Analysis and Design of Authentication Watermarking

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]Taxas A&M University, 3128 TAMU, College Station, TX USA 77843-3128

ABSTRACT

This paper focuses on the use of nested lattice codes for effective analysis and design of semi-fragile watermarking schemes for content authentication applications. We provide a design framework for digital watermarking which is semi-fragile to any form of acceptable distortions, random or deterministic, such that both objectives of robustness and fragility can be effectively controlled and achieved. Robustness and fragility are characterized as two types of authentication errors. The encoder and decoder structures of semi-fragile schemes are derived and implemented using nested lattice codes to minimize these two types of errors. We then extend the framework to allow the legitimate and illegitimate distortions to be modelled as random noise. In addition, we investigate semi-fragile signature generation methods such that the signature is invariant to watermark embedding and legitimate distortion. A new approach, called MSB signature generation, is proposed which is shown to be more secure than the traditional dual subspace approach. Simulations of semi-fragile systems on real images are provided to demonstrate the effectiveness of nested lattice codes in achieving design objectives.

Keywords: Semi-fragile Watermarking, Authentication, Lattice Codes

1. INTRODUCTION

This paper addresses the problem of content authentication using coding-based digital watermarking in which a watermark is used to assist in verifying the integrity of its associated multimedia data. The embedding of an invisible watermark in 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, the authentication system should indicate that the signal is authentic. Conversely, when it undergoes illegitimate tampering, the distorted signal should be rejected as inauthentic. Applications of authentication watermarking include trusted cameras, digital insurance claim evidence, medical image archiving, journalistic photography, and digital rights management systems.

Initially proposed digital watermarking techniques for authentication were highly fragile [1,2] often detecting slight modifications to the signal in a similar way to traditional digital signatures. In order to exploit the benefits of a data embedding approach to authentication, semi-fragile watermarking methods were later introduced [3–7]. The primary advantage of employing semi-fragile watermarking over digital signature and fragile watermarking 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. One of the first approaches to semi-fragile watermarking called telltale tamper-proofing was proposed by Kundur and Hatzinakos [8] to determine the extent of modification both in the spatial and frequency domains of a signal using a statistics-based tamper assessment function. In *semi-fragile* watermarking, the watermark, often a host-dependent signature message, must survive legitimate distortions, but be destroyed by illegitimate modifications applied to the signal. Most proposed schemes to date are either designed for robustness to specific distortions (usually compression) using ad hoc development measures, or borrow from the robust watermarking literature and tune down the resilience of the watermark [5,7].

One influential semi-fragile system is the self-authentication-and-recovery image (SARI) method developed by Lin and Chang [3,4] in which a semi-fragile signature is designed to survive JPEG compression up to a certain level. To distinguish JPEG compression from malicious manipulation, two invariant properties of quantization are used to generate the signature and embed the watermark. The first property shows that a pre-quantized

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

Security, Steganography, and Watermarking of Multimedia Contents VI, edited by Edward J. Delp III, Ping W. Wong, Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 5306 © 2004 SPIE and IS&T · 0277-786X/04/$15

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 the invariant relationships between a pair of coefficients before and after JPEG compression, and is used to generate the watermark signature.

Although many techniques, such as SARI, work well under a class of attacks, their ad hoc design nature focusing on resilience to a specific distortion limits their portability to different applications. A more general formulation and design framework would be of interest for emerging multimedia applications.

We provide a design framework for semi-fragile signature watermarking where the legitimate distortions can be any form, random or deterministic, such that both objectives of robustness and fragility can be effectively controlled and achieved. We analyze semi-fragile watermarking using a coding approach, which is similar to more recent robust watermarking based on the concept of communications with side information [9, 10]. We demonstrate how this coding approach gives better characterization of the semi-fragile system and hence provides a superior way to control robustness and fragility for application-specific design.

The contributions of this paper are summarized below:

1. We provide a methodology to balance robustness and fragility objectives with respect to legitimate and illegitimate distortions that is superior to existing approaches.
2. We implement the semi-fragile requirements in a practical watermarking scheme using lattice codes. We demonstrate how lattice codes have regular structure with efficient encoding and decoding algorithms and are easy to analyze.
3. We investigate the security risk of traditional dual subspace approach of semi-fragile signature watermarking, and propose a novel approach, called MSB, which we show is more secure than the traditional dual subspace approach. Security analysis is provided based on tool-sets for content authentication.

We describe the general model of semi-fragile signature watermarking in Section 2. Section 3 derives the encoder and decoder structures to achieve robustness and fragility properties of the semi-fragile system. Section 4 focuses on signature generation to ensure security of the authentication system and proposes our novel MSB approach. In Section 5, we provide the simulation results of the novel MSB approach in the design of semi-fragile signature watermarking systems. Finally, conclusions are drawn in Section 6.

2. SEMI-FRAGILE SIGNATURE WATERMARKING

We consider the following authentication scenario as shown in Fig. 1 where a signature m is generated from a source signal, then embedded into the source as a watermark. In order for Bob to be assumed that the signal did originate from Alice, Alice derives a signature m from the source S with a secret key k using a signature generation function denoted by $m = H_k(S)$. She then embeds the signature m as a watermark into the host source S. The watermark embedding process is described by a function $X = f(S, m)$ where X is the watermarked signal. The watermarked signal X must satisfy imperceptability condition with the source S. Embedding distortion is often measured using L2-norm, $D = \frac{1}{n}\|S - X\|^2$ where n is the length of the signal.

The received signal Y could be tampered by some opponent Oscar in the public channel. As stated in the introduction, the semi-fragile system is designed to be robust to legitimate distortions such as incidental signal processing distortions, and fragile to illegitimate malicious tampering. These legitimate distortions include minor modifications such as high rate JPEG compression, and geometric distortions such as rotation, scaling and translation (RST). For geometric distortions, invariant watermark techniques [11,12] have been proposed recently in which watermarking takes place in some transform domain which is RST-invariant. The same techniques can be employed for semi-fragile problem to tolerate geometric distortions. The other type of legitimate distortions in authentication are minor modifications such as high rate JPEG compression. In [13], authentic and inauthentic regions of the original signal are specified as spheres in some suitable metric space. In this paper, we use a deterministic set of any shape, called admissible set Ω, to characterize the legitimate minor modifications which may occur on the authenticated signals. Denote the additive distortion $N = Y - X$. If the distortion $N \in \Omega$, it is legitimate and the distorted signal $Y = X + N$ is authentic; otherwise, the distortion is illegitimate and the

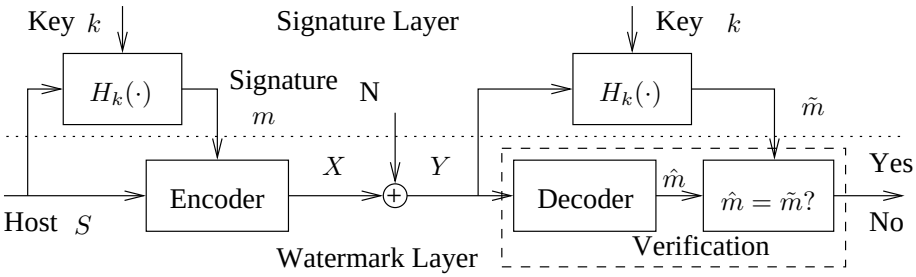


Figure 1. Semi-fragile signature watermarking for authentication.

distorted signal Y is inauthentic. Since Ω characterizes minor modifications, we assume the admissible set Ω is bounded and independent of the authenticated signal X.

