarked signal X. In quantization-based schemes, different watermarks are represented by different quantizers, so the embedded watermark is uniquely determined from the quantized signal X. Therefore, quantization-based schemes have $H(M|X) = 0$.
- minimize $H(X)$; The entropy of X is reduced if X is quantized after watermark embedding. A larger quantization step will result in less entropy of X. However, a larger quantization step will also result in larger embedding distortion D. Therefore, there is a tradeoff in determining the quantization step.
- maximize $H(M)$; In other word, the watermark should be uniformly distributed.

In spread spectrum watermarking, a watermark-related spread sequence $W(M)$ is embedded in the original host S, so the watermarked signal $X = S + W(M)$. Therefore, $H(X|M) = H(S)$. The original host serves as an interference to channel hypothesis testing. Such case is equivalent to the worst case that no watermark is embedded and used for verification since the known watermark M does not give any help to reduce the interference from the host signal.

From the above analysis, we see that the quantization-based method embeds the watermark by quantizing the source, thus reduces the interference of the watermarked signal X to distortion channel differentiation. Therefore, the quantization-based embedding method is better than the spread spectrum method in achieving semi-fragility of multimedia authentication.

4. ANALYSIS OF AWGN CHANNELS WITH A GAUSSIAN SOURCE

To support the conclusion of the superiority of quantization-based embedding, we analyze in this section a simple case of AWGN channels with a Gaussian source. The legitimacy of an AWGN channel is specified as follows. An AWGN channel is legitimate if its variance $\sigma^2 < a$ for a constant a, and illegitimate if $\sigma^2 > b$ for a constant $b \geq a$. We also assume that the host signal is Gaussian distributed with zero mean and variance σ_s^2. We use the generalized likelihood ratio test to derive the optimal decision region for three different schemes: non-blind, spread spectrum, and quantization-index-modulation (QIM). These methods are assessed and compared using the relative entropy between two hypothesis distributions as well as the receiver operating characteristic (ROC).

4.1. Non-blind Authentication

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

$$H_0: \quad \mathbf{y} = \mathbf{x} + \mathbf{z}(\sigma^2) \text{ for } \sigma^2 < a \tag{3a}$$

$$H_1: \quad \mathbf{y} = \mathbf{x} + \mathbf{z}(\sigma^2) \text{ for } \sigma^2 > b, \tag{3b}$$

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 using the following generalized log-likelihood ratio test $GLLR = \frac{1}{n} \log \sup_{\sigma^2 < a} f(\mathbf{z}, \sigma^2) - \frac{1}{n} \log \sup_{\sigma^2 > b} f(\mathbf{z}, \sigma^2) > T$ for some threshold T. Here the likelihood function $f(\mathbf{z}, \sigma^2)$ for Gaussian noise $\mathbf{z}$ with variance σ^2 is given by $f(\mathbf{z}, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{z_i^2}{2\sigma^2}) = \Big(2\pi\sigma^2 \exp(\bar{\mathbf{z}^2}/\sigma^2)\Big)^{-n/2}$, where $\bar{\mathbf{z}^2} = \frac{1}{n}\|\mathbf{z}\|^2$. For a fixed $\bar{\mathbf{z}^2}$, $2\pi\sigma^2 \exp(\bar{\mathbf{z}^2}/\sigma^2)$ is minimized at $\sigma^2 = \bar{\mathbf{z}^2}$. Then, the generalized log-likelihood ratio is given by

$$GLLR = \begin{cases} \frac{1}{2}(\frac{\bar{\mathbf{z}^2}}{b} - \log \bar{\mathbf{z}^2} + \log b - 1) & \text{if } \bar{\mathbf{z}^2} < a \\ \frac{1}{2}(\frac{\bar{\mathbf{z}^2}}{b} - \frac{\bar{\mathbf{z}^2}}{a} + \log b - \log a) & \text{if } a \le \bar{\mathbf{z}^2} \le b \\ \frac{1}{2}(\log \bar{\mathbf{z}^2} - \frac{\bar{\mathbf{z}^2}}{a} + 1 - \log a) & \text{otherwise,} \end{cases} \tag{4}$$

which is a strictly deceasing function of $\bar{\mathbf{z}^2}$. The decision test $GLLR > T$ can be simplified to $\bar{\mathbf{z}^2} = \frac{1}{n}\|\mathbf{z}\|^2 < r^2$ for some positive constant r^2, which is related to T, and the parameters a and b. Therefore, the optimal decision region A_n is given by

$$\|\mathbf{y} - \mathbf{x}\|^2 \le nr^2 \tag{5}$$

for some constant r^2. For a detailed derivation of the optimal decision region, refer to [9].

