Published in the Proceedings of the IEEE, 87, 7, 1127-1141, 1999.

Watermarking as communications with side information

Ingemar J. Cox⋆, Matt L. Miller† and Andrew L. McKellips‡

Abstract

Several authors have drawn comparison between embedded signaling or watermarking and communications, especially spread spectrum communications. We examine the similarities and differences between watermarking and traditional communications. Our comparison suggests that watermarking most closely resembles communications with side information at the transmitter and or detector, a configuration originally described by Shannon. This leads to several novel characteristics and insights regarding embedded signaling which are discussed in detail.

1 Introduction

Watermarking is a process in which a signal is hidden or embedded into another signal, usually a photograph, video or music. There are a variety of possible uses for embedded signaling, ranging from covert signaling applications that encompass classical steganography, to recent commercial interest in providing copyright and copy control information. In the latter case, an advantage of embedded signaling is that the copy control information is embedded directly into the media to be protected and is therefore independent of the broadcast or transmission format and remains present even after decryption [BCK$^+$99]. The reader is directed to [CM97] and to articles in this current issue for a review of watermarking methods.

Watermarking is, of course, a form of communications. The requirement that the fidelity of the media content must not be impaired implies that the magnitude of the watermark signal must be very small in comparison to the content signal, analagous to a stringent power constraint in traditional communications. This characteristic, together with the widespread view that, from the perspective of watermark detection, the content is noise, has led several authors to think of watermarking as a form of spread spectrum communications.

When media content is viewed purely as noise, no advantage is taken of the fact that the content is completely known to the watermark embedder (and detector, if the original unwatermarked content is available as part of the detection process). We therefore prefer to view watermarking as an example of communication with side information. This form of communication was originally introduced by Shannon [Sha58] who was interested in calculating the capacity of a channel when the transmitter

[0] ⋆Post: NEC Research Institute, 4 Independence Way, Princeton, NJ 08540.
Email: ingemar@research.nj.nec.com

[0] †Post: Signafy Inc., 4 Independence Way, Princeton, NJ 08540.
Email: mlm@signafy.com

[0] ‡Post:Work was performed while the author was at Department of Electrical Engineering, Princeton University, Princeton, NJ 08544. The author is currently at MIT Lincoln Laboratory, 244 Wood Street, Lexington, MA 02420-9108
Email: mckellips@ll.mit.edu

and/or receiver had information regarding the state of channel noise. We believe that modeling watermarking as communication with side information allows more effective watermark insertion and detection methods to be designed.

In Section 2 we provide a high level view of three paradigms for watermarking in order to motivate the remainder of the paper. In Section 3, we introduce a basic framework for the watermarking process, including notation used throughout the paper. In Section 4, we compare watermarking and traditional communication models, demonstrating a strong similarity between the former and a communication channel with side information. In order to utilize this communications model in the design of a watermark insertion algorithm, it is necessary to have knowledge of the underlying statistics of the content and the distortions it is likely to experience. These issues are discussed in Section 5. In Sections 6 and 7 we analyze watermark insertion and detection algorithms from the point of view of communications with side information. Conclusions are presented in Section 8.

2 Three paradigms for watermarking

We believe that prior work in watermarking can be divided into two distinct categories. Much early work in watermarking falls into Category 1, which essentially consists of directly adding a watermark signal, $\mathbf{w}$ to the content, $\mathbf{C_0}$ to produce a watermarked version $\mathbf{C_w}$. Category 1 watermarking is characterized at the detector by considering the content $\mathbf{C_0}$ as noise. When the original content is available to the detector, this "noise" is subtracted from the received signal prior to detection.[1] Insertion of the watermark in a Category 1 system is characterized by the fact that the content modification is independent of the content being modified. Thus, for example, perceptual modeling is not applied to the content and the signal that is added is not adjusted based on this information.