At the receiver, to authenticate that the received signal is from Alice, Bob extracts the watermark from the received signal Y using the corresponding extraction function $\hat{m} = g(Y)$. If the extracted watermark is the same as the original signature, Bob accepts the signal as authentic, otherwise, rejects it.

Thus, a semi-fragile system consists two layers: the signature layer and the watermark layer. In the signature layer, a signature is generated to ensure the security of the authentication system. The signature generation must tolerate the legitimate noise. In other words, for the same key, $H_k(S) = H_k(Y)$ must hold for any legitimate noise. Since $Y = X + N = f(S, m) + N$, the signature must be generated from the features of the source that are invariant to embedding distortion and legitimate noise in the channel. In the watermark layer, the watermark is designed to be semi-fragile such that legitimate and illegitimate channel distortions can be distinguished. The received signal must be accepted as authentic if the channel distortion is legitimate and rejected otherwise.

We consider the following requirements necessary for the design of a semi-fragile signature watermarking authentication system.

1. Robustness and fragility objectives should be simultaneously addressed. When both cannot be completely achieved, one must have a quantitative mechanism to tradeoff between these objectives.
2. The authentication system must be secure to intentional tampering. For security, it must be computationally infeasible for the attacker to devise a fraudulent message without knowing the key.
3. The watermark must be imperceptible. Given a watermark payload, the watermarking scheme should be well-designed such that the embedding distortion is minimized.
4. The watermark embedding and authentication algorithms must be computationally efficient, especially for real time authentication applications.

3. WATERMARK LAYER DESIGN

Now we focus on the design of the watermark layer in Fig. 1 to analyze the requirements of the encoder and decoder structures to satisfy both robustness and fragility properties. It is assumed in this section that the same signature is derived using the same secret key at the receiver for authentication.

3.1. Robustness and Fragility Error Analysis

We define robustness and fragility as:

- Robustness: the ability to extract correctly the watermark m in the face of legitimate noise, $N \in \Omega$;
- Fragility: the inability to detect the presence of the watermark m in the face of illegitimate noise, $N \notin \Omega$.

In order to characterize robustness and fragility, we define two types of authentication errors. The Type I error, called false negative event, is the error that an authentic signal is wrongly regarded by the system as inauthentic. In other words, application of the legitimate noise results in failure to extract the correct watermark.

This type of authentication error characterizes the robustness of the semi-fragile watermarking system. The type II error, called false positive error, occurs when an inauthentic signal is wrongly authenticated. This corresponds to situations in which an authentic signal has been illegitimately tampered yet the watermark is extracted properly. This type of authentication error, considered more serious in authentication applications, characterizes the fragility of the semi-fragile system. Our overall objective then is to minimize the false negative and positive authentication errors, and optimize authentication performance such as embedding distortion, embedding and verification complexity.

For convenience, we assume the signals S, X, Y, and N all take values in n-dimensional Euclidean space, although our analysis can also be applied to other linear spaces such as the n-dimensional binary space. Given a derived signature message m, $m \in \{1, 2, \cdots, M\}$ where M is the total number of signature messages, we define the *encoding region* $\mathcal{C}(m)$ to be the set of signals containing the signature/watermark message m. That is, $\mathcal{C}(m) = \{f(S, m) \in \mathbb{R}^n | \forall S \in \mathbb{R}^n\}$, which is the set the transmitter Alice uses to send authentic signals. The set $\mathcal{C}(m)$ can also be regarded as the reconstruction point set of a quantizer corresponding to the message m; for example, the set of points marked with +'s in Fig. 2(a) corresponds to the encoding region $\mathcal{C}(1)$ of message 1, and the set of those with $\times$'s corresponds to $\mathcal{C}(2)$ of message 2. To minimize the induced embedding distortion, the nearest neighbor in $\mathcal{C}(m)$ around the source S is selected as the watermarked signal X. Thus the embedding of signature message m is to find $X = \arg\min_{x \in \mathcal{C}(m)} \|S - x\|$.

Let the *decoding region* $\mathcal{E}(m)$ be the set of signals from which the signature message m can be extracted, $\mathcal{E}(m) = \{Y \in \mathbb{R}^n | g(Y) = m\}$. This is the set the receiver Bob uses to accept authentic signals. Thus a received signal Y is viewed as authentic only if $Y \in \mathcal{E}(m)$.

In this work, we do not consider joint authentication in which two or more watermarks can be extracted from a signal. Therefore, from a received signal, only one signature message can be extracted. The decoding regions associated with different individuals must satisfy the disjoint condition:

$$\mathcal{E}(i) \bigcap \mathcal{E}(j) = \varnothing \quad \forall i \neq j. \tag{1}$$

Since the semi-fragile watermark should survive legitimate noise within the admissible set Ω, for any watermarked signal $X \in \mathcal{C}(m)$ and any legitimate distortion $N \in \Omega$, the received signal $Y = X + N$ should be regarded as authentic. Thus for robustness, the signature m should be extracted from the signals in the region $\mathcal{C}(m) + \Omega$ where the set summation is defined as $A + B = \{a + b | a \in A \text{ and } b \in B\}$.