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 given by $\alpha_n(\sigma^2) = P\big[\mathbf{y} \in A_n^c\big] = P\big[\chi^2(n) > \frac{nr^2}{\sigma^2}\big]$ for $\sigma^2 < a$, $\beta_n(\sigma^2) = P\big[\mathbf{y} \in A_n\big] = P\big[\chi^2(n) < \frac{nr^2}{\sigma^2}\big]$ for $\sigma^2 > b$, respectively, where $\chi^2(n)$ denotes chi-square distribution with degree n, and A_n^c is the complement set of A_n.

When $a \le r^2 \le b$, the Type I and II error probabilities decrease to zero as the length n increases, so their error exponents exist. From large deviation theory, the error exponent function $I(x)$ for χ^2 distribution function is $I(x) = 0.5(x - \ln x - 1)$ [10]. Therefore, the Type I and II error exponents are given by $E_\alpha(\sigma^2) = 0.5(\frac{r^2}{\sigma^2} - \ln\frac{r^2}{\sigma^2} - 1)$ for $\sigma^2 < a$, and $E_\beta(\sigma^2) = 0.5(\frac{r^2}{\sigma^2} - \ln\frac{r^2}{\sigma^2} - 1)$ for $\sigma^2 > b$.

4.2. Spread Spectrum Scheme

The embedding function for spread spectrum scheme is given by $\mathbf{x} = \mathbf{s} + \mathbf{w}(m)$ where $\mathbf{w}(m)$ is an additive watermark signal related to watermark message m. The verification procedure for the spread spectrum embedding is the following hypothesis testing

$$H_0: \quad \mathbf{y} = \mathbf{s} + \mathbf{w}(m) + \mathbf{z}(\sigma^2) \text{ for } \sigma^2 < a \tag{6a}$$

$$H_1: \quad \mathbf{y} = \mathbf{s} + \mathbf{w}(m) + \mathbf{z}(\sigma^2) \text{ for } \sigma^2 > b. \tag{6b}$$

The receiver knows the watermark m, thus the spread spectrum signal $\mathbf{w}(m)$. However, the original signal $\mathbf{s}$ is not known to the receiver, thus serves as noise to the hypothesis testing of two channels. Using a similar procedure of the generalized log-likelihood ratio test, the optimal decision region A_n is given by $\|\mathbf{y} - \mathbf{w}(k)\|^2 < nr^2$ for some positive constant r^2. This optimal decision criterion gives the best authentication verification structure for spread spectrum watermarking, which is a distance detector to the embedded watermark. By contrast, in robust watermarking for AWGN channels with a Gaussian source, a correlation detector is the best structure to test the existence of the embedded watermark. We can see from this example the distinct nature of robust watermarking and semi-fragile watermarking for authentication. The best detector for robust watermarking may not be good for authentication watermarking in terms of semi-fragility characterized by two types of error probabilities.

Given the best decision region, the Type I and II error probabilities are given by $\alpha_n(\sigma^2) = P\big[\mathbf{y} \in A_n^c\big] = P\big[\chi^2(n) > \frac{nr^2}{(\sigma_s^2 + \sigma^2)}\big]$ for $\sigma^2 < a$, and $\beta_n(\sigma^2) = P\big[\mathbf{y} \in A_n\big] = P\big[\chi^2(n) < \frac{nr^2}{(\sigma_s^2 + \sigma^2)}\big]$ for $\sigma^2 < b$, 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 above results, we can see that the additive spread spectrum signal $\mathbf{w}(k)$ does not help tradeoff two error probabilities. One would get the same results if no signal $\mathbf{w}(k)$ is embedded. This observation confirms our intuitive explanation of the spread spectrum method in Section 3.

When the constant r^2 is chosen such that $\sigma_s^2 + a \le r^2 \le \sigma_s^2 + b$, both error probabilities asymptotically decay to zero as n approaches infinite. The Type I and II error exponents exist and are given by $E_\alpha(\sigma^2) = 0.5(\frac{r^2}{\sigma_s^2 + \sigma^2} - \ln\frac{r^2}{\sigma_s^2 + \sigma^2} - 1)$ for $\sigma^2 < a$, and $E_\beta(\sigma^2) = 0.5(\frac{r^2}{\sigma_s^2 + \sigma^2} - \ln\frac{r^2}{\sigma_s^2 + \sigma^2} - 1)$ for $\sigma^2 > b$, respectively.