Cox *et al* [CKLS97] suggested that perceptual modeling had significant utility for watermarking. Their early paper highlighted an idea that is somewhat counter intuitive. If a watermark is to be robust, i.e. survive, common signal processing that is routinely performed on the content, then the watermark signal must be placed in the perceptually significant components of the content. This is clearly true if, for example, a watermark is intended to survive lossy compression, since the goal of such compression algorithms is to remove the perceptually insignificant, i.e. redundant, components of the content. Of course, such a requirement conflicts with the desire that the watermark also be invisible. In order to meet both these requirements, Cox *et al* proposed a form of spread spectrum coding such that each perceptually significant component is only changed by a small amount. Subsequently, there have been a variety of very interesting papers describing more sophisticated perceptual models and coding schemes, for example [SZT98, PZ98].

In our taxonomy, such systems are considered Category 2 systems. The watermark inserter is characterized by the fact that the watermark signal to be embedded is now modified, prior to insertion, based on a function of the content itself. However, such modification is performed primarily for fidelity considerations, and not specifically with the intention of improved detector performance.

[1] The use of the original content at the detector is a case of detection with side information. However, in the paper, we are primarily concerned with embedding with side information.

At this point, it is useful to discuss a simple example. Consider a content vector $\mathbf{C_0}$ and a watermark vector $\mathbf{w}$, where the magnitude of $\mathbf{w}$ is considerably smaller than that of $\mathbf{C_0}$. In a Category 1 watermarking system, the embedding process forms a watermarked content vector $\mathbf{C_w}$ as the simple sum of $\mathbf{C_0}$ and $\mathbf{w}$. Category 2 watermarking takes advantage of a content-dependent fidelity model to balance a perceptual distortion constraint with effective embedding techniques. While the particular form of fidelity model is not important to this paper, assume for illustration that the output from such a model indicates that each element of the content can be changed by no more than $\pm y$ units. Thus, a very simple Category 2 embedder might simply scale the watermark vector, $\mathbf{w}$ by a scalar value, α such that the magnitude of the largest element of $\alpha\mathbf{w}$ equals y. This scaled vector is then added to the content to obtain the watermarked content, $\mathbf{C_w}$. A variety of other embedders are possible. For example, the embedder might choose a value for α such that the average magnitude of the elements of the watermark is scaled to y and elements of the watermark that exceed this magnitude are then truncated to y. Many other possibilities are imaginable, which begs the question of whether there is an optimium embedding strategy, which is the focus of Category 3 watermarking.

This paper describes a third category of watermarking, Category 3, in which side information in the form of knowledge of the actual content at the embedder is exploited for improved detector performance. The result is that novel watermark embedding and detection algorithms can be developed which, we believe, are (or at least have the potential to be) superior to the prior art. The basic idea is to choose a watermarked content vector $\mathbf{C_w}$ from within an acceptable distortion region in order to maximize probability of detection. This maximization may have further constraints, such as maximizing detectability after MPEG compression, but the rest of this paper does not develop this further.

To perform this optimization, we must know the form of the detector. Matched filtering is a detection technique frequently employed in many communications applications, particularly spread spectrum systems, which often form a model for watermarking systems. When matched filter detection is applied, the gain to be had in one-shot detection performance from side information at the embedder is largely dependent on the form of fidelity and distortion models, and can be limited. For instance, under a content-independent power constraint on the strength of the embedding vector and an assumption of no distortion encountered subsequent to the embedding process, an optimal embedding scheme is formed by adding the watermark vector (against which the matched filter correlates) scaled to maximum-allowable power; clearly, this scheme takes no advantage of the knowledge of content realization at the embedder[2].

While a matched filtering approach to detection would seem natural for a watermarking system, recent designs have favored another detection technique, whereby the normalized coefficient between a given content and watermark vector pair is evaluated and compared to a threshold[3]. This form of

[2] However, if detection is performed over multiple blocks of content vectors, rather than as one-shot detection over a single content vector, it might be possible to exploit side-information at the embedder to make a more sophisticated system. While detector complexity becomes a serious issue in such a design, this could indeed form the focus of future investigations.