On the other hand, to minimize the false positive authentication error, the hidden watermark m must become undetectable from the signals if the additive distortion N is illegitimate. For the region outside $\mathcal{C}(m) + \Omega$, m cannot be extracted. This requires

$$\mathcal{C}(m) + \Omega = \mathcal{E}(m) \quad \forall m. \tag{2}$$

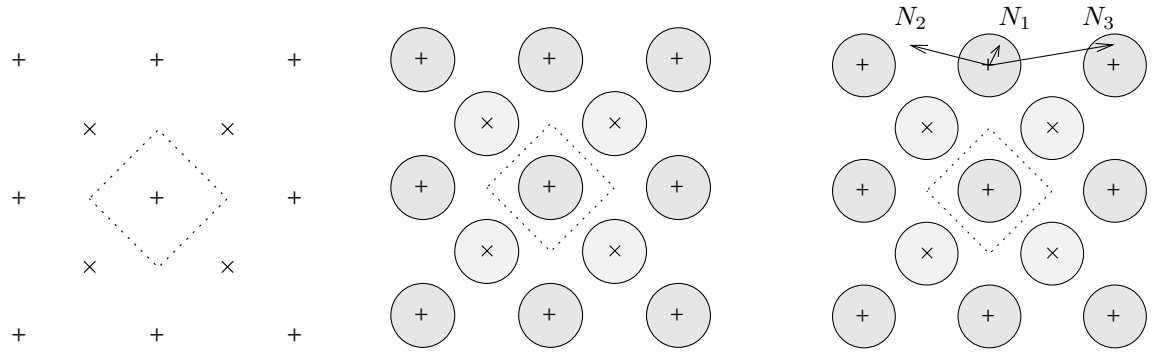
Figs. 2(a) and 2(b) illustrate one simple example of the encoding and decoding structures specified by Eqn. (2). From the disjoint constraint of Eqn. (1), the codes $\mathcal{C}(m), m = 1, 2, \cdots, M$ for different signature messages must be far enough such that they are still distinguishable even when legitimate noises occur. This guarantees that the hidden signature message can be extracted correctly if the noise is legitimate, making false negative authentication error impossible, and resulting in ideal robustness for the semi-fragile system.

Fragility, however, is not completely addressed because it is possible that for some $N \notin \Omega$, it just so happens that from the received signal $Y = X + N$, the hidden signature m is still extracted. In our framework, this corresponds to the following event

$$Y = X + N \in \mathcal{C}(m) + \Omega \ \& \ N \notin \Omega \text{ given } X \in \mathcal{C}(m) \tag{3}$$

which is illustrated as N_3 in Fig. 2(c). This error event leads to security concerns for malicious tampering as we discuss in Section 4. This false positive error event cannot be eliminated due to the *blind* nature of the watermarking system in which the original source is not available at the receiver. Essentially, it is unavoidable to have some error probability that N will push Y such that it yields an authentic results, but is not in the set Ω. However, we can reduce this error probability by increasing the total number of signature messages.

With the decoding region $\mathcal{E}(m)$ in Eqn. (2), the authentication process is conducted by verifying if the derived signature m can be extracted from the received signal Y. That is, to verify if $Y \in \mathcal{E}(m) = \mathcal{C}(m) + \Omega$.



(a) Encoding Regions. (b) Decoding Regions. (c) Legitimate and illegitimate distortions.

Figure 2. Nested lattice codes for semi-fragile watermarking. (a) All points marked with +'s and ×'s form the fine lattice Λ_1 and the points marked only with +'s form the sublattice Λ_2. Λ_1 is partitioned into cosets of Λ_2: one is Λ_2 itself, the other is a shifted version of Λ_2, points marked with ×'s. The fundamental Voronoi region of Λ_1, $\mathcal{V}_0(\Lambda_1)$, is shown by the dotted shape. (b) The fundamental Voronoi region $\mathcal{V}_0(\Lambda_1)$ covers the admissible set Ω, which is a disk. Each coset corresponds to a watermark message. The points marked with +'s and ×'s are the encoding region $\mathcal{C}(m)$ for $m = 1$ and $m = 2$, respectively. The shadowed region around points marked with +'s and ×'s are the decoding region $\mathcal{E}(m)$ for $m = 1$ and $m = 2$, respectively. (c) Three types of distortions: i. $N_1 \in \Omega$; ii. $N_2 \notin \Omega$ and found to be inauthentic; iii. $N_3 \notin \Omega$ and found authentic, leading to false positive error.

3.2. Nested Lattice Code Implementation

So far, we have derived the encoding and decoding structures of the semi-fragile authentication system. We note a similarity to recent robust watermarking schemes such as QIM [9], or SCS [10] and more generally, algorithms based on the concept of communications with side information [14]. From the coding approach for semi-fragile authentication, we see that the code-based structure is the same in principle. A code consists of several subsets corresponding to different messages (or keys in authentication) and these subsets must be far enough such that they can be reconstructed in the presence of legitimate noises. In other words, the decoding region must cover the area of legitimate noises around the watermarked signals. The idea of channel capacity of robust watermarking can also be used to estimate the maximum number of signature messages for semi-fragile authentication. However, the primary distinction between watermarking and authentication is that robust watermarking is essentially a communication problem while authentication is a verification problem.

Because the code structures for robust watermarking and semi-fragile watermarking are the same, we can borrow some of the implementation insights from the robust watermarking literature as well as more general communications with side information schemes [9, 14]. Lattice codes have shown potential in the robust watermarking research field because of their regular structure that makes faster decoding possible while achieving capacity. Though authentication is basically a verification problem, lattice codes have advantages in distinguishing effectively legitimate or illegitimate distortions which may occur on the watermark signal. For this reason, we, too, make use of lattice codes.

We first provide some basic definitions and notions of lattice and nested lattice codes to help understand our semi-fragile watermarking system implementation. For detailed definitions and properties, the reader is referred to [14–16]. Roughly speaking, a lattice Λ is simply a regular array of points in n-dimensional Euclidean space $\mathbb{R}^n$. Mathematically, an n-dimensional lattice Λ is obtained as the set of all integer multiples of a group of n basic vectors, $g_1, g_2, \cdots, g_n \in \mathbb{R}^n$, i.e., $\Lambda = \left\{ \sum_{i=1}^{n} a_i g_i | a_i \in \mathbb{Z} \right\}$.

A nearest neighbor quantizer $Q(\cdot)$ associated with lattice Λ is defined by $Q(x) = \arg\min_{\lambda \in \Lambda} \|x - \lambda\|$ where ties on equal distance are broken in a systematic fashion [17]. The modulo-Λ operation is defined as $x \bmod \Lambda = x - Q(x)$ which is the quantization error of x with respect to Λ. The fundamental Voronoi region of Λ is the set of points that are closest to the origin and all quantize to the same value, i.e. $\mathcal{V}_0(\Lambda) = \{x \in \mathbb{R}^n | Q(x) = 0\}$.