4.3. Quantization-based Embedding

In quantization-based schemes, a watermark m is embedded by quantizing the host $\mathbf{s}$ using a quantization function associated with the watermark. The embedding function is described as follows, $\mathbf{x} = Q(\mathbf{s}, m)$ where $Q(\cdot, m)$ is the quantization function corresponding the watermark m. Let $\mathcal{C}(m)$ be the reconstruction point set of the quantizer associated with the watermark m. Then $\mathbf{x}$ is the quantized value of $\mathbf{s}$ using a nearest neighbor quantizer associated with $\mathcal{C}(m)$. The authenticated signal $\mathbf{x}$ is discretely distributed over the code set $\mathcal{C}(m)$. Its probability distribution $p(\mathbf{x}|m)$ for $\mathbf{x} \in \mathcal{C}(m)$ conditioned on watermark m can be derived from the distribution of the source $\mathbf{s}$ as follow, $p(\mathbf{x}|m) = P[\mathbf{X} = x|m] = P[Q(\mathbf{s}, m) = \mathbf{x}|m] = P[\mathbf{s} \in \mathcal{V}(\mathbf{x})] = \int_{\mathcal{V}(\mathbf{x})} f(\mathbf{s}) d\mathbf{s}$ where $\mathcal{V}(x)$ is the Voronoi region around $\mathbf{x}$ associated with $\mathcal{C}(m)$ and $f(\mathbf{s})$ is the PDF of the host signal $\mathbf{s}$.

The composite hypothesis testing problem for a quantization-based embedding scheme becomes the following:

$$H_0 : \quad \mathbf{y} = \mathbf{x} + \mathbf{z}(\sigma^2) \text{ for } \sigma^2 < a \tag{7a}$$

$$H_1 : \quad \mathbf{y} = \mathbf{x} + \mathbf{z}(\sigma^2) \text{ for } \sigma^2 > a \tag{7b}$$

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

$$GLLT = \frac{1}{n} \log \sup_{\sigma^2 < a} \sum_{\mathbf{x} \in \mathcal{C}(m)} p(\mathbf{x}) f(\mathbf{y} - \mathbf{x}, \sigma^2) - \frac{1}{n} \log \sup_{\sigma^2 > b} \sum_{\mathbf{x} \in \mathcal{C}(m)} p(\mathbf{x}) f(\mathbf{y} - \mathbf{x}, \sigma^2) > T \tag{8}$$

for some constant T. The above test statistic needs to find the optimal solution of σ^2 to maximize the summation of a weighted likelihood. It is hard to find an explicit form for the solution of σ^2. Since the term $f(\mathbf{y} - \mathbf{x}, \sigma^2) = \left(2\pi\sigma^2 \exp(\frac{1}{n}\|\mathbf{y} - \mathbf{x}\|^2/\sigma^2)\right)^{-n/2}$ is a deceasing function of $\|\mathbf{y} - \mathbf{x}\|$, the codeword closest to $\mathbf{y}$ has the smallest distance $\frac{1}{n}\|\mathbf{y} - \mathbf{x}\|$, thus is the dominant term in the summation, especially for large n. Therefore, we use the dominant term to approximate the test statistic. Because of the flat shape of the distribution of $p(\mathbf{x})$ over $\mathcal{C}(m)$ and the dominance of the closest codeword, this approximation leads to a suboptimal solution which is very close to the optimal one. Let $\mathbf{x}_c$ is the closest codeword in $\mathcal{C}(m)$ to the received signal $\mathbf{y}$. The generalized likelihood ratio test is simplified to the following by just using the dominant term of $\mathbf{x}_c$, $\frac{1}{n} \log \sup_{\sigma^2 < a} p(\mathbf{x}_c) f(\mathbf{y} - \mathbf{x}_c, \sigma^2) - \frac{1}{n} \log \sup_{\sigma^2 > b} p(\mathbf{x}_c) f(\mathbf{y} - \mathbf{x}_c) > T$. Using a similar calculation in the non-blind authentication case, the decision region is obtained as follows $\|\mathbf{y} - \mathbf{x}_c\| < nr^2$ for some positive constant r^2. For QIM schemes in which $\mathcal{C}(m)$ is a dithered uniform quantizer, the closest codeword $\mathbf{x}_c$ to $\mathbf{y}$ is given by $\mathbf{x}_c = Q(\mathbf{y} - d(m)) + d(m)$. The decision region for the QIM scheme is illustrated in Fig. 2. The decision region can be represented by $A_n = \mathcal{C}(m) + Q_n$ where Q_n is an n-dimensional sphere with radius r, i.e. $Q_n = \{\mathbf{z} \in \mathbb{R}^n | \|\mathbf{z}\|^2 < nr^2\}$.

