[From *Proc. of SPIE: Multimedia Systems and Applications II*, vol. 3845]

Provably robust digital watermarking

Brian Chen and Gregory W. Wornell

Research Laboratory of Electronics and
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139

ABSTRACT

Copyright notification and enforcement, authentication, covert communication, and hybrid transmission are examples of emerging multimedia applications for digital watermarking methods, methods for embedding one signal (*e.g.*, the digital watermark) within another "host" signal to form a third, "composite" signal. The embedding is designed to achieve efficient trade-offs among the three conflicting goals of maximizing information-embedding rate, minimizing distortion between the host signal and composite signal, and maximizing the robustness of the embedding.

Quantization index modulation (QIM) methods are a class of watermarking methods that achieve provably good rate-distortion-robustness performance. Indeed, QIM methods exist that achieve performance within a few dB of capacity in the case of a (possibly colored) Gaussian host signal and an additive (possibly colored) Gaussian noise channel. Also, QIM methods can achieve capacity with a type of postprocessing called distortion compensation. This capacity is independent of host signal statistics, and thus, contrary to popular belief, the information-embedding capacity when the host signal is not available at the decoder is the same as the case when the host signal is available at the decoder.

A low-complexity realization of QIM called dither modulation has previously been proven to be better than both linear methods of spread spectrum and nonlinear methods of low-bit(s) modulation against square-error distortion-constrained intentional attacks. We introduce a new form of dither modulation called spread-transform dither modulation that retains these favorable performance characteristics while achieving better performance against other attacks such as JPEG compression.

Keywords: digital watermarking, dither modulation, quantization index modulation, capacity, spread transform, information embedding

1. INTRODUCTION

Digital watermarking and information embedding systems have a number of important multimedia applications.[1] These systems embed one signal, sometimes called an "embedded signal" or "watermark", within another signal, called a "host signal". The embedding must be done such that the embedded signal causes no serious degradation to its host. At the same time, the embedding must be robust to common degradations to the composite host and watermark signal, which in some applications result from deliberate attacks. Ideally, whenever the host signal survives these degradations, the watermark also survives.

Many of these applications relate to copyright notification and enforcement for multimedia content such as audio, video, and images that are distributed in digital formats. For example, the watermark may (1) notify a recipient of any copyright or licensing restrictions (copyright notification), (2) identify the original purchaser of the copyrighted work for future tracing of the source of illicit copies (digital fingerprinting), (3) identify the creator of the work (so that a web crawler can search for this watermark, for example), or (4) allow a standards-compliant player or duplication device to determine whether or not to duplicate or play the host signal (in a DVD copyright protection system,[2,3] for example). Other applications include (1) authentication, where the watermark may be a digital signature embedded within the host signal, (2) covert communication, where the watermark is embedded in such a way that it is difficult for an adversary to detect its presence, (3) hybrid transmission[4,5] of two different signals

The authors' email addresses and web pages are:
B. Chen: bchen@mit.edu, http://web.mit.edu/bchen/www/home.html
G. W. Wornell: gww@allegro.mit.edu, http://allegro.mit.edu/dspg/gww.html

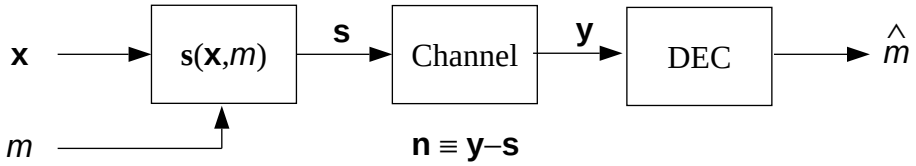

Figure 1. General information-embedding problem model. A message m is embedded in the host signal vector $\mathbf{x}$ using some embedding function $\mathbf{s}(\mathbf{x}, m)$. A perturbation vector $\mathbf{n}$ corrupts the composite signal $\mathbf{s}$. The decoder extracts an estimate $\hat{m}$ of m from the noisy channel output $\mathbf{y}$.

over the same channel in the same bandwidth, which may be used for backwards-compatible upgrading of existing analog communication systems, and (4) automated monitoring of airplay of advertisements, where advertisers embed a digital watermark within their ads and count the number of times the watermark occurs during a given broadcast period.

A number of information-embedding algorithms have been proposed[1] in this still emerging field. One class of nonlinear methods involves a quantize-and-replace strategy: after first quantizing the host signal, these systems change the quantization value to embed information. A simple example of such a system is so-called low-bit(s) modulation (LBM), where the least significant bit(s) in the quantization of the host signal are replaced by a binary representation of the embedded signal. These methods range from simple replacement of the least significant bit(s) of the pixels of an image to more sophisticated methods that involve transformation of the host signal before quantization and adjustment of the quantization step sizes.[6] Such methods have been shown to be inherently less efficient than the quantization index modulation methods discussed in this paper in terms of the amount of embedding-induced distortion for a given rate and robustness.[4,5,7] Linear classes of methods such as spread-spectrum methods embed information by linearly combining the host signal with a small pseudo-noise signal that is modulated by the embedded signal. Although these methods have received considerable attention in the literature,[8–11] linear methods in general and spread-spectrum methods in particular are limited by host-signal interference when the host signal is not known at the decoder, as is typical in many of the applications mentioned above. Intuitively, the host signal in a spread spectrum system is an additive interference that is often much larger, due to distortion constraints, than the pseudo-noise signal carrying the embedded information.