A pair of lattices (Λ_1, Λ_2) is nested if Λ_2 is a sublattice of Λ_1, i.e. $\Lambda_2 \subset \Lambda_1$. We say that Λ_1 and Λ_2 are the fine and coarse lattices, respectively. The points of the set $[\Lambda_1/\Lambda_2] = \Lambda_1 \cap \mathcal{V}_0(\Lambda_2)$ are called the coset leaders of Λ_2 relative to Λ_1. For each coset leader $v \in [\Lambda_1/\Lambda_2]$, the shift lattice $\Lambda_2 + v$ is called a coset of Λ_2 relative to Λ_1. By elementary group theory, the fine lattice Λ_1 is partitioned into cosets of the coarse lattice Λ_2, each of which is a version of Λ_2 shifted by a coset leader in $[\Lambda_1/\Lambda_2]$. This is called a coset decomposition of Λ_1 and is represented as $\Lambda_1 = \Lambda_2 + [\Lambda_1/\Lambda_2]$.

For semi-fragile watermarking authentication, we design a nested lattice code (Λ_1, Λ_2) such that the admissible set of distortions Ω is contained in the fundamental Voronoi region of the fine lattice Λ_1, $\mathcal{V}_0(\Lambda_1)$. Fig. 2 illustrates the use of nested lattice codes for semi-fragile watermarking in 2-dimensional Euclidean space where the fine and coarse lattices are $\Lambda_1 = \{(i+j, i-j) \in \mathbb{R}^2 | i, j \in \mathbb{Z}\}$ and $\Lambda_2 = \{(2i, 2j) \in \mathbb{R}^2 | i, j \in \mathbb{Z}\}$, respectively. The fundamental Voronoi region of Λ_1, $\mathcal{V}_0(\Lambda_1) = \{(x, y) \in \mathbb{R}^2 | -1 < x + y \leq 1 \ \& \ -1 < x - y \leq 1\}$, covers the admissible set Ω, which is a disk of radius $\frac{1}{2}$. The sublattice Λ_2 induces a partition of Λ_1 to two cosets: one is Λ_2 itself, the other is $\Lambda_2 + (1, 1)$, a shifted version of Λ_2 by a coset leader $(1, 1)$. Each coset of Λ_2 corresponds to a watermark message. For watermark message m, the coset $\Lambda_2 + v(m)$ is used as the encoding set $\mathcal{C}(m)$ where $v(m) \in [\Lambda_1/\Lambda_2]$ is the coset leader corresponding to m. In the example, $v(1) = (0, 0)$ and $v(2) = (1, 1)$.

The encoder embeds a watermark message m by quantizing the n-dimensional source S using the quantizer associated with the coset $\Lambda_2 + v(m)$. The embedding function is $X = Q_2(S - v(m)) + v(m)$ where S is the source, X is the watermarked signal, $Q_2(\cdot)$ is the quantizer associated with the coarse lattice Λ_2, and $v(m) \in [\Lambda_1/\Lambda_2]$ is the coset leader corresponding to message m. The average embedding distortion is $D = \frac{1}{n} E\{(X - S)^2\} = \frac{1}{n} E\{(Q_2(S - v(m)) - (S - v(m)))^2\}$, which is the energy of the quantization noise per dimension. Thus to reduce the embedding distortion, one would like to choose the fine lattice such that its fundamental Voronoi region is as small as possible to cover the admissible set Ω.

Since the fundamental Voronoi region $\mathcal{V}_0(\Lambda_1)$ covers the admissible set Ω, authentication is conducted by using the quantizer associated with the coset $\Lambda_2 + v(m)$, as follows:

1. Given the received signal Y and the derived signature m, first find the nearest point in the coset $\Lambda_2 + v(m)$ around Y. This is equivalent to computing $\hat{X} = Q_2(Y - v(m)) + v(m)$.
2. Estimate the distortion $\hat{N} = Y - \hat{X}$. If $\hat{N} \in \Omega$, the received signal is authentic. Otherwise, it is inauthentic.

When Y is considered inauthentic, with the knowledge of the coset $\Lambda_2 + v(m)$, there is potential to characterize the illegal tampering. The estimated illegitimate distortion $\hat{N} = Y - \hat{X}$ can help determine the extent of illegal tampering using the telltale tamper-proofing technique in [8]. In particular, if the amplitude of malicious tampering is less than one half of the minimum distance of the coset $\Lambda_2 + v(m)$, the illegal distortion can be correctly characterized. We see that the coarse lattice Λ_2 determines the ability to recover illegitimate tampering.

3.3. Random Legitimate Noises

In the last section, we analyzed the encoder and decoder structures by using a deterministic admissible set Ω. However, in most cases, legitimate and illegitimate noises are specified by random variables. For instance, one may like to design a semi-fragile system such that the watermark can survive additive white Gaussian noise (AWGN) with energy below a certain threshold and be totally destroyed by AWGN with energy above this value. In this example, the admissible set of legitimate noises is a set of Gaussian random variables with variance less than a certain value and the inadmissible set is a set of Gaussian random variables with variance greater than that value.

We consider distortion noises modelled as parameterized random variables. In general, let $N(p)$ be a random variable with a parameter p such that $N(p)$ is legitimate if $p < a$ and illegitimate if $p > a$ where a is a given value. In such a situation, our approach is to find a deterministic set to distinguish the set of legitimate random noises from the set of illegitimate ones, then use the deterministic set to design the encoder and decoder structures as discussed in the previous section.

Given an i.i.d sequence $N^n = [N_1, N_2, \ldots, N_n]$ of length n drawn from a random variable N, we find a deterministic set $\Omega_n \subset \mathbb{R}^n$ such that if the random variable N is legitimate, the sequence is in the set Ω_n with high probability; otherwise if N is illegitimate, the sequence is not in the set Ω_n with high probability.

The approximation of a random set by a deterministic set inevitably introduces two types of errors. Type I error is one in which the sequence N^n drawn from a legitimate noise is actually outside the deterministic set Ω_n. Type II error involves the case in which the sequence N^n of an illegitimate noise is within the admissible set Ω_n. Type I set approximation error results in the failure to extract the correct watermark, thus contributes to Type I authentication error. Similarly, Type II set approximation error results in Type II authentication error. Therefore, a well-designed Ω_n should optimize and balance both types of approximation error. The shape of the set Ω_n should also be considered for easy implementation using using nested lattice codes.

We assert that the following two properties are necessary for the deterministic set Ω_n,

- Asymptotic zero probability of Type I approximation error: For any legitimate random noise, the probability that $N^n \notin \Omega_n$ approaches zero when the dimension n goes to infinity.
- Asymptotic zero probability of Type II approximation error: For any illegitimate random noise, the probability that $N^n \in \Omega_n$ approaches zero when the dimension n goes to infinity.