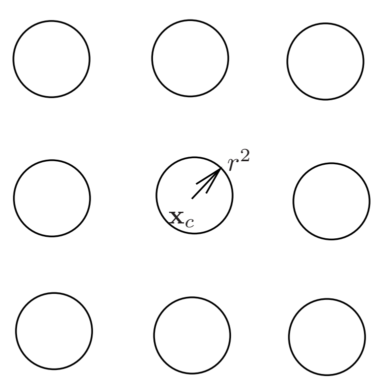


Figure 2. The decision region for the QIM scheme contains blocks around the reconstruction set of the quantizer associated with the watermark m.

Now we compute Type I and II error probabilities associated with the above derived decision region. The Type I error probability for a legitimate noise with variance $\sigma^2 < a$ is given by $\alpha_n(\sigma^2) = P[\mathbf{y} \notin A_n] = P[\mathbf{x} + \mathbf{z}(\sigma^2) \notin \mathcal{C}(m) + Q] = P[\mathbf{z}(\sigma^2) \notin (\mathcal{C}(m) - \mathbf{x}) + Q_n]$, which is the probability that $\mathbf{z}(\sigma^2)$ is not in any of the blocks in Fig. 2. This error probability is less than the probability that $\mathbf{z}(\sigma^2)$ is just not in the closest block around $\mathbf{x}_c$. So we have $\alpha_n(\sigma^2) \le P[\mathbf{z}(\sigma^2) \notin Q_n] = P\big[\chi^2(n) > \frac{nr^2}{\sigma^2}\big]$. When the constant $r^2 > a$, the above error probability decays to zero as n approaches infinite for all $\sigma^2 < a$, so the Type I error exponent exists and is lower bounded as follows, $E_\alpha(\sigma^2) \ge 0.5(\frac{r^2}{\sigma^2} - \ln\frac{r^2}{\sigma^2} - 1)$ for $\sigma^2 < a$.

The Type II error probability for an illegitimate noise with variance $\sigma^2 > b$ is given by $\beta_n(\sigma^2) = P[\mathbf{y} \in A_n] = P[\mathbf{x}+\mathbf{z}(\sigma^2) \in \mathcal{C}(m)+Q_n] = P\Big[\mathbf{z}(\sigma^2) \in \sum_{\mathbf{x}_c \in \mathcal{C}(m)}(\mathbf{x}_c - \mathbf{x} + Q_n)\Big]$, where $(\mathbf{x}_c - \mathbf{x}) + Q_n$ is the decision block around $\mathbf{x}_c - \mathbf{x}$ for $\mathbf{x}, \mathbf{x}_c \in \mathcal{C}(m)$ as shown in Fig. 2. For QIM schemes where the encoding set is a dithered version of a base quantizer Λ, the above Type II probability is represented by $\beta_n(\sigma^2) = \sum_{\boldsymbol{\lambda} \in \Lambda} P\Big[\mathbf{z}(\sigma^2) \in \mathcal{V}(\boldsymbol{\lambda}) \cap Q_n\Big]$ where $\mathcal{V}(\boldsymbol{\lambda})$ denotes the Voronoi region around $\boldsymbol{\lambda}$. Given the Type II error probability represented in a summation over all decision blocks, the Type II error exponent is difficult to have a explicit form. We give an approximation here. When n is sufficiently large, from information theory, the Gaussian sequence $\mathbf{z}(\sigma^2)$ is uniformly distributed in its typical set, which is an n-dimensional ball with a radius of $\sqrt{n\sigma^2}$. When σ^2 is small enough that the error $\mathbf{z}(\sigma^2)$ only falls in the fundamental Voronoi block $\mathcal{V}_0$ corresponding to the origin, so $\beta_n(\sigma^2) = P\Big[\mathbf{z}(\sigma^2) \in \mathcal{V}_0 \cap Q_n\Big] \le P\big[\chi^2(n) < \frac{nr^2}{\sigma^2}\big]$, and the corresponding error exponent is given by $E_\beta(\sigma^2) = 0.5(\frac{r^2}{\sigma^2} - \ln\frac{r^2}{\sigma^2} - 1)$ for $\sigma^2 > b$ and $r^2 < b$. When σ^2 is large, the errors that $\mathbf{z}(\sigma^2)$ falls in other blocks cannot be ignored. Since for large n, $\mathbf{z}(\sigma^2)$ is uniformly distributed, the Type II error probability is therefore approximately equal to the volume ratio of the decision region to the whole signal space. This volume ratio is given by

$$\beta_n(\sigma^2) \approx \frac{\mathrm{Vol}(\Omega_n \cap \mathcal{V}_0)}{\mathrm{Vol}(\mathcal{V}_0)} \le \frac{\mathrm{Vol}(\Omega_n)}{\mathrm{Vol}(\mathcal{V}_0)} = \left(\frac{r^2 G_n(\Lambda)}{D G_n(B)}\right)^{n/2}, \tag{9}$$