In Sec. 2 we formulate a general model of information embedding problems. Quantization index modulation (QIM) methods, which we review in Sec. 3, reject host-signal interference and, as a result, have very favorable performance characteristics in terms of their achievable trade-offs among the amount of data embedded (rate), the degradation to the host signal caused by the embedding, and the robustness of the embedding. Spread-transform dither modulation (STDM), a low-complexity realization of QIM introduced in Sec. 4, also achieves favorable rate-distortion-robustness trade-offs. Indeed, an SNR advantage of STDM over spread spectrum is demonstrated in Sec. 5, and robustness against intentional, distortion-constrained attacks are discussed in Sec. 6. Information-embedding capacities are discussed in Sec. 7, and one result in this section is that QIM methods are within a few dB of capacity in the case of a Gaussian host signal and an additive Gaussian noise channel. Furthermore, distortion-compensated QIM, a postprocessing enhancement of QIM, achieves capacity. These results apply even when the host signal and/or channel noise are colored. In fact, the capacity does not depend on the host signal statistics, and thus, the capacity in this case is the same regardless of whether or not the host signal is available during watermark decoding. Finally, simulation results are reported in Sec. 8.

2. PROBLEM MODEL

Although information-embedding applications described in Sec. 1 are quite diverse, the simple problem model of Fig. 1 captures most of their fundamental features. We wish to embed some digital information or watermark m in some host signal vector $\mathbf{x} \in \Re^N$. This host signal could be a vector of pixel values or Discrete Cosine Transform (DCT) coefficients from an image, for example. Alternatively, the host signal could be a vector of samples or transform coefficients, such as Discrete Fourier Transform (DFT) or linear prediction coding coefficients, from an audio or speech signal. We wish to embed at a rate of R_m bits per dimension (bits per host signal sample) so we can think of m as an integer, where

$$m \in \left\{1, 2, \ldots, 2^{NR_m}\right\}. \tag{1}$$

An embedding function maps the host signal $\mathbf{x}$ and embedded information m to a composite signal $\mathbf{s} \in \Re^N$. The embedding should not unacceptably degrade the host signal, so we have some distortion measure $D(\mathbf{s}, \mathbf{x})$ between the composite and host signals. For example, one might choose the square-error distortion measure

$$D(\mathbf{s}, \mathbf{x}) = \frac{1}{N}\|\mathbf{s} - \mathbf{x}\|^2. \tag{2}$$

In some cases we may measure the expected distortion $D_{\mathbf{s}} = E[D(\mathbf{s}, \mathbf{x})]$. The composite signal $\mathbf{s}$ is subjected to various common signal processing manipulations such as lossy compression, addition of random noise, and resampling, as well as deliberate attempts to remove the embedded information. These manipulations occur in some channel, which produces an output signal $\mathbf{y} \in \Re^N$. For convenience, we define a perturbation vector $\mathbf{n} \in \Re^N$ to be the difference $\mathbf{y} - \mathbf{s}$. Thus, this model is sufficiently general to include both random and deterministic, and both signal-independent and signal-dependent, perturbation vectors. The decoder forms an estimate $\hat{m}$ of the embedded information m based on the channel output $\mathbf{y}$. The robustness of the overall embedding-decoding method is characterized by the class of perturbation vectors over which the estimate $\hat{m}$ is reliable, where reliable means either that $\hat{m} = m$ deterministically or that $\Pr[\hat{m} \neq m] < \epsilon$. In general, one can specify this class of tolerable perturbation vectors in terms of a set of possible outputs $\mathcal{P}\{\mathbf{y}|\mathbf{s}\}$ for any given input in the deterministic case or in terms of a conditional probability law $p_{\mathbf{y}|\mathbf{s}}(\mathbf{y}|\mathbf{s})$ in the probabilistic case. In some special cases, several examples of which appear later in this paper, one can conveniently characterize the size of this tolerable class of perturbations, and hence the robustness, with a single parameter.

One desires the embedding system to have high rate, low distortion, and high robustness, but in general these three goals tend to conflict. Thus, the performance of an information embedding system is characterized in terms of its achievable rate-distortion-robustness trade-offs.

3. QUANTIZATION INDEX MODULATION METHODS

One class of embedding methods that achieves very good, and in some cases optimal, rate-distortion-robustness trade-offs are so-called quantization index modulation (QIM) methods.[4,5,7] In this section, we review the basic principles behind this class of methods.

One can view the embedding function $\mathbf{s}(\mathbf{x}, m)$ as an ensemble of functions of $\mathbf{x}$, each function in the ensemble indexed by m. We denote the functions as $\mathbf{s}(\mathbf{x}; m)$ to emphasize this view. If the embedding-induced distortion is to be small, then each function must be an approximate identity function in some sense so that $\mathbf{s}(\mathbf{x}; m) \approx \mathbf{x}$ for all m. If these approximate identities are quantizers, then the embedding method is a QIM method.

Thus, quantization index modulation refers to embedding information by first modulating an index or sequence of indices with the embedded information and then quantizing the host signal with the associated quantizer or sequence of quantizers. Fig. 2 illustrates QIM in the case where one bit is to be embedded so that $m \in \{1, 2\}$. Thus, we require two quantizers, and their corresponding sets of reconstruction points in $\Re^N$ are represented in Fig. 2 with $\times$'s and $\circ$'s. If $m = 1$, for example, the host signal is quantized with the $\times$-quantizer, *i.e.*, $\mathbf{s}$ is chosen to be the $\times$ closest to $\mathbf{x}$. If $m = 2$, $\mathbf{x}$ is quantized with the $\circ$-quantizer. The sets of reconstruction points are non-intersecting as no $\times$ point is the same as any $\circ$ point. This non-intersection property leads to host-signal interference rejection. As $\mathbf{x}$ varies, the composite signal value $\mathbf{s}$ varies from one $\times$ point ($m = 1$) to another or from one $\circ$ point ($m = 2$) to another, but it never varies between a $\times$ point and a $\circ$ point. Thus, even with an infinite energy host signal, one can determine m if channel perturbations are not too severe. The $\times$ points and $\circ$ points are both quantizer reconstruction points for $\mathbf{x}$ and signal constellation points for communicating m. (One *set* of points, rather than one individual point, exists for each possible value of m.) Thus, we may view design of QIM systems as the simultaneous design of an ensemble of quantizers (or source codes) and signal constellations (or channel codes).