[3] This threshold can be determined based on the desired level of false positives and false negatives and is well known

detector is favored primarily because of its robustness to many of the distortive effects encountered in a watermarking environment, which often exhibit a strong correlation with the embedded content vector. For instance, a simple brightness scaling of a video frame has the potential to significantly degrade matched filter detector performance, while the preserving of linear relationships would tend to render a normalized correlation detector more robust. Section 5 forms a detailed discussion of watermark distortion issues, while detector design is addressed in Section 6, including a detailed analysis of the normalized correlation detector. When normalized correlation detection is employed, the potential exists for significant performance improvement through the use of side information at the embedder, even in one-shot detection.

Returning to the example outlined above, assume that the detector operates by evaluating the normalized correlation between a given content vector $\mathbf{C_w}$ and the watermark $\mathbf{w}$ and comparing to a threshold. A Category 3 embedder would then pick a content vector $\mathbf{C_w}$ from within the allowable distortion region in order to maximize such correlation, perhaps subject to post-embedding distortion considerations. In so doing, knowledge of the original content vector is used explicitly by the embedder to produce a watermark with a greater probability of detection.

The remainder of this paper explores how a Category 3 watermarking system can be developed.

3 Framework and Definitions

In this section, we model a generic watermarking procedure and describe the terms we will use to analyze it. The framework described here is sufficiently detailed for analyzing watermarking algorithms that employ side information at the embedder, but it does not specify many details of the algorithm, such as perceptual models, data transforms, registration methods, etc. In the framework, these details are abstracted in the form of a small set of functions that need not be specifically defined here. A wide variety of different algorithms can be created by defining the functions in different ways. The analysis presented in the present article will apply to all of these.

The terms, functions and variables are summarized in Table 1, and some of them are used as annotations to Figures 2 and 3. We describe these terms below, and then give a pair of examples of how specific functions might be defined to correspond to components of watermarking technology found in the literature.

3.1 Terms, functions, and variables

We follow [Pfi96] in referring to the media content that is to be watermarked as the "cover data". This is described by a vector, which we usually denote by $\mathbf{B}$ or $\mathbf{C}$, in a "media space" of M dimensions. For example, in a system designed to watermark images that are represented as p by q arrays of pixels, the media space is a $M = p \times q$ dimensional pixel space.[4]

in detection theory, especially in a Neyman-Pearson framework.

[4] Some types of cover data are not naturally represented by vectors of predetermined numbers of dimensions. For example, different audio clips might be different lengths, yielding different numbers of samples. In such cases, we make the simplifying assumption that the watermark detector will examine the cover data for a fixed amount of time, t, before making a decision about the presence or absence of a watermark. If the sampling rate per unit time is S,

"cover data"	a piece of media that is to be watermarked
"media space"	a space in which each piece of cover data can be represented as a vector
"watermark space"	a space in which each watermark can be represented as a vector
M	the number of dimensions in media space
K	the number of dimensions in watermark space
$\mathbf{B}$, $\mathbf{C}$	vectors in media space
$\mathbf{B_0}$, $\mathbf{C_0}$	unwatermarked cover data
$\mathbf{B_w}$, $\mathbf{C_w}$	watermarked cover data
$\mathbf{B_u}$, $\mathbf{C_u}$	cover data that might or might not contain a watermark
$\mathbf{B}'$, $\mathbf{C}'$	possibly distorted cover data
$\mathbf{N}$	distortion vector in media space ($\mathbf{C}' - \mathbf{C}$)
$D(\mathbf{B}, \mathbf{C})$	perceptual distance in media space
$\mathbf{w}$, $\mathbf{r}$, $\mathbf{s}$	vectors in watermark space ($\mathbf{w}$ is usually used for a vector that specifies a watermark)
$\mathbf{r_0}$, $\mathbf{s_0}$	signals extracted from unwatermarked cover data
$\mathbf{r_w}$, $\mathbf{s_w}$	signals extracted from watermarked cover data
$\mathbf{r_u}$, $\mathbf{s_u}$	signals extracted from cover data that might or might not contain a watermark
$\mathbf{r}'$, $\mathbf{s}'$	signals extracted from possibly distorted cover data
$\mathbf{n}$	distortion vector in watermark space ($\mathbf{r}' - \mathbf{r}$)
$d(\mathbf{r}, \mathbf{s})$	perceptual distance in watermark space
$\mathcal{S}(\mathbf{r})$	region of watermark space with acceptable perceptual difference from $\mathbf{r}$
$\mathcal{R}(\mathbf{w})$	region of watermark space that will be detected as containing watermark $\mathbf{w}$
$X(\mathbf{C})$	extraction function, which projects from media space into watermark space
$Y(\mathbf{r}, \mathbf{C})$	inverse extraction function, defined such that, if $\mathbf{B} = Y(\mathbf{r}, \mathbf{C})$, then $\mathbf{r} = X(\mathbf{B})$, and $D(\mathbf{B}, \mathbf{C})$ is small
$\mathbf{m}$	watermark message
$E(\mathbf{m})$	encoded and modulated watermark message (a vector in watermark space)
$f(\mathbf{r}, \mathbf{w})$	"mixing" function, which yields a vector in the intersection of $\mathcal{S}(\mathbf{r})$ and $\mathcal{R}(\mathbf{w})$