Only when the above properties are satisfied, is it possible to design a semi-fragile watermarking system that is guaranteed to be robust to legitimate noises in the admissible set, but fragile to illegitimate noises in the inadmissible set. Another implicit requirement for semi-fragile watermarking implementation is that the deterministic set Ω_n must be bounded or at least bounded after projecting to some subspace. This requirement is always satisfied since the legitimate noise always implies the minor modification as discussed in Section 2.

4. SIGNATURE LAYER DESIGN

The semi-fragile watermarking in the watermark layer is not secure because as stated in Eqn. (3), it is very easy for an opponent Oscar to devise an illegitimate distortion N which will push Y such that it still yields an authentic results. The signature layer is incorporated to enforce security for the authentication system. Traditionally, the notion of "security" for digital watermarking refers to its ability to resist intentional tampering [18]. To apply digital watermarking for authentication, cryptographic security measures also need to be considered.

4.1. Security Requirement

To analyze the security requirement of the semi-fragile systems, we need to identify possible attacks which the opponent Oscar may use to defeat the semi-fragile authentication system. The following two types of attacks are therefore critical for semi-fragile systems. To analyze these possible attacks, it is always assumed that the attacker has the knowledge of the semi-fragile watermarking scheme except the secret key.

1. The first type of attack occurs when an opponent tries to devise a fraudulent message to the receiver based on his/her knowledge of the general authentication scheme. For this type of attack, commonly referred to as impersonation, the opponent is successful if the devised signal is wrongly regarded by the receiver as authentic and from the transmitter .
2. The second type of attack occurs when an opponent intercepts a legitimate message from the transmitter, then alters the message in illegitimate manner such that the tampered signal is accepted by the receiver as authentic.

From the system design process of semi-fragile signature watermarking systems, we know the received signal Y is regarded as authentic if and only if the derived signature $\tilde{m} = H_k(Y)$ can be extracted from Y. From the code structure of the watermark embedding and extraction in Section 3, there is only one watermark which can be extracted from Y. Denote this extracted watermark $\hat{m} = g(Y)$. Thus the first type of attack is successful if a signal devised by the attacker happens to extract the watermark $\hat{m} = g(Y)$ equal to the signature $\tilde{m} = H_k(Y)$. Since the attacker does not know the secret key of the transmitter, his probability of choosing an authentic signal is less than $\frac{1}{M}$ where M is the total number of signature messages. Therefore, in the design of a semi-fragile system, the signature space must be large enough or the length of the signature bit sequence must be long enough to reduce the probability of a successful impersonation attack.

For the second type of attack, suppose the attacker intercepts a copy of authentic signal X from the transmitter, then he can extract the original signature message m from X because he has the knowledge of the

watermarking scheme. He tries to find an illegitimate copy X' such that the same watermark and signature are preserved. This illegitimate copy X' must satisfy $\hat{m} = m$ and $\tilde{m} = m$ where m is the original signature message from X, $\tilde{m}$ is the signature generated from X', and $\hat{m}$ is the extracted watermark from X'. A semi-fragile system is secure against this type of attack if given one authentic signal X, it is computational infeasible for the attacker to find such illegitimate signal X'.

4.2. Dual Subspace Approach

In some semi-fragile systems, the legitimate noise only takes place in part of the source, leaving the remaining unchanged. For example, legitimate high quality compression only affects the perceptually insignificant components of an image. Thus one can generate the signature from the perceptually significant components of the signal, and then embed the signature in the perceptually insignificant components of the signal. This approach is very common in semi-fragile signature system design such as [3, 18, 19]. We call this approach dual subspace approach. In the dual subspace approach, the whole signal space is divided into two subspaces and watermarking and signature generation occur in different subspaces. The main advantage of this approach is that both watermarking and signature generation processes can be designed and implemented separately. However, this approach leads to security concern since the signature technique only protects the the signature subspace, leaving the watermark subspace vulnerable for the above two types of attacks.

Since part of the source signal is used for signature generation, the total number of signature messages which can be embedded in the remaining region is reduced, thus increasing the probability of a successful first type attack.

For the second type of attack, the signature subspace is protected by the signature technique with a key, so it is computationally infeasible to devise another illegitimate signal X' from which the same signature is generated. However, the watermark subspace is not protected. Since the attacker knows the watermarking scheme, from a watermarked signal X, he can extract the hidden signature m, thus can easily construct an illegitimate signal X' such that the watermark m is preserved. Thus in the dual subspace approach, a key in signature layer is not enough, there must be some secret key in the watermarking scheme to protect the watermark subspace such that given a watermarked signal X, it is computationally infeasible to find another illegitimate signal X' which will produce the same watermark.

Thus from the above analysis, though signature generation and watermark extraction processes are independently designed in the dual subspace approach, both processes must be secure to illegal tampering. A secret key in the signature layer is not enough to guarantee the security of the overall semi-fragile system,

4.3. MSB Approach

Based on the security risk of the dual subspace approach, we propose a novel approach that uses a nested lattice code for both the watermark and signature layers, and the secret key in the signature layer is used to protect both layers. We call this approach the most significant bit (MSB) signature generation method, which originates from the least significant bit (LSB) strategy of watermark embedding. In LSB watermarking, the LSB is replaced by the watermark bit. Thus the remaining most significant bits (MSBs) are the features that are invariant to watermark embedding and legitimate channel noise. Walton [20] proposed to hide the checksums of the seven MSBs in the LSBs of pixels. We generalize this idea into n-dimensional Euclidean space using a nested lattice code. For reasons of security, a keyed hash message authentication code is used instead to generate signatures.

Instead of subspace decomposition of the source, the MSB approach uses a decomposition property of nested lattices for signature generation and watermark embedding. Given an n-dimensional nested lattices (Λ_1, Λ_2), the Euclidean space can be decomposed as follows [16], $\mathbb{R}^n = \Lambda_1 + \mathcal{V}_0(\Lambda_1) = \Lambda_2 + [\Lambda_1/\Lambda_2] + \mathcal{V}_0(\Lambda_1)$ where $[\Lambda_1/\Lambda_2]$ is the set of coset leaders of Λ_1 relative to Λ_2 and $\mathcal{V}_0(\Lambda_1)$ is the fundamental Voronoi region of the lattice Λ_1. From the above decomposition property, for any point $S \in \mathbb{R}^n$, there exist unique $\lambda_1 \in \Lambda_1$, $\lambda_2 \in \Lambda_2$, $v \in [\Lambda_1/\Lambda_2]$, and $r \in \mathcal{V}_0(\Lambda_1)$ such that $S = \lambda_1 + r = \lambda_2 + v + r$.