The structure of QIM systems is convenient from an engineering perspective since properties of the quantizer ensemble can be connected to the performance parameters of rate, distortion, and robustness. For example, the number of quantizers in the ensemble determines the number of possible values of m, or equivalently, the rate. The sizes and shapes of the quantization cells, one of which is represented by the dashed polygon in Fig. 2, determines the amount of embedding-induced distortion, all of which arises from quantization error. Finally, for many classes of channels the minimum distance $d_{\min}$ between the sets of reconstruction points of different quantizers in the ensemble determines the robustness of the embedding. We define the minimum distance to be

$$d_{\min} \triangleq \min_{(i,j):i\neq j} \min_{(\mathbf{x}_i, \mathbf{x}_j)} \|\mathbf{s}(\mathbf{x}_i; i) - \mathbf{s}(\mathbf{x}_j; j)\|. \tag{3}$$

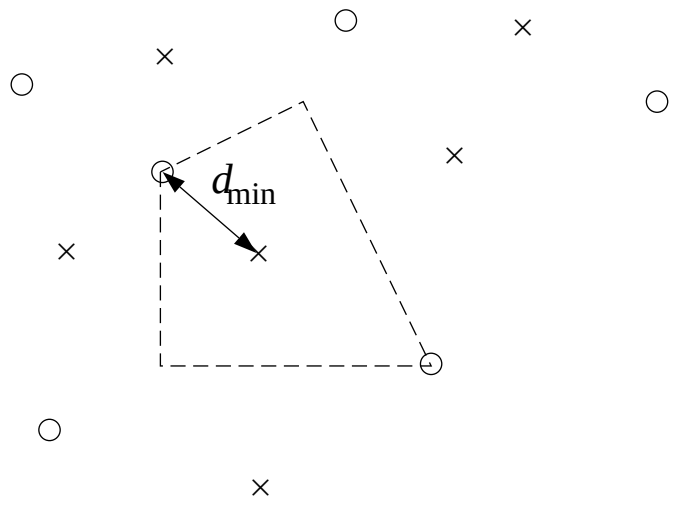

Figure 2. Quantization index modulation for information embedding. The points marked with ×'s and ∘'s belong to two different quantizers, each with its associated index. The minimum distance $d_{\min}$ measures the robustness to perturbations, and the sizes of the quantization cells, one of which is shown in the figure, determine the distortion. If $m = 1$, the host signal is quantized to the nearest ×. If $m = 2$, the host signal is quantized to the nearest ∘.

Intuitively, the minimum distance measures the size of perturbation vectors that can be tolerated by the system. For example, as long as the length of the perturbation vector is less than half the minimum distance,

$$d_{\min} > 2\|\mathbf{n}\|, \tag{4}$$

then a minimum distance decoder will not make an error. In the case of an additive white Gaussian noise channel with a noise variance of σ_n^2, at high signal-to-noise ratio the minimum distance also characterizes the error probability of the minimum distance decoder,[12]

$$\Pr[\hat{m} \neq m] \sim Q\left(\sqrt{\frac{d_{\min}^2}{4\sigma_n^2}}\right).$$

The minimum distance decoder to which we refer simply chooses the reconstruction point closest to the received vector, *i.e.*,

$$\hat{m}(\mathbf{y}) = \arg\min_m \min_{\mathbf{x}} \|\mathbf{y} - \mathbf{s}(\mathbf{x}; m)\|. \tag{5}$$

If, which is often the case, the quantizers $\mathbf{s}(\mathbf{x}; m)$ map $\mathbf{x}$ to the nearest reconstruction point, then (5) can be rewritten as

$$\hat{m}(\mathbf{y}) = \arg\min_m \|\mathbf{y} - \mathbf{s}(\mathbf{y}; m)\|. \tag{6}$$

4. SPREAD-TRANSFORM DITHER MODULATION

Dithered quantizers[13] are quantizer ensembles where the quantization cells and reconstruction points of any given quantizer in the ensemble are shifted versions of the quantization cells and reconstruction points of any other quantizer in the ensemble. Dither modulation systems[7] embed information by modulating the amount of the shift, which is called the dither vector, by the embedded signal, *i.e.*, each possible embedded signal maps uniquely onto a different dither vector $\mathbf{d}(m)$. The host signal is quantized with the resulting dithered quantizer to form the composite signal. Specifically, we start with some base quantizer $\mathbf{q}(\cdot)$, and the embedding function is $\mathbf{s}(\mathbf{x}; m) = \mathbf{q}(\mathbf{x} + \mathbf{d}(m)) - \mathbf{d}(m)$.

A simple example of dither modulation that will be of interest in this paper is called binary spread-transform dither modulation (STDM) with uniform, scalar quantization of step size Δ. We assume that $1/N \leq R_m \leq 1$. Specifically, STDM involves the following steps:

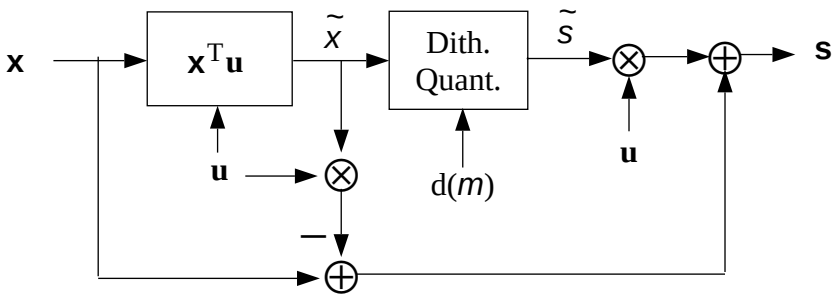

Figure 3. Spread-transform dither modulation. Information is embedded in the projection of a block $\mathbf{x}$ of the host signal onto $\mathbf{u}$, which is typically a pseudorandom vector. Components of $\mathbf{x}$ orthogonal to $\mathbf{u}$ are added back to the signal after dithered quantization to form the corresponding block of the composite signal $\mathbf{s}$.