where $G_n(B)$ and $G(\Lambda)$ are the normalized second moments of the n-dimensional ball and the base lattice Λ for $\mathcal{C}(m)$, respectively [11]. $G_n(B)$ converges to $1/2\pi e$ as $n \to \infty$ and $G(\Lambda) \ge 1/2\pi e$ [12]. Therefore, the error exponent $E_\beta(\sigma^2) \ge \frac{1}{2}\log(\frac{D}{r^2}) - \log(2\pi e G(\Lambda))$. If the lattice Λ is also a "good" lattice, $G(\Lambda) = 1/2\pi e$, then $E_\beta(\sigma^2) \ge \frac{1}{2}\log(\frac{D}{r^2})$.

4.4. Comparison Results

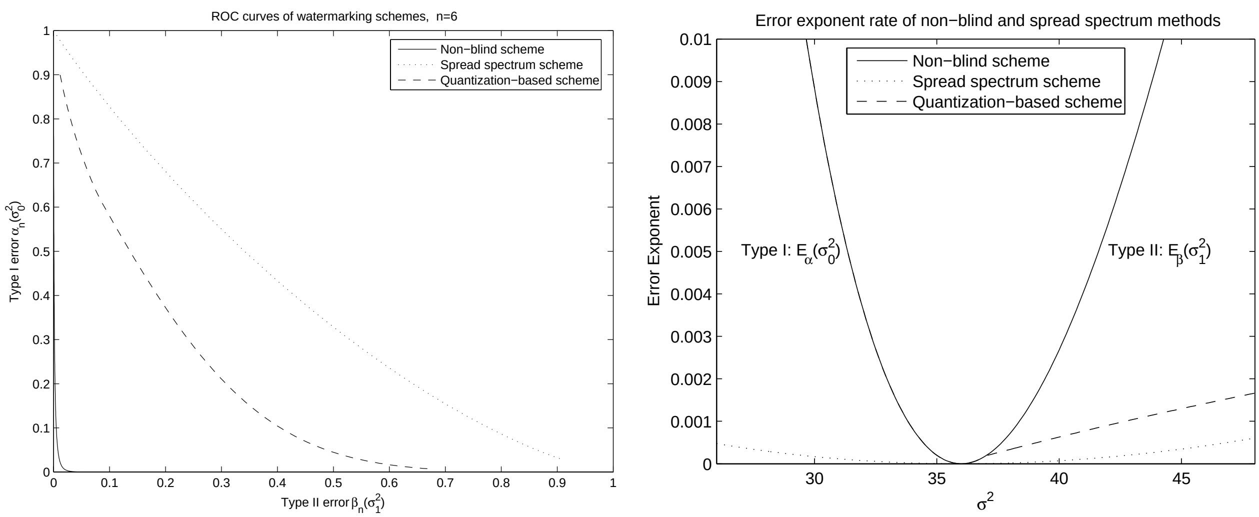
In this section, we compares three scenarios by computing the relative entropy between two hypothesis distributions and two families of error probabilities using the generalized likelihood ratio test. We assume a Gaussian host $\mathbf{s}$ with variance $\sigma_s^2 = 200$. In our simulation, we set $a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

	$\sigma_1^2 = 64$			$\sigma_1^2 = 100$		
	NB	SS	QIM	NB	SS	QIM
$\sigma_0^2 = 4$	1.3237	0.0220	0.1548	1.6294	0.0474	0.1792
$\sigma_0^2 = 10$	0.7304	0.0175	0.0208	1.0118	0.0409	0.0434

Table 1. The relative entropy between legitimate and illegitimate AWGN channels for non-blind (NB), spread spectrum (SS) and quantization-index-modulation (QIM) schemes. The host variance $\sigma_s^2 = 200$. The values of the relative entropy for QIM are numerically computed due to quantization.

First, we compute the relative entropy $D(p_{Y|M}||q_{Y|M})$ between an legitimate channel $p_{Y|X}(y)$ and an illegitimate channel $q_{Y|X}(y)$ in three 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 $\sigma_1^2 = 100$. The relative entropy $D(p_{Y|M}||q_{Y|M})$ for different legitimate and illegitimate channels are shown in Table 1 in which the QIM scheme uses scalar uniform quantizer with step size 8. The values of the relative entropy in the table show that the ideal non-blind scenario expectedly has the largest values, and spread spectrum has the smallest values. The quantization-based scheme achieves relative entropy greater than the spread spectrum scheme. This results show that the quantization-based scheme can achieve better balance between the two types of error probabilities than the spread spectrum scheme.