Table 1: Definition of terms

Where it is necessary to distinguish between cover data that contains a given watermark and cover data that does not, we use subscripts. The vector $\mathbf{C_0}$ denotes cover data that has not been watermarked while $\mathbf{C_w}$ denotes cover data that contains a watermark $\mathbf{w}$. Cover data that might or might not contain a watermark is denoted by $\mathbf{C_u}$, where the subscipt $\mathbf{u}$ denotes that presence or absence of a watermark is unknown.

Cover data, either containing or not containing a watermark, may experience a variety of distortions due to compression, signal processing, etc. A prime superscipt is used to denote cover data that has possibly been distorted, for instance $\mathbf{C_w}'$ is a possibly distorted version of $\mathbf{C_w}$. A "distortion vector" $\mathbf{N}$ is an M dimensional additive distortion applied to a given piece of cover data, *i.e.* $\mathbf{N} = \mathbf{C_w}' - \mathbf{C_w}$.

The real-valued function $D(\mathbf{B}, \mathbf{C})$ represents a perceptual distance function which yields a numerical measure of the perceptual distance between content vectors $\mathbf{B}$ and $\mathbf{C}$. This is typically used to measure the perceptual distance between watermarked and unwatermarked versions of a piece of cover data, $D(\mathbf{C_w}, \mathbf{C_0})$.

A "watermark message" is an arbitrary set of bits that will be encoded in the watermark. The watermark embedding process involves an encoding and modulation function, $\mathbf{w} = E(\mathbf{m})$, which maps bit sequences into vectors in a K-dimensional "watermark space". When we refer to a "watermark", "watermark vector", or "watermark signal", we are referring to a vector in this space. The dimensionality of watermark space is assumed to be less than or equal to the dimensionality of media space.

The number of possible watermarks that a system might embed is determined by the number of bits in its watermark messages. During watermark detection, many systems can decode the bits of the message, in effect determining which of the many possible messages is most likely to have been embedded. Other systems can test for the presence or absence of only one watermark at a time. These more limited systems can be made to identify the most likely of many messages by exhaustively testing for all the possible watermarks. While this approach is less efficient than the approaches taken in typical, multi-bit algorithms, it is conceptually the same. For this reason, we will consider only systems which perform single-watermark detection.

The media and watermark spaces need not be of the same dimension. For example, we may choose to watermark a frequency band that is significantly smaller than the spectrum of the the media. Media space and watermark space are related by a pair of functions, which we refer to as the "extraction function", $X(\cdot)$, and the "inverse extraction function", $Y(\cdot)$.