- The NR_m information bits $\{b_1, b_2, \ldots, b_{NR_m}\}$ representing the embedded message m are error correction coded using a rate-k_u/k_c code to obtain a coded bit sequence $\{z_1, z_2, \ldots, z_{N/L}\}$, where
$$L = \frac{1}{R_m}(k_u/k_c).$$
- A spreading vector $\mathbf{u} \in \Re^L$ is chosen along with two sequences $d(\cdot, 0)$ and $d(\cdot, 1)$ of N/L dither values satisfying the constraint
$$d(i,1) = \begin{cases} d(i,0) + \Delta/2, & d(i,1) < 0 \\ d(i,0) - \Delta/2, & d(i,1) \geq 0 \end{cases}$$
for $1 \leq i \leq N/L$. For example, one could choose $d(i,0)$ pseudorandomly with a uniform distribution over $[-\Delta/2, \Delta/2]$.*
- The projection of the i-th length-L subvector of $\mathbf{x}$ onto $\mathbf{u}$ is quantized with a dithered quantizer using dither value $d(i, z_i)$.

A block diagram of this embedding process is shown in Fig. 3.

STDM has a number of advantages over earlier forms[7] of dither modulation. One advantage is that the STDM signal constellation has fewer "nearest neighbors",[4] which usually results in a lower probability of decoding error. Another advantage is that one can easily convert existing amplitude-modulation spread spectrum (AM-SS) systems, a class of previously proposed spread spectrum methods that have embedding functions of the form

$$\mathbf{s}(\mathbf{x}, m) = \mathbf{x} + a(m)\mathbf{u},$$

into spread-transform dither modulation systems by replacing addition with quantization. This property is useful if one has already invested considerable effort in optimizing a spread spectrum system, for example, by exploiting perceptual properties of the human visual and auditory systems or designing receiver front-ends to mitigate effects of geometric distortion. Specifically, the AM-SS embedding function can be re-written in the form

$$\mathbf{s}(\mathbf{x}, m) = (\tilde{x} + a(m))\mathbf{u} + (\mathbf{x} - \tilde{x}\mathbf{u}),$$

where $\tilde{x} = \mathbf{x}^T\mathbf{u}$. We see that AM-SS is equivalent to adding a value $a(m)$ to the projection $\tilde{x}$ of the host signal onto the spreading vector $\mathbf{u}$. Thus, if one has designed a good spread spectrum system, for example, by designing a $\mathbf{u}$ that has good perceptual distortion properties, but would like to gain the advantages of dither modulation, one can do so simply by replacing the addition step of AM-SS,

$$\tilde{s} = \tilde{x} + a(m), \tag{7}$$

by the quantization step of STDM,

$$\tilde{s} = q(\tilde{x} + d(m)) - d(m). \tag{8}$$

*A uniform distribution for the dither sequence implies that the quantization error is statistically independent of the host signal and leads to fewer "false contours", both of which are generally desirable properties from a perceptual viewpoint.[13]

5. SNR ADVANTAGE OF STDM OVER AM SPREAD SPECTRUM

The close coupling of STDM and AM spread spectrum allows a direct comparison that results in a provable robustness advantage of STDM in terms of a "signal-to-noise ratio at the decision device". This performance advantage results from the host signal interference rejection properties of QIM methods in general.

We consider embedding one bit in a length-L block $\mathbf{x}$ using STDM and AM spread spectrum methods with the same spreading vector $\mathbf{u}$, which is of unit length. Because the embedding occurs entirely in the projections of $\mathbf{x}$ onto $\mathbf{u}$, the problem is reduced to a one-dimensional problem with the embedding functions (7) and (8). For AM-SS (7), $a(m) = \pm\sqrt{LD_{\mathbf{s}}}$ so that

$$|a(1) - a(2)|^2 = 4LD_{\mathbf{s}}. \tag{9}$$

For STDM (8),

$$\min_{(\tilde{x}_1,\tilde{x}_2)} |\tilde{s}(\tilde{x}_1, 1) - \tilde{s}(\tilde{x}_2, 2)|^2 = \Delta^2/4 = 3LD_{\mathbf{s}}, \tag{10}$$

where $\Delta = \sqrt{12LD_{\mathbf{s}}}$ so that the expected distortion in both cases is the same, and where we have used the fact that $d(1)$ and $d(2)$ are chosen such that $|d(1) - d(2)| = \Delta/2$. Because all of the embedding-induced distortion occurs only in the direction of $\mathbf{u}$, the distortion in both cases also has the same time or spatial distribution and frequency distribution. Thus, one would expect that any perceptual effects due to time/space masking or frequency masking are the same in both cases. Therefore, square-error distortion may be a more meaningful measure of distortion when comparing STDM with AM-SS than one might expect in other more general contexts where square-error distortion may fail to capture certain perceptual effects.

The decoder in both cases makes a decision based on $\tilde{y}$, the projection of the channel output $\mathbf{y}$ onto $\mathbf{u}$. In the case of AM-SS,

$$\tilde{y} = a(m) + \tilde{x} + \tilde{n},$$

while in the case of STDM,

$$\tilde{y} = \tilde{s}(\tilde{x}, m) + \tilde{n},$$

where $\tilde{n}$ is the projection of the perturbation vector $\mathbf{n}$ onto $\mathbf{u}$. We let $P(\cdot)$ be some measure of energy. For example, $P(x) = x^2$ in the case of a deterministic variable x, or $P(x)$ equals the variance of the random variable x. The energy of the interference or "noise" is $P(\tilde{x} + \tilde{n})$ for AM-SS, but only $P(\tilde{n})$ for STDM, *i.e.*, the host signal interference for STDM is zero. Thus, the signal-to-noise ratio at the decision device is

$$\mathrm{SNR}_{\mathrm{AM-SS}} = \frac{4LD_{\mathbf{s}}}{P(\tilde{x} + \tilde{n})}$$