We also compute two families of error probabilities associated with best decision regions with various thresholds. A common approach in the assessment of hypothesis testing scheme is the receiver operating characteristic (ROC) curves. The ROC curve is a curve of Type I error probability vs. 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 a 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, thus a ROC curve is obtained as $\alpha_n(\sigma_0^2)$ vs. $\beta_n(\sigma_1^2)$ as the threshold parameter T or r^2 varies.



(a) ROC. (b) Error exponents.

Figure 3. The ROC curves when $n = 6$ and the error exponent curves for non-blind, SS and QB methods.

Fig. 3(a) shows the ROC curves of different schemes for sequence length $n = 6$. We can see from the figure that the spread spectrum scheme is the worst scheme, and the ideal non-blind scheme is the best. The quantization-based scheme is worse than the non-blind one, but better than the spread spectrum scheme. This simulation results confirm our analysis that the signal $\mathbf{x}$ is localized in quantization-based schemes, thus interfering less in the channel identification of the two types. Fig. 3(b) illustrates the error exponent curves as the noise variance σ^2 varies for the non-blind, the spread spectrum, and the quantization-based schemes. In order that the corresponding error exponents exist, the threshold r^2 is set to be $\sigma_s^2 + a$ for the spread spectrum scheme, and a for the non-bind and quantization based schemes. We see in the figure that the non-blind scheme has much larger error exponents than the spread spectrum scheme since the host signal has a significant impact on the differentiation of two channel sets. The quantization-based scheme has the same error exponent as the non-blind case when the noise variance σ^2 is small. However, when σ^2 increases, the Type II error exponent approaches to a constant related to the embedding distortion D. These simulation results show that the quantization-based method achieves significant improvement over the spread spectrum method in the ability to distinguish the legitimacy of a distortion channel.

5. COMMON IMAGE PROCESSING ATTACKS

Our hypothesis testing approach on the case study of AWGN noise with Gaussian source confirms our intuitive explanation that the quantization-based method allows the best tradeoff in semi-fragility. 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 result 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 our work.

In the quantization-based scheme, a watermark is embedded by quantizing the host. The structure of the quantizer should tradeoff among semi-fragility, embedding distortion, and security [13]. 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 legitimacy of a 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 accidental changes. Based on the hypothesis testing model, we examine the relevant test statistic for specific distortions: JPEG compression, filtering, and geometric distortions.

5.1. JPEG Compression

JPEG compression is the most common incidental modification since most images are stored and transmitted in compressed format to save storage and bandwidth. Therefore, many watermarking systems have been proposed to be semi-fragile to JPEG compression up to a certain degree [3, 4, 14]. They all utilize a common property of uniform scalar quantizer that the quantization error due to JPEG compression in DCT domain is bounded in the range of $[-\frac{\Delta}{2}, \frac{\Delta}{2}]$ where Δ is the quantization step for compression quantization. However, in these systems, fragility against illegitimate JPEG compression is not investigated. The Type II authentication error probability cannot be eliminated since illegitimate severe quantization may still result in small quantization error in the detection region. In this section, we view quantization noise as uniformly distributed signal, analyze both error probabilities, and derived the test statistic to identify legitimacy of the JPEG compression distortion channel.

Given the watermarked image coefficients $\mathbf{x}$ in 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 compressed signal. The quantization error is defined as $\mathbf{z}(\Delta) = \mathbf{y} - \mathbf{x} = Q_\Delta(\mathbf{x}) - x$. For JPEG compression attack, high quality factor down to certain level is regarded as legitimate but low quality factor is illegitimate. The composite hypothesis testing problem for JPEG compression is described as the following,

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

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

Suppose $\mathbf{x}$ is known, we need to make a decision on the hypotheses based on the observed quantization error $\mathbf{z}$. Often the channel input $\mathbf{x}$ has larger variance than the quantization step, so $\mathbf{z}$ can be assumed to be uniformly distributed in the range $[-\frac{\Delta}{2}, \frac{\Delta}{2}]$ for high rate quantization.