The extraction function $X(\cdot)$ maps cover data into vectors in watermark space. The resulting vectors are referred to as "extracted signals", and are denoted by $\mathbf{r}$ or $\mathbf{s}$. As such, we have that $\mathbf{r} = X(\mathbf{C})$. We use the same subscript and prime conventions for extracted signals as we use for cover data, thus $\mathbf{r_0} = X(\mathbf{C_0})$, $\mathbf{r_w}' = X(\mathbf{C_w}')$, etc. Distortion vectors can also be expressed in watermark space: $\mathbf{n} = \mathbf{r}' - \mathbf{r}$.

The inverse extraction function, $\mathbf{B} = Y(\mathbf{r}, \mathbf{C})$, maps vectors from watermark space into media

then the portion of the cover data examined by the detector can be represented as a vector in $M = t \times S$ dimensional sample space.

space. The first argument is a vector in watermark space that usually corresponds to watermarked cover data. The second argument is necessary because, if watermark space has fewer dimensions than media space, the mapping from watermark space to media space is one-to-many, so the original, unwatermarked cover data is required to disambiguate the function. Thus, the inverse extraction function finds a unique piece of media, $\mathbf{B}$, such that $\mathbf{r} = X(\mathbf{B})$ and $D(\mathbf{B}, \mathbf{C})$ is small. An optimal version of $Y(\cdot)$ would find the value of $\mathbf{B}$ that minimizes $D(\mathbf{B}, \mathbf{C})$. However, a system need not be optimal in order to fit into the present framework. Note that the inverse extraction procedure does not represent the entire embedding process. A further key element, which we refer to as "mixing", is also required, as discussed shortly.

The function $d(\cdot)$ estimates perceptual distances in watermark space. This must correspond to the distance function in media space. That is, the distance between two vectors in watermark space should be equal to the distance between the two media space vectors that result from applying the inverse extraction function. Since the inverse extraction function requires an instance of cover data as a reference, the distance function in watermark space can be strictly defined only with respect to a given piece of cover data, as $d(\mathbf{r}, \mathbf{s}, \mathbf{C}) = D(Y(\mathbf{r}, \mathbf{C}), Y(\mathbf{s}, \mathbf{C}))$.

When $d(\cdot)$ is used, the reference cover data will typically be the original, unwatermarked cover data, $\mathbf{s}$ will be the watermark signal extracted from it, and $\mathbf{r}$ will be the signal in watermark space corresponding to the watermarked cover data. Thus, we will most often be interested in the distance

$$d(\mathbf{s_w}, \mathbf{s_0}, \mathbf{C_0}) = D(Y(\mathbf{s_w}, \mathbf{C_0}), \mathbf{C_0})$$

where $\mathbf{s_0} = X(\mathbf{C_0})$ and $Y(\mathbf{s_0}, \mathbf{C_0}) = \mathbf{C_0}$.

For many possible extraction and inverse extraction functions, the value of $d(\cdot)$ is independent of the reference cover data. Consider, for example, an extraction function that projects an M dimensional media vector into a K dimensional extracted signal by simply ignoring $M-K$ dimensions of the input vector. This is the case when we choose to watermark only a low frequency subset of the image spectrum, for example. So, if $\mathbf{r} = X(\mathbf{C})$, then $r_i = C_i$ for all $i \leq K$, where $K < M$. A natural inverse extraction function, $\mathbf{B} = Y(\mathbf{r}, \mathbf{C})$, would fill in the $M - K$ dimensions required to go from watermark space to media space by copying them from $\mathbf{C}$. So $B_i = r_i$ for all $i \leq K$, and $B_i = C_i$ for all $K < i \leq M$. If $D(\cdot)$ is defined to be Euclidean distance in media space, then it is easy to see that $d(\cdot)$ will be independent of its reference cover data. In fact, $d(\cdot)$ will simply be Euclidean distance in watermark space.

Because the reference cover data used in $d(\cdot)$ is often irrelevant, we will ignore it in the rest of this paper. Thus, we assume that we can have a function in watermark space, $d(\mathbf{r}, \mathbf{s})$, which corresponds exactly with $D(\mathbf{B}, \mathbf{C})$ in media space. If the value of $d(\cdot)$ is not independent of the reference cover data, then we assume that the reference cover data is always the data being watermarked.