for AM-SS and

$$\mathrm{SNR}_{\mathrm{STDM}} = \frac{3LD_{\mathbf{s}}}{P(\tilde{n})}$$

for STDM, where the "signal" energies $P(a(1) - a(2))$ and $P\left(\min_{(\tilde{x}_1,\tilde{x}_2)} |\tilde{s}(\tilde{x}_1, 1) - \tilde{s}(\tilde{x}_2, 2)|\right)$ are given by (9) and (10). Thus, the advantage of STDM over AM-SS is

$$\frac{\mathrm{SNR}_{\mathrm{STDM}}}{\mathrm{SNR}_{\mathrm{AM-SS}}} = \frac{3}{4}\frac{P(\tilde{x} + \tilde{n})}{P(\tilde{n})}, \tag{11}$$

which is typically very large since the channel perturbations $\tilde{n}$ are usually much smaller than the host signal $\tilde{x}$ if the channel output $\tilde{y}$ is to be of reasonable quality. For example, if the host signal-to-channel noise ratio is 30 dB and $\tilde{x}$ and $\tilde{n}$ are uncorrelated, then the SNR advantage (11) of STDM over AM spread spectrum is 28.8 dB. On the other hand, even if the host signal interference were zero, for example, such as would be the case if $\tilde{x}$ were known at the decoder and thus could be subtracted from $\tilde{y}$, then STDM would be worse than AM-SS by only 4/3 or 1.25 dB. (This gap can be eliminated for general QIM methods, which may involve non-dithered vector quantizers, at least in the case of additive Gaussian noise channels.[4])

6. ROBUSTNESS AGAINST INTENTIONAL ATTACKS

Earlier forms of dither modulation have been shown to achieve provably better rate-distortion-robustness performance than spread spectrum and low-bit(s) modulation (LBM) methods against square-error distortion-constrained attacks.[7] It can be shown[4] that the spread-transform dither modulation method of Sec. 4 has the same minimum distance for a given amount of embedding-induced distortion as the earlier form[7] of dither modulation, and thus, STDM methods also exhibit these same performance advantages. We discuss these results below.

6.1. Bounded Perturbation Attacks

One way to quantify the distortion constraint faced by an attacker is to bound the size of the perturbation vectors, *i.e.*,

$$\|\mathbf{n}\|^2 \leq N\sigma_n^2, \tag{12}$$

where σ_n^2 is the maximum perturbation energy per dimension. This bounded perturbation channel model describes a maximum distortion† or minimum SNR constraint between the channel input and output. When facing these bounded perturbation attacks, the robustness of a digital watermarking system is conveniently characterized by the largest σ_n^2 such that one can deterministically guarantee that $m = \hat{m}$ for every $\mathbf{n}$ satisfying (12).

By calculating the minimum distance (3) of an embedding method in terms of the rate and embedding-induced distortion, one can use (12) and the error-free decoding condition (4) to determine achievable rate-distortion-robustness trade-offs of the particular embedding method against bounded perturbation attacks. Such an analysis has been done for earlier forms of dither modulation[7] and can straightforwardly be extended[4] to show that STDM with uniform scalar quantization can achieve the following rate-distortion-robustness trade-offs:

$$R_m < \gamma_c \frac{3}{4} \frac{D_{\mathbf{s}}}{N\sigma_n^2}, \tag{13}$$

where γ_c is the error correction coding gain (the product of the Hamming distance and rate of the error correction code). This expression gives an achievable set of embedding rates for a given expected distortion $D_{\mathbf{s}}$ and channel perturbation energy per dimension σ_n^2 when one wishes to deterministically guarantee error-free decoding with finite length signals. Thus, one can view (13) as a deterministic counterpart to the conventional, information-theoretic notion of the capacity[14] of a random channel. Spread spectrum methods in contrast offer no such guaranteed robustness to bounded perturbation attacks because their minimum distance is zero.[7] Finally, the achievable rate-distortion-robustness trade-offs of coded LBM with uniform scalar quantization are 2.43 dB worse than those of (13).[4,7]

6.2. Bounded Host-distortion Attacks

Some attackers may work with a distortion constraint between the host signal, rather than the channel input, and the channel output since this distortion is the most direct measure of degradation to the host signal. For example, if an attacker has partial knowledge of the host signal, which may be in the form of a probability distribution so that he or she can calculate this distortion, then it may be appropriate to bound the expected distortion $D_{\mathbf{y}} = E[D(\mathbf{y}, \mathbf{x})]$, where this expectation is taken over the probability density of $\mathbf{x}$ given the channel input $\mathbf{s}$. We refer to this channel model as the bounded host-distortion channel to emphasize that the attacker's distortion is measured relative to the the host signal.

For these bounded host-distortion channels, it can be shown[4,7] that an in-the-clear attacker, one who knows everything about the embedding and decoding processes including any keys, can remove spread spectrum and LBM embedded watermarks and improve the signal quality ($D_{\mathbf{y}} \leq D_{\mathbf{s}}$) at the same time. In contrast, to remove a watermark embedded with QIM methods (including STDM and earlier forms[7] of dither modulation), the in-the-clear attacker's distortion $D_{\mathbf{y}}$ must be greater than the embedding-induced distortion $D_{\mathbf{s}}$.

7. RANDOM CHANNELS

Probabilistic channel models are an alternative to the deterministic channel models of Sec. 6, especially in cases where perturbations may be unbounded. In these cases, an alternative notion of robustness can be developed if the channel can be modeled as random and memoryless, with some conditional probability density function (pdf) $p_{y|s}(y|s)$ of the channel output y given the channel input s, which in turn determines the conditional pdf $p_{n|s}(n|s)$ of the channel perturbation. As is well known,[14] one can use long channel codes to communicate over such channels robustly in the sense that the probability of error $\Pr[\hat{m} \neq m]$ can approach zero asymptotically with long signal lengths even if there is no guarantee that $\hat{m} = m$ with finite signal lengths. In this section we explore the achievable performance limits of information-embedding methods over these random, memoryless channels.

†Some types of distortion, such as geometric distortions can be large in terms of square error, yet still be small perceptually. However, in some cases these distortions can be mitigated either by preprocessing at the decoder or by embedding information in parameters of the host signal that are less affected (in terms of square error) by these distortions. For example, a simple delay or shift may cause large square error, but the magnitude of the DFT coefficients are relatively unaffected.