The decision region is obtained if $GLLR = \frac{1}{n}\log\sup_{\Delta<a} f_Z(\mathbf{z}, \Delta) - \frac{1}{n}\log\sup_{\Delta>b} f_Z(\mathbf{z}, \Delta) > T$ for some threshold T, where the likelihood of uniformly distributed signal is given by $f_Z(\mathbf{z}, \Delta) = (\frac{1}{\Delta})^n$ if $\max_i |z_i| < \frac{\Delta}{2}$, and 0 otherwise. The optimal decision A_n from GLLR can be simplified to $\max_i |z_i| < \frac{1}{2}\Delta_T$ where $\Delta_T = \min\{a, be^{-T}\}$ [9]. The Type I error probability $\alpha_n(\Delta) = 1 - \left(\frac{\Delta_T}{\Delta}\right)^n$ for $\Delta > \Delta_T$, and 0 otherwise when $\Delta > a$. The Type II error probability $\beta_n(\Delta) = \left(\frac{\Delta_T}{\Delta}\right)^n$ for $\Delta_T < a$ when $\Delta > b$. The above hypothesis testing results show that from the observed quantized noise $\mathbf{z}$, the test statistic $2\max_i |z_i|$ is an estimate of Δ, therefore measures the degree of JPEG compression. Legitimacy of the JPEG compression attacks should be judged by comparing the test statistic with a threshold.

5.2. Image Filtering

The objective of image filtering is to remove noise from an image, but still keep a good visual quality of the image. Real images have energy concentrated in low frequency but noises often occur in high frequency. Therefore, image filtering is always a low pass filter to filter out high frequency components. For filtering, it is better to represent the image signals in the frequency domain to investigate the effects of filtering distortions. In frequency domain, we have $Y(U, V) = X(U, V)H(U, V)$ 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 additive model by taking a logarithm on the magnitude of the frequency as follows, $\log|Y(U, V)| = \log|X(U, V)| + \log|H(U, V)|$. The additive filtering distortion item $\log|H(U, V)|$ is small for smooth filtering, and large for severe filtering, we then can apply our approach to semi-fragile watermarking to detect the degree or legitimacy of filtering distortions. The idea here is to apply quantization-based watermarking in signal $\log|X(U,V)|$ at frequency band (U,V) such that the severity of the filtering can be measured from the quantization error of the test signal $\log|Y(U,V)|$.

To better illustrate the idea, we use a Gaussian filter as an example and control its degree of degradation using a single parameter. Gaussian filter is a linear filter whose low-pass filter curve is a Gaussian function with a single degradation parameter. One advantage of Gaussian filters is that there are no sidelobes in both spatial and frequency domain. The frequency spectrum of a Gaussian filter is approximated by [15] $H(U,V) \approx e^{-2\pi^2\sigma^2(U^2+V^2)}$ for $|U|,|V| < 1/2$, which is also a Gaussian function. The parameter σ^2 controls the shapes of frequency responses, which represents the degree of filtering degradation. Since large σ^2 may blur edges, low values of σ^2 is preferred when applying a Gaussian filer to remove noises. Therefore, we assume Gaussian filters are legitimate when $\sigma^2 < a$, but legitimate when $\sigma^2 > b$ where a, b are two positive constant and $b \geq a$. The composite hypothesis testing problem of Gaussian filtering is described as the following,

$$H_0: \quad \log|Y(U,V)| = \log|X(U,V)| - 2\pi^2\sigma^2(U^2+V^2) \text{ for } \sigma^2 < a \tag{11a}$$

$$H_1: \quad \log|Y(U,V)| = \log|X(U,V)| - 2\pi^2\sigma^2(U^2+V^2) \text{ for } \sigma^2 > b. \tag{11b}$$

We apply a QIM scheme to embed a watermark by quantizing the signal $\log|X(U,V)|$. In frequency band (U,V), the dithered quantizer set $C(m)$ is designed to be those of $X(U,V)$ satisfying $\log|X(U,V)| = \Delta(U,V)(i+d(m))$ for some integer i and a given dither value $d(m)$ where $\Delta(U,V)$ is the quantization step size in frequency (U,V) and $\Delta(U,V) > 2\pi^2(U^2+V^2)a$. After applying a quantization scheme, the quantized signal $\log|X(U,V)|$ can be recovered from $\log|Y(U,V)|$ under legitimate filtering. We then can estimate σ^2 from $\log|Y(U,V)|$ and the recovered $\log|X(U,V)|$ as follows,

$$\hat{\sigma}^2(U,V) = \frac{\log|X(U,V)| - \log|Y(U,V)|}{2\pi^2(U^2+V^2)}. \tag{12}$$

Legitimacy decision is made from the estimated degree of degradation $\hat{\sigma}^2$ over all frequency (U,V).

5.3. Geometric Distortions

Geometric distortions are the most complex one among all possible distortions [16]. In our analysis so far, we consider a value-metric model that uses the magnitude of additive changes to determine their severity. Geometric distortion instead does not change the value of the input. Rather it eliminates the synchronization between the input and the output. For images, geometric distortions include rotation, scaling and transformation.