We can, therefore, design the nested lattice code (Λ_1, Λ_2) such that the fine lattice Λ_1 resists channel noise, the coset of the coarse lattice Λ_2 relative to Λ_1 represents the watermark message, and the quantized value of the source by the coarse lattice Λ_2 is the invariant feature from which to generate the signature. In other words, the

fundamental Voronoi region $\mathcal{V}_0(\Lambda_1)$ must cover the legitimate noise region Ω_n such that the quantized value of the source by Λ_1 can be recovered from legitimate noise. Given a source S, let $I = Q_2(S)$ be the quantized value of S where $Q_2(\cdot)$ is the quantization operation using the coarse lattice Λ_2. $I = Q_2(S)$ corresponds to the most significant bits in binary case of LSB watermarking. Then the signature m is generated from the invariant feature I using a keyed hash function or message authentication code (MAC) for security and then embedded into the source S by shifting the quantized value $I = Q_2(S)$ with an offset $v(m)$ where $v(m)$ is a one-to-one mapping from watermark message set to coset representative set $[\Lambda_1/\Lambda_2]$. We therefore have the following watermark embedding function $X = Q_2(S) + v(m)$ such that the watermarked signal $X \in \Lambda_1$. The average embedding distortion is $D = \frac{1}{n}E\{(X-S)^2\} = \frac{1}{n}E\{(Q_2(S) - S + v(m))^2\}$. In case of high rate quantization, the quantization noise is uniformly distributed and symmetric, independent of $v(m)$, then $D = \frac{1}{n}E\{(Q_2(S) - S)^2\} + \frac{1}{n}E\{v(m)^2\}$. The first term is the energy of quantization noise per dimension, and the second term is the energy of the dither values per dimension.

The received signal $Y = X+N = Q_2(S)+v(m)+N$. If N is legitimate, $N \in \mathcal{V}_0(\Lambda_1)$. Because of the uniqueness of the lattice decomposition, the watermarked signal X, the invariant feature I, the coset representative $v(m)$, and even the legitimate noise N all can be reconstructed from the received signal Y using the corresponding lattice quantizers Λ_1 and Λ_2 as follows,

$$\hat{X} = Q_1(Y), \tag{4}$$

$$\hat{I} = Q_2(\hat{X}), \tag{5}$$

$$\hat{v}(m) = \hat{X} \bmod \Lambda_2 = \hat{X} - Q_2(\hat{X}) \tag{6}$$

where $Q_1(\cdot)$ and $Q_2(\cdot)$ are the quantization operations with respect to the fine lattice Λ_1 and the coarse lattice Λ_2, respectively.

The signature is then generated from the reconstructed invariant feature $\hat{I} = Q_2(Q_1(Y))$ by applying the same keyed hash or MAC. The embedded watermark is extracted from the reconstructed $\hat{v}(m)$. Authentication is conducted by verifying if the derived signature is equal to the extracted watermark and the estimated noise $\hat{N} = Y - \hat{X}$ is legitimate, i.e. $\hat{N} \in \Omega_n$.

4.4. Security Analysis of MSB Approach

Now we investigate the security of the MSB approach using nested lattice codes. The nested lattice code (Λ_1, Λ_2) is designed such that the fine lattice Λ_1 resists channel noise, the coset of the coarse lattice Λ_2 relative to Λ_1 represents the watermark message, and the quantized value of the source by the coarse quantizer Λ_2 is the invariant feature to generate the signature.

Suppose the watermark signal X sent from the transmitter has the following decomposition, $X = \lambda_2 + v(m)$ where $\lambda_2 \in \Lambda_2$ is the quantized value of the source S by Λ_2, $v(m) \in [\Lambda_1/\Lambda_2]$. The signature m is generated from λ_2 using a keyed hash function, and then embedded to $v(m)$ using one-to-one mapping function. In order to find an authentic signal $X' \in \mathbb{R}^n$ from which the same signature and watermark are derived as from X, we represent X' as the following unique decomposition, $X' = \lambda_2' + v' + r'$ where $\lambda_2' \in \Lambda_2$, $v' \in [\Lambda_1/\Lambda_2]$, and $r' \in \mathcal{V}_0(\Lambda_1)$.

The signature is generated from the first term of the decomposition, i.e. a point in the coarse lattice Λ_2. Because the signature generation function is a hash function, it is computationally infeasible for the attacker to find a $\lambda_2' \in \Lambda_2$ from which the same signature is generated as from λ_2 of the watermarked signal X. Therefore λ_2' must be equal to λ_2. Because the mapping from coset leaders to watermark messages is an one-to-one function, and the same watermark is extracted from v' and $v(m)$, so $v' = v(m)$. Since X' must be authentic, r' must be admissible. Therefore, the difference between X and X' is only r', which is admissible. Thus given one watermarked signal X, it is computational infeasible for the attacker to find an illegitimate authentic signal X' such that the same signature and watermark are derived as from X. The MSB approach of watermarking and signature generation is secure against the second type of attack.

From the above analysis, we see that MSB approach places watermark and signature messages in different regions of lattice decomposition, making it secure against the second type of attack even using just one secret key in the signature layer. The security of the MSB approach is determined by the security of the signature generation function.

5. SIMULATION RESULTS OF MSB APPROACH

In this section, two simple examples are illustrated to show how to choose deterministic set to distinguish random legitimate and illegitimate distortions, then use nested lattice codes to implement MSB approach.

5.1. Uniformly Distributed Noise

Suppose N is uniformly distributed in the range of $[-\frac{\Delta_N}{2}, \frac{\Delta_N}{2}]$. The system is designed such that it is robust to uniformly distributed noise with parameter $\Delta_N < 1$, but fragile to that with $\Delta_N > 1$.

The sequence $N^n = [N_1, N_2, \cdots, N_n]$ is uniformly distributed in an n-dimensional cube. Therefore it is natural to choose the deterministic set Ω_n to be an n-dimensional cube, that is,

$$\Omega_n = \{[x_1, x_2, \cdots, x_n] \in \mathbb{R}^n | x_i \in [-\frac{1}{2}, \frac{1}{2}], \forall i = 1, 2, \cdots, n\}. \tag{7}$$

When the parameter $\Delta_N < 1$, the sequence N^n is totally contained in Ω_n, thus Type I error probability $P_I(\Delta_N)$ is zero. When $\Delta_N > 1$, Type II error probability $P_{II}(\Delta_N) = (\frac{1}{\Delta_N})^n$, the probability that N^n is contained in Ω_n. Since $\frac{1}{\Delta_N} < 1$, Type II error probability thus approaches zero as $n \to \infty$.