Next, we define two regions in watermark space. The region $\mathcal{S}(\mathbf{r_0})$ is the set of signals $\mathbf{s}$ for which $d(\mathbf{s}, \mathbf{r_0})$ is less than a given fidelity threshold. This represents the set of signals that correspond to cover data that is perceptually similar to the data from which $\mathbf{r_0}$ was extracted. The region $\mathcal{R}(\mathbf{w})$ is the set of signals that are sufficiently similar to the watermark $\mathbf{w}$ so as to be considered as watermarked. To convey the presence of a watermark in cover data $\mathbf{C_0}$, the embedder chooses a signal $\mathbf{r_w}$ from within the intersection of $\mathcal{S}(\mathbf{r_0})$ and $\mathcal{R}(\mathbf{w})$, and maps this signal into media space. To determine if a given piece of cover data $\mathbf{C_u}$ is watermarked, the watermark detector tests whether .

During watermark detection, the extraction function is applied to the cover data being tested, and the resulting vector is compared with the watermark being tested for. Thus, the steps for watermark detection are:

1. $r_u = X(C_u)$
2. if $r_u \in \mathcal{R}(\mathbf{w})$, then C_u contains watermark w

During watermark embedding, knowledge of the content vector is exploited by applying the extraction function and subsequently modifying the extracted signal so as to maximize the probability of a detection at the receiver. The modification is performed by the "mixing function", $f(\cdot)$. This function is given an extracted signal and a watermark vector. $\mathbf{r_w} = f(\mathbf{r_0}, \mathbf{w})$ yields a signal in watermark space that is within the detection region for the given watermark vector, $\mathcal{R}(\mathbf{w})$, and that is perceptually similar to the extracted signal, so $d(\mathbf{r_0}, \mathbf{r_w})$ is small. We call this process "mixing" because it "mixes" the desired watermark, $\mathbf{w}$, with the extracted vector, $\mathbf{r_o}$. We apply the mixing function in watermark space because, if the extraction and inverse extraction functions are suitably defined, the task can be substantially easier than it is in media space, as discussed in Section 5.

Watermark embedding is completed by applying the inverse extraction function to convert the modified signal from watermark space into media space. Thus, the steps for watermark embedding are:

1. $\mathbf{r_0} = X(\mathbf{C_0})$
2. $\mathbf{r_w} = f(\mathbf{r_0}, \mathbf{w})$
3. $\mathbf{C_w} = Y(\mathbf{r_w}, \mathbf{C_0})$.

3.2 Examples

At this point, a couple of examples are supplied in order to illustrate how the functions $X(\cdot)$, $Y(\cdot)$, $D(\cdot)$, $d(\cdot)$, and $f(\cdot)$, and the regions of watermark space $\mathcal{S}(\cdot)$ and $\mathcal{R}(\cdot)$ might be defined for some image watermarking algorithms.

First, we begin with a simple system in which each possible watermark message is encoded as an independent pseudorandom pattern equal in size to that of the image. A watermark is embedded into an image by simply adding the pattern to it, attenuated so-as not to cause too much fidelity impact. The detector works by matched filtering, that is, it computes the inner product between the image being tested and the watermark pattern, and then compares the result against a threshold, T_c.

In such a system, the media space and the watermark space are the same. The watermark extraction function is simply an identity mapping, i.e. the extracted watermark $\mathbf{r}$ is simply the received cover data or content $\mathbf{C}$, and

$$\mathbf{r} = \mathbf{C} = X(\mathbf{C}) \; .$$

The inverse extraction function is similarly an identity mapping, in which the reference cover-data, $\mathbf{B}$, is ignored:

$$Y(\mathbf{r}, \mathbf{B}) = \mathbf{r} = \mathbf{C} .$$

A natural perceptual distance metric in such a system is a weighted Euclidean distance in pixel space, in which the weight of each pixel is determined by a local fidelity function. If the weight for pixel C_i is G_i, then we have

$$D(\mathbf{B}, \mathbf{C}) \quad = \quad \sqrt{\sum_i ((C_i - B_i)G_i)^2}$$

The distance function in watermark space, $d(\cdot)$, is the same as $D(\cdot)$. The region of acceptable fidelity, $\mathcal{S}(\mathbf{r})$, is just a sphere centered around $\mathbf{r}$.