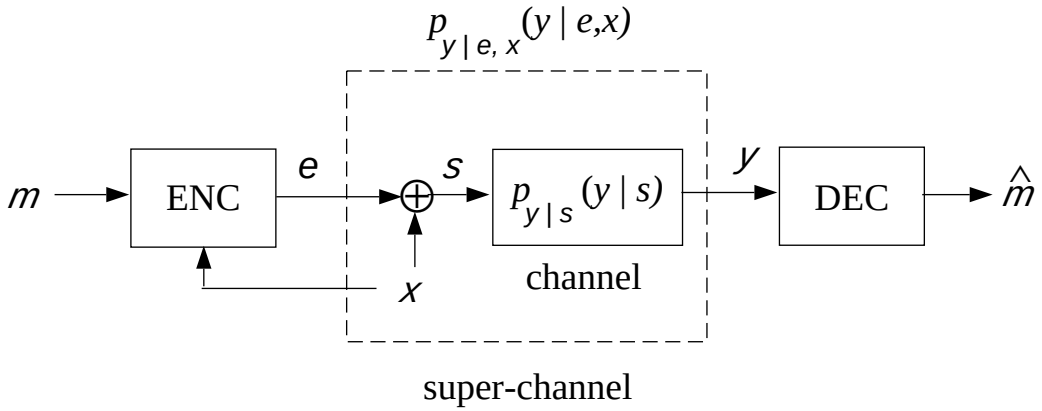

Figure 4. Equivalent super-channel for information embedding. The composite signal is the sum of the host signal, which is the state of the super-channel, and a host-dependent distortion signal.

Our approach is to consider the host signal to be the state of a channel that is known at the encoder, as illustrated in Fig. 4. Any embedding function can be written as a sum of the host signal $\mathbf{x}$ and a host-dependent distortion signal $\mathbf{e}$,

$$\mathbf{s}(\mathbf{x}, m) = \mathbf{x} + \mathbf{e}(\mathbf{x}, m),$$

simply by defining the distortion signal to be $\mathbf{e} \stackrel{\Delta}{=} \mathbf{s} - \mathbf{x}$. Thus, one can interpret the embedding process as mapping the embedded information m onto a host-dependent distortion signal subject to a distortion constraint such as

$$\frac{1}{N}\sum_{i=1}^{N} e^2 \leq D_{\mathbf{s}}.$$

This distortion signal is the input to a super-channel that has the host signal as a state. The capacity[14] of this super-channel is the reliable information-embedding rate R_m that is asymptotically achievable with long signal lengths N.

When the elements of the state vector $\mathbf{x}$ are independent and identically distributed (iid) random variables and the encoder sees the entire state vector before choosing the channel input $\mathbf{e}$, this capacity is[15]

$$C = \max_{p_{u,e|x}(u,e|x)} I(u;y) - I(u;x), \tag{14}$$

where $I(\cdot;\cdot)$ denotes mutual information and u is an auxiliary random variable. In the case of watermarking, the maximization (14) is subject to a distortion constraint $E[e^2] \leq D_{\mathbf{s}}$.

In the case where the channel is an additive white Gaussian noise (AWGN) channel and the host signal is also white and Gaussian, the capacity (14) is[16]

$$C_{\text{Gauss}} = \frac{1}{2}\log_2\left(1 + \frac{D_{\mathbf{s}}}{\sigma_n^2}\right) = \frac{1}{2}\log_2(1 + \text{DNR}), \tag{15}$$

where DNR is the (embedding-induced) distortion-to-noise ratio $D_{\mathbf{s}}/\sigma_n^2$. One can show[5] that (15) also gives the capacity in the case of colored host signals, colored channel noise, and an appropriate definition of DNR that reflects the colored nature of the channel noise. Thus, rather remarkably, the capacity is independent of host signal statistics, implying that an infinite energy host signal causes no decrease in capacity in the Gaussian case and that one can do just as well when the host signal is not known at the decoder as when the host signal is known at the decoder.

This principle suggests that optimal and near-optimal digital watermarking methods have some sort of host-signal interference rejection capability. Recall, that QIM methods possess such a capability, and as a result, one can show[4] that there exist near capacity-achieving QIM methods, as illustrated in Fig. 5, which shows an upper bound on the "gap" between QIM and capacity.[4] This gap is the additional amount of DNR that the best possible QIM system

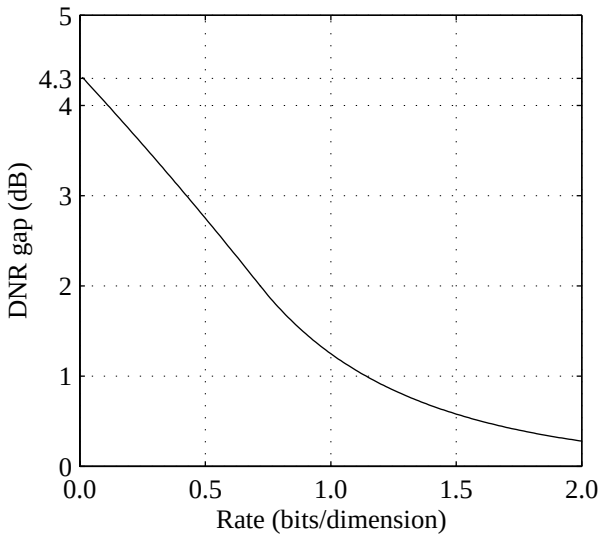

Figure 5. Gap between QIM and Gaussian capacity. The maximum gap is a factor of e, which is approximately 4.3 dB, and approaches 0 dB asymptotically at high rates.

needs to achieve the same rate as a capacity-achieving system. Thus, for example, to embed at a rate of 1 bit per host signal sample with a QIM system, one would need to accept at most about 1.3 dB more embedding-induced distortion than would otherwise be necessary using a capacity-achieving system on a channel with a given noise variance. (Equivalently, for a fixed amount of embedding-induced distortion, one could tolerate at most 1.3 dB more noise with the capacity-achieving system than with the QIM system.) This gap is at most a factor of $e \approx 4.3$ dB at any finite rate and asymptotically approaches 0 dB at high rates. Furthermore, even at all finite rates this gap can be eliminated using a type of postprocessing called distortion compensation,[4,17] *i.e.*, capacity-achieving distortion-compensated QIM systems exist.