We choose the fine lattice to be $\Lambda_1 = \mathbb{Z}^n$ whose fundamental Voronoi region $\{x^n \in \mathbb{R}^n | -0.5 < x_i \leq 0.5 \ \forall i\}$ exactly covers the admissible set. The choice of the coarse lattice should trade off between signature payload capacity and the allowed embedding distortion. For simplicity, we choose $\Lambda_2 = 2\mathbb{Z}^n$ such that both lattices are uniform scalar quantizers. This nested lattice code (Λ_1, Λ_2) has message rate $R = 1$, in other words, one bit can be embedded in one dimension.

Suppose n is the length of the source, given a source $S = [s_1, s_2, \cdots, s_n]$, the quantized value $I = Q_2(S)$ is the invariant feature for signature generation. The signature $m = [m_1, m_2, \cdots, m_n]$ is then generated from I using a keyed Hash message Authentication Code (HMAC) [21]. Since $v(m) \in [\Lambda_1/\Lambda_2]$ is a binary sequence, we choose the message mapping to be $v(m) = m$. Let $X = [x_1, x_2, \cdots, x_n]$ be the watermarked signal, then the watermark embedding function is the following: $x_i = Q_2(s_i) + m_i, \ i = 1, 2, \cdots, n$ where $Q_2(\cdot)$ is the standard uniform scalar quantizer with quantization step 2. The expected distortion induced by the embedding is $D = \frac{5}{6}$. The distortion can be further reduced to $D = \frac{7}{12}$ using the coset leader mapping $v(0) = -0.5, v(1) = 0.5$ such that the average is centered at zero.

Suppose the received signal is $Y = [y_1, y_2, \cdots, y_n]$. Then the watermark signal X, the invariant feature I, and the coset representative $v(m)$ all can be reconstructed from the received signal Y using the corresponding lattice quantizers Λ_1 and Λ_2 if the distortion N is legitimate. The whole authentication process is as follows: First, from the received signal Y, reconstruct the watermarked signal, the invariant feature $I = [I_1, I_2, \cdots, I_n]$ and extract the watermark by $\hat{x}_i = Q_1(y_i), \quad \hat{I}_i = Q_2(\hat{x}_i), \quad \hat{m}_i = \hat{x}_i - Q_2(\hat{x}_i)$ where $Q_1(\cdot)$ and $Q_2(\cdot)$ are the standard uniform scalar quantizers with quantization step 1 and 2, respectively. Then generate the signature $\tilde{m}$ from the reconstructed invariant feature $\hat{I}$ using the same HMAC and the same key. If the derived signature is equal to the extracted watermark, $\tilde{m} = \hat{m}$, the received signal is authentic, otherwise, it is inauthentic.

We simulate the semi-fragile signature scheme on 8×8 blocks of image Lenna. The length n is set to be 64. Fig. 3 shows the error probability curves. The probability of two types of authentication error is plotted as the parameter Δ_N of the uniformly distributed noise varies. The error curve when $\Delta_N < 1$ describes the robustness property and the curve when $\Delta_N > 1$ describes the fragility property. Both analytical results derived from the deterministic set and simulation results are computed and shown as solid lines in Fig. 3. We can see that the unit cubic set has no Type I error probability, so the robustness objective is fully achieved. However, for all $\Delta_N > 1$, the unit cubic set has Type II error probability. Improvement can be made by choosing a deterministic set with balanced error probabilities associated with robustness and fragility. For example, we choose the other deterministic set $\Omega'_n = \{[x_1, x_2, \cdots, x_n] \in \mathbb{R}^n | |x_i| < 0.4945, \forall i = 1, 2, \cdots, n\}$, which is a little smaller than the unit cubic set. Its analytical and simulation error curves are also plotted as dotted lines in Fig. 3, which have balanced error probabilities.

Please note that around the point $\Delta_N = 1$, it is hard to distinguish between the legitimate and illegitimate random variables, so the error probability is quite large. Actually at $\Delta_N = 1$, $P_I(\Delta_N) + P_{II}(\Delta_N) = 1$, thus we cannot reduce the error probabilities at $\Delta_N = 1$. Our choice of deterministic set is to make both error curves as sharp as possible around $\Delta_N = 1$. Even sharper error curves are possible by increasing the dimension n.

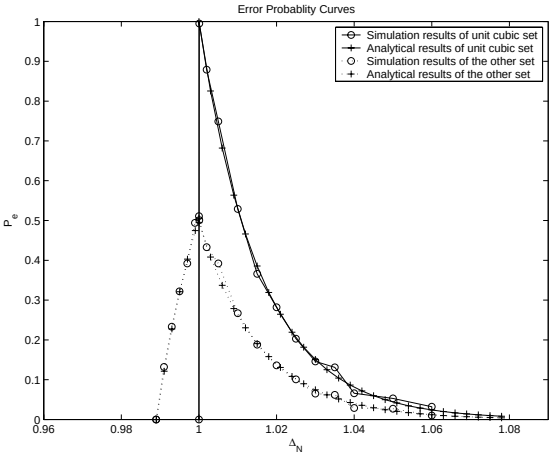


Figure 3. Uniformly distributed noise.

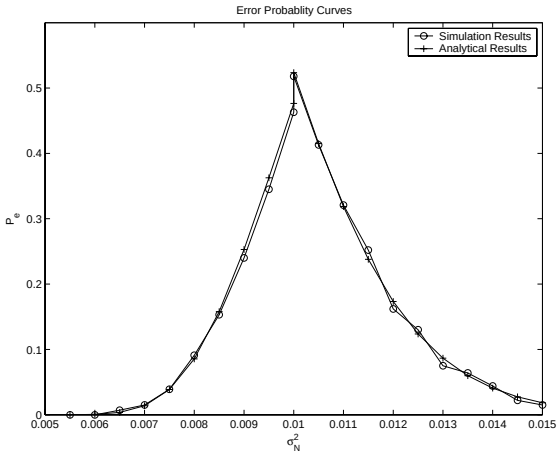


Figure 4. Error probability curves: AWGN.

5.2. AWGN

In the second example, the system is designed such that it is robust to additive white Gaussian noise with energy $\sigma_N^2 < 0.01$, but fragile to that with $\sigma_N^2 > 0.01$.

From information theory, we know that for large n, the n-dimensional Gaussian noise sequence is within an n-dimensional sphere of radius $\sqrt{0.01n}$ if $\sigma_N^2 < 0.01$, and outside the sphere if $\sigma_N^2 > 0.01$ with high probability. Thus a good choice of the deterministic set is an n-dimensional sphere of radius $\sqrt{0.01n}$. That is,

$$\Omega_n = \{[x_1, x_2, \cdots, x_n] \in \mathbb{R}^n | x_1^2 + x_2^2 + \cdots + x_n^2 \leq 0.01n\}. \tag{8}$$

When $\sigma_N^2 < 0.01$, Type I error probability $P_I(\sigma_N^2) = P[N^n \notin \Omega_n] = P[\chi^2(n) > \frac{0.01n}{\sigma_N^2}]$ where $\chi^2(n)$ is the chi-squared distribution with n degrees of freedom. When $\sigma_N^2 > 0.01$, Type II error probability $P_{II}(\sigma_N^2) = P[N^n \in \Omega_n] = P[\chi^2(n) < \frac{0.01n}{\sigma_N^2}]$. Both error probabilities approach zero when n goes to infinity.