The legitimacy of geometric distortions depends on specific applications. In applications where any geometric distortion is not acceptable, geometric distortions belong to the illegitimate set. Traditional quantization-based schemes still work since geometric distortion, like malicious tampering, pushes the quantized signal out of the detection region, so the distorted images are detected as illegitimate with high probability. For some image applications, global geometric distortions are considered acceptable since they do not change the visual meaning of images. In these cases, geometric distortions belong to the legitimate set. In robust watermarking literature, several watermarking systems resilient to geometric distortions have been proposed in which watermarking takes place in some transform domain which is invariant to rotation, scaling and translation (RST) [17]. The same approach can be employed in authentication watermarking. Authentication embedding and verification take place in these RST-invariant domains where geometric distortion does not change the image component.

The most complicated specification of geometric distortion is that small amount of rotation, scaling or transformation produces a similar image, thus is legitimate. However, large amount of rotation, scaling or transformation will change placement of image objects, so is illegitimate. This specification is reasonable for some practical applications such as digital checks or digital medical images. For such specifications on geometric distortions, we can apply a similar idea of value-metric model that the embedded watermark should retain all geometric information of the watermarked image. Then any geometric distortion can be assessed by comparing the geometrically distorted signal with the original watermark. To achieve this, we can embed a periodic reference watermark pattern on the entire image using the QIM scheme. At the receiver end, the embedded reference watermark is extracted, and checked with the original reference pattern to estimate possible geometric distortion and judge if the geometric distortion is acceptable or not. To check the extracted watermark with the original one, one has to do an exhaustive search on all possible legitimate RST distortions. Since only small amount of geometric distortion is allowed, such exhaustive search is feasible.

6. CONCLUSIONS

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

Practical semi-fragile authentication applications may involve several composite attacks. For example, the legitimate set contains minor additive Gaussian noise, high quality compression, and mild image filtering while the illegitimate set includes severe Gaussian noise, low quality compression, severe filtering, and object substitution. For these types of composite sets, it is much harder to distinguish them, even in the non-blind case. Such challenge is due to the nature of multimedia perception and classification, which is beyond the scope of authentication watermarking. Further work will consider multimedia perception for authentication purposes.

REFERENCES

1. P. W. Wong, "A public key watermark for image verification and authentication," in *Proc. IEEE Int. Conf. on Image Processing*, **I**, May 98.
2. J. Dittmann, A. Steinmetz, and R. Steinmetz, "Content-based digital signature for motion pictures authentication and content-fragile watermarking," in *Proc. IEEE Int. Conf. Multimedia Computing and Systems*, **2**, pp. 209–213, 1999.
3. D. Kundur and D. Hatzinakos, "Digital watermarking for telltale tamper-proofing and authentication," *Proc. IEEE* **87**, pp. 1167–1180, July 1999.
4. 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.
5. I. J. Cox, M. L. Miller, and J. A. Bloom, *Digital watermarking*, Morgan Kaufmann Publishers, 2002.
6. T. Liu and P. Moulin, "Error exponents for one-bit watermarking," in *Proc. ICASSP '03*, Apr. 2003.
7. S. M. Kay, *Fundamentals of Statistical Signal Processing: Detection Theory*, Prentice Hall, NJ, 1993.
8. U. M. Maurer, "Authentication theory and hypothesis testing," *IEEE Trans. Inform. Theory* **46**, pp. 1350–1356, July 2000.
9. C. Fei, *Analysis and Design of Watermark-based Multimedia Authentication Systems.* PhD thesis, University of Toronto, Toronto, Ontario, Canada, 2006.
10. R. R. Bahadur, *Some limit theorems in statistics*, Society for Industrial and Applied Mathematics, 1972.
11. R. Zamir, S. Shamai, and U. Erez, "Nested linear/lattice codes for structured multiterminal binning," *IEEE Trans. Inform. Theory* **48**, pp. 1250–1275, June 2002.
12. J. H. Conway and N. J. A. Slane, *Sphere Packings, Lattices and Groups*, Springer-Verlag, 3 ed., 1999.
13. 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 of Multimedia Content VII*, **5681**, (San Jose, CA), Jan. 2005.
14. M. Wu and B. Liu, "Watermarking for image authentication," in *Proc. IEEE Int. Conf. on Image Processing*, **II**, (Chicago,IL), 1998.
15. 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., ch. 3.1, pp. 71–79, Academic Press, 2000.
16. V. Licks and R. Jordan, "Geometric attacks on image watermarking systems," *IEEE Multimedia* **12**(3), pp. 68–78, 2005.
17. C.-Y. Lin, M. Wu, J. A. Bloom, M. L. Miller, I. J. Cox, and Y.-M. Lui, "Rotation, scale, and translation resilient public watermarking for images," *IEEE Trans. Image Processing* **10**, May 2001.