In contrast, spread spectrum methods do not reject host signal interference. Thus, the achievable rate of a spread spectrum method is the Gaussian channel capacity, treating both x and n as interference sources. As is well-known,[4,10] when both x and n are white, this capacity is

$$C_{\mathrm{ss}} = \frac{1}{2}\log_2\left(1 + \frac{D_{\mathsf{s}}}{\sigma_x^2 + \sigma_n^2}\right) = \frac{1}{2}\log_2\left(1 + \frac{\mathrm{DNR}}{\mathrm{SNR}_x + 1}\right), \tag{16}$$

where SNR_x is the ratio between the host signal variance and the channel noise variance. (This rate is also the capacity when n is non-Gaussian, but still independent of s, and a correlation detector is used for decoding.[18]) By comparing (16) to (15) we see that the gap to capacity of spread-spectrum is $\mathrm{SNR}_x + 1$. Typically, SNR_x is very large since the channel noise is not supposed to degrade signal quality too much. Thus, in these cases the gap to capacity of spread-spectrum is much larger than the gap to capacity of QIM, which again is never larger than about 4.3 dB.

8. SIMULATION RESULTS

Some simulation results for dither modulation implementations are reported below for both Gaussian and non-Gaussian channels.

8.1. Gaussian Channel

It can be shown[4] that the bit-error probability of *uncoded* STDM with uniform, scalar quantization is about 10^{-6} on the AWGN channel at a rate-normalized distortion-to-noise ratio ($\mathrm{DNR}_{\mathrm{norm}}$) of 15 dB, where

$$\mathrm{DNR}_{\mathrm{norm}} \triangleq \frac{\mathrm{DNR}}{R_{\mathsf{m}}},$$

Figure 6. Composite (left) and AWGN channel output (right) images. The composite and channel output images have peak signal-to-distortion ratios of 34.9 dB and 22.6 dB, respectively. DNR $= -12.1$ dB, yet all bits were extracted without error. $R_m = 1/512$ and $\text{DNR}_{\text{norm}} = 15.0$ dB so the actual bit-error probability is 10^{-6}.

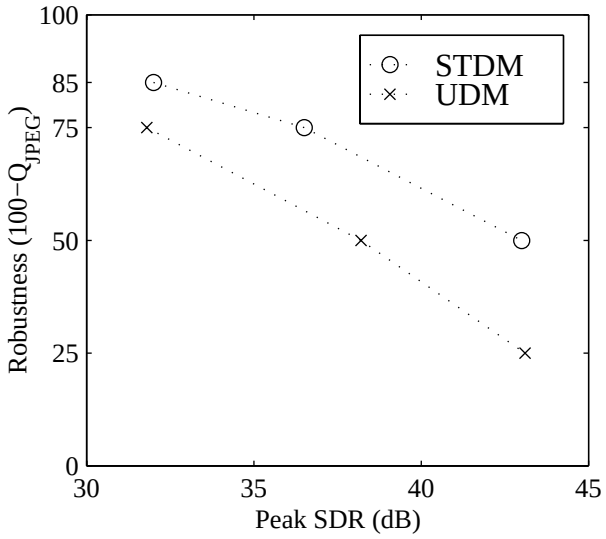

Figure 7. Achievable robustness-distortion trade-offs of dither modulation on the JPEG channel. $R_m = 1/320$. The bit-error rate is less than 5×10^{-6}.

Thus, one can embed using uncoded STDM for very noisy AWGN channels by choosing sufficiently low rates,

$$R_m \leq \frac{\text{DNR}}{\text{DNR}_{\text{norm}}}.$$

This case is illustrated in Fig. 6, where despite the fact that the channel has degraded the composite image by over 12 dB, all 512 embedded bits are recovered without any errors from the 512-by-512 image. The actual bit-error probability is about 10^{-6}.

8.2. JPEG Channel

The robustness of digital watermarking algorithms to common lossy compression algorithms such as JPEG is of considerable interest. A natural measure of robustness is the worst tolerable JPEG quality factor (The JPEG quality factor is a number between 0 and 100, 0 representing the most compression and lowest quality, and 100 representing the least compression and highest quality.) for a given bit-error rate at a given distortion level and embedding rate. We experimentally determined achievable rate-distortion-robustness operating points for particular uncoded implementations of both STDM and an earlier form[7] of dither modulation, which we will refer to for convenience as unspread dither modulation (UDM).

These achievable distortion-robustness trade-offs at an embedding rate of $R_m = 1/320$ bits per grayscale pixel are shown in Fig. 7 at various JPEG quality factors (Q_{JPEG}). The peak signal-to-distortion ratio (SDR) is defined as the

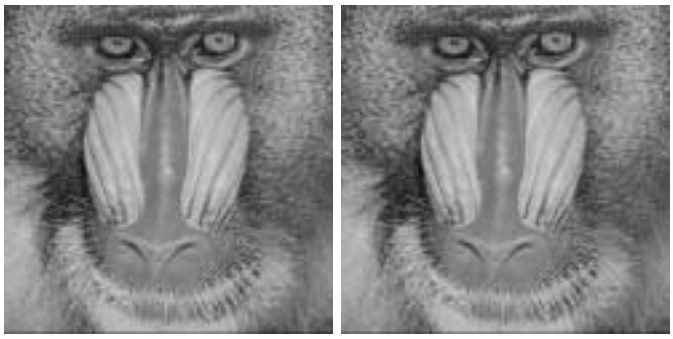

Figure 8. Host (left) and composite (right) image. After 25%-quality JPEG compression of the composite image, all bits were extracted without error. $R_m = 1/320$. Peak SDR of composite image is 36.5 dB.