We again use a scaled version of the simple nested lattice code, $\Delta(\mathbb{Z}/2\mathbb{Z})$ in which the fine lattice is a scalar uniform quantizer with step size Δ and the coarse lattice is a scalar uniform quantizer with step size 2Δ. The fundamental Voronoi region $\mathcal{V}_1$ of the fine quantizer Λ_1 is $\{x^n \in \mathbb{R}^n | |x_i| < \frac{\Delta}{2} \ \forall i\}$. Set $\Delta = 2\sqrt{0.01n}$ such that the fundamental Voronoi region $\mathcal{V}_1$ covers the admissible set Ω_n. The watermark embedding and extraction, the signature generation processes are the same as the previous example. Because the fundamental Voronoi region is bigger than the admissible set Ω_n, the received signal is authentic only if the derived signature is equal to the extracted watermark, $\tilde{m} = \hat{m}$, and the estimated noise $\hat{N} = Y - \hat{X}$ is admissible.

Let the length of the source $n = 64$, Fig. 4 shows both analytical and simulation results of the error probability curve associated with the admissible set. The expected embedding distortion $D = 1.4933$ using the coset leader mapping $v(0) = -0.5, v(1) = 0.5$. To further decrease the embedding distortion, trellis codes [16,22] can be used because trellis codes can achieve more spherical fundamental Voronoi region. In our simulation, an Ungerboeck's 4-state trellis code using conventional code $G(D) = [D \quad 1 + D^2]$ and lattice partition chain $\mathbb{Z}/2\mathbb{Z}/4\mathbb{Z}$ can reduce the embedding distortion to $D = 0.7461$. Further reduction of embedding distortion is possible with compromise of the error probabilities associated with robustness and fragility.

6. CONCLUSIONS

The results of this paper show that our coding approach provides useful and constructive tools in the analysis and design of authentication watermarking. In particular, we demonstrate the effectiveness of nested lattice codes in achieving both robustness and fragility design objectives.

We also analyze the security requirements of semi-fragile watermarking systems. We show that the new MSB signature generation approach is more secure than the traditional dual subspace approach. Although a symmetric keyed HMAC is used for signature generation in MSB approach, it is possible to incorporate asymmetric keyed digital signature schemes in semi-fragile watermarking systems. Future work considers asymmetric keyed schemes.

REFERENCES

1. M. M. Yeung and F. Mintzer, "An invisible watermarking technique for image verification," in *Proc. IEEE Int. Conf. on Image Processing*, (Santa Barbara, CA), Oct. 1997.
2. P. W. Wong, "A public key watermark for image verification and authentication," in *Proc. IEEE Int. Conf. on Image Processing*, **I**, May 98.
3. C.-Y. Lin and S.-F. Chang, "Semi-fragile watermarking for authenticating JPEG visual content," in *SPIE Security and Watermarking of Multimedia Content II*, Jan. 2000.
4. C.-Y. Lin and S.-F. Chang, "SARI: self-authentication-and-recovery image watermarking system," in *ACM Multimedia*, **4518**, Oct. 2001.
5. G.-J. Yu, C.-S. Lu, and H.-Y. M. Liao, "Mean quantization blind watermarking for image authentication," *Optical Engineering* **40**(7), pp. 1396–1408, 2001.
6. F. Alturki and R. Mersereau, "Secure fragile digital watermarking technique for image authentication," in *Proc. IEEE Int. Conf. on Image Processing*, 2001.
7. E. T. Lin, C. I. Podilchuk, and E. J. Delp, "Detection of image alterations using semi-fragile watermarks," in *Proc. of SPIE*, **3971**, pp. 152–163, 2000.
8. D. Kundur and D. Hatzinakos, "Digital watermarking for telltale tamper-proofing and authentication," *Proc. IEEE* **87**, pp. 1167–1180, July 1999.
9. B. Chen and G. W. Wornell, "Quantization index modulation: a class of provably good methods for digital watermarking and information embedding," *IEEE Trans. Inform. Theory* **47**, pp. 1423–1443, May 2001.
10. J. J. Eggers, R. Bauml, R. Tzschoppe, and B. Girod, "Scalar Costa scheme for information embedding," *IEEE Trans. Signal Processing* **51**, pp. 1003–1019, Apr. 2003.
11. J. J. K. O'Ruanaidh and T. Pun, "Rotation, scale and translation invariant spread spectrum digital image watermarking," *Signal Processing* **6**, pp. 303–317, 1998.
12. C.-Y. Lin, M. Wu, J. Bloom, M. Miller, I. Cox, and Y.-M. Lui, "Rotation, scale, and translation resilient public watermarking for images," *IEEE Trans. Image Processing* **10**, May 2001.
13. C. W. Wu, "On the design of content-based multimedia authentication systems," *IEEE Trans. Multimedia* **4**, pp. 385–393, Sept. 2002.
14. 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.
15. J. H. Conway and N. J. A. Slane, *Sphere Packings, Lattices and Groups, 3rd Edition*, Springer-Verlag, 1999.
16. G. D. Forney, Jr., "Coset codes - part I: Introduction and geometrical classification," *IEEE Trans. Inform. Theory* **34**, pp. 1123–1151, Sept. 1988.
17. G. D. Forney, Jr., "Multidimensional constellations - Part II: Voronoi constellations," *IEEE Journal on Selected Areas in Communications* **7**, pp. 941–958, Aug. 1989.
18. I. J. Cox, M. L. Miller, and J. A. Bloom, *Digital watermarking*, Morgan Kaufmann Publishers, 2002.
19. Y. Zhao, P. Campisi, and D. Kundur, "Dual domain watermarking for authentication and compression of cultural heritage images," *IEEE Trans. Image Processing* **13**, Feb. 2004.
20. S. Walton, "Image authentication for a slippery new age," *Dr. Dobb's Journal* **20**, pp. 18–26, Apr. 1995.
21. D. R. Stinson, *Cryptography Theory and Practice*, Chapman & Hall/CRC, 2 ed., 2002.
22. C. Schlegel, *Trellis coding*, IEEE Press, Piscataway, NJ, 1997.