ratio between the square of the maximum possible pixel value and the average embedding-induced distortion per pixel. The host and composite signals, both 512-by-512 images, are shown in Fig. 8. The actual embedding is performed in the DCT domain using 8-by-8 blocks ($f_1, f_2 \in \{0, 1/16, \ldots, 7/16\}$) and low frequencies $\left(\sqrt{f_1^2 + f_2^2} \leq 1/4\right)$, with 1 bit embedded across 5 DCT blocks. STDM is better than UDM by about 5 dB at $(100 - \mathrm{Q_{JPEG}})$ of 50 and 75. As mentioned in Sec. 4, one explanation for this performance advantage involves the smaller number of "nearest neighbors"[4] of STDM.

Although no bit errors occurred during the simulations used to generate Fig. 7, we estimate the bit-error rate to be at most 5×10^{-6}. At an embedding rate of 1/320, one can only embed 819 bits in the host signal image, which is not enough to measure bit-error rates this low. However, one can estimate an upper bound on the bit-error rate by measuring the bit-error rate ϵ at an embedding rate five times higher ($R_m = 1/64$) and calculating the coded bit-error probability of a rate-1/5 repetition code when the uncoded error probability is ϵ assuming independent errors, which can approximately be obtained by embedding the repeated bits in spatially separated places in the image. This coded bit-error probability is

$$P_{\mathrm{rep}} = \sum_{k=3}^{5} \left(\begin{array}{c} 5 \\ k \end{array} \right) \epsilon^k (1-\epsilon)^{5-k} \tag{17}$$

If $\epsilon \leq 32/4096$, then (17) implies $P_{\mathrm{rep}} \leq 4.7 \times 10^{-6}$. Thus, to obtain Fig. 7, we first embedded at a rate of 1/64 adjusting the SDR until $\epsilon \leq 32/4096$. Then, we embedded at a rate of 1/320 using a rate-1/5 repetition code to verify that no bit errors occurred.

ACKNOWLEDGMENTS

This work has been supported in part by the Office of Naval Research under Grant No. N00014-96-1-0930, by the Air Force Office of Scientific Research under Grant No. F49620-96-1-0072, by the MIT Lincoln Laboratory ACC, and by a National Defense Science and Engineering Graduate Fellowship.

REFERENCES

1. M. D. Swanson, M. Kobayashi, and A. H. Tewfik, "Multimedia data-embedding and watermarking technologies," *Proceedings of the IEEE* **86**, pp. 1064–1087, June 1998.
2. I. J. Cox and J.-P. M. G. Linnartz, "Some general methods for tampering with watermarks," *IEEE Journal on Selected Areas in Communications* **16**, pp. 587–593, May 1998.
3. J.-P. Linnartz, T. Kalker, and J. Haitsma, "Detecting electronic watermarks in digital video," in *Proc. of the 1999 IEEE Int. Conf. on Acoustics, Speech, and Signal Proc.*, vol. 4, pp. 2071–2074, (Phoenix, AZ), Mar. 1999.
4. B. Chen and G. W. Wornell, "Quantization index modulation: A class of provably good methods for digital watermarking and information embedding," submitted to *IEEE Transactions on Information Theory*, 1999.
5. B. Chen and G. W. Wornell, "Quantization index modulation methods for digital watermarking and information embedding," submitted to *Journal of VLSI Signal Proc. Systems*, 1999.

6. M. D. Swanson, B. Zhu, and A. H. Tewfik, "Data hiding for video-in-video," in *Proceedings of the 1997 IEEE International Conference on Image Processing*, vol. 2, pp. 676–679, (Piscataway, NJ), 1997.
7. B. Chen and G. W. Wornell, "Dither modulation: A new approach to digital watermarking and information embedding," in *Proc. of SPIE: Security and Watermarking of Multimedia Contents*, vol. 3657, pp. 342–353, (San Jose, CA), Jan. 1999.
8. W. Bender, D. Gruhl, N. Morimoto, and A. Lu, "Techniques for data hiding," *IBM Systems Journal* **35**(3-4), pp. 313–336, 1996.
9. I. J. Cox, J. Killian, T. Leighton, and T. Shamoon, "A secure, robust watermark for multimedia," in *Information Hiding. First International Workshop Proceedings*, pp. 185–206, June 1996.
10. J. R. Smith and B. O. Comiskey, "Modulation and information hiding in images," in *Information Hiding. First International Workshop Proceedings*, pp. 207–226, June 1996.
11. J. R. Hernandez, F. Perez-Gonzalez, J. M. Rodriguez, and G. Nieto, "Performance analysis of a 2-D-multipulse amplitude modulation scheme for data hiding and watermarking of still images," *IEEE Journal on Selected Areas in Communications* **16**, pp. 510–524, May 1998.
12. E. A. Lee and D. G. Messerschmitt, *Digital Communication*, Kluwer Academic Publishers, 2nd ed., 1994.
13. N. S. Jayant and P. Noll, *Digital Coding of Waveforms: Principles and Applications to Speech and Video*, Prentice-Hall, 1984.
14. T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, Inc., 1991.
15. S. I. Gel'fand and M. S. Pinsker, "Coding for channel with random parameters," *Problems of Control and Information Theory* **9**(1), pp. 19–31, 1980.
16. M. H. M. Costa, "Writing on dirty paper," *IEEE Trans. on Information Thy.* **29**, pp. 439–441, May 1983.
17. B. Chen and G. W. Wornell, "System, method, and product for distortion-compensated information embedding using an ensemble of non-intersecting embedding generators." U.S. patent pending. Licensing info: MIT Technology Licensing Office.
18. A. Lapidoth, "Nearest neighbor decoding for additive non-Gaussian noise channels," *IEEE Transactions on Information Theory* **42**, pp. 1520–1529, Sept. 1996.