The detection region is defined by the correlation test described above. Thus we have

$$\mathcal{R}(\mathbf{w}) \quad = \quad \{\mathbf{s} : \mathbf{s}^T\mathbf{w} > T_c\}$$

Finally, the mixing function, $f(\mathbf{r}, \mathbf{w})$, finds a scaling value, k, such that $k\sqrt{\sum_i (w_i G_i)^2}$ is within the radius of the sphere $\mathcal{S}(\mathbf{r})$, and adds the watermark, scaled by this factor, to the image. So

$$f(\mathbf{r}, \mathbf{w}) \quad = \quad \mathbf{r} + k\mathbf{w}$$

A more interesting example would be a system similar to the one above, but in which only the low frequencies of the image are watermarked. Here, a watermark is embedded by applying a frequency transform to the image, adding the frequency-domain watermark to the low-frequency coefficients, and then taking the inverse frequency transform of the result. A watermark is detected by applying the frequency transform to the image, and then computing the inner product between it's low-frequencies and the frequency-domain watermark. If the result surpasses a given threshold, T_c, then the watermark is determined to be present.

When this system is fit into our framework, the extraction function, $X(\cdot)$, consists of the frequency transform and the removal of high-frequency coefficients. So

$$X(\mathbf{C}) \quad = \quad \mathit{Low}(\mathcal{F}(\mathbf{C}))$$

where $\mathcal{F}(\cdot)$ is a frequency transform such as the Fourier or discrete cosine ransform, and $\mathit{Low}(\cdot)$ truncates an M-dimensional frequency-domain image to its K lowest frequencies. Since K is less than M, $X(\cdot)$ maps a higher-dimensional media space down to a lower-dimensional watermark space.

Since $X(\cdot)$ performs a many-to-one mapping, the second argument of $Y(\cdot)$ will be required for disambiguation. The inverse extraction process, $Y(\cdot)$, can be defined to append the high frequencies from its second argument to the low frequencies present in its first argument, forming a complete, frequency-domain image, and then to apply the inverse frequency transform, $\mathcal{F}^{-\infty}(\cdot)$, to obtain a spatial-domain image. So

$$Y(\mathbf{r}, \mathbf{B}) \quad = \quad \mathcal{F}^{-1}(\mathit{Append}(\mathbf{r}, \mathit{High}(\mathcal{F}(\mathbf{B}))))$$

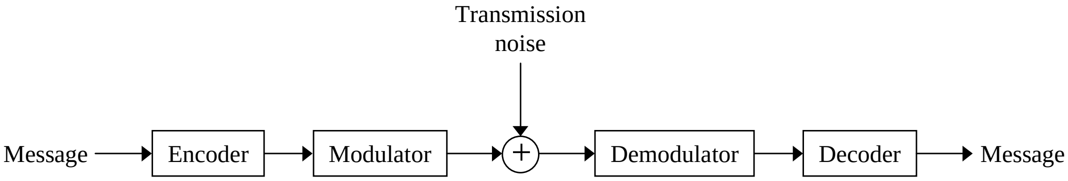


Figure 1: Standard model of a communications channel.

where $High(\cdot)$ gives the $M-K$ high frequencies of its M-dimensional, frequency-domain argument, and $Append(\cdot)$ appends an $M-K$-dimensional vector (second argument) onto a K-dimensional vector (first argument) to produce a single, M-dimensional vector.

The perceptual distance metric, $D(\cdot)$, can be defined in a manner similar to that in the first example, but it will be most convenient to define it in the frequency domain. So

$$D(\mathbf{B},\mathbf{C}) \quad = \quad \sqrt{\sum_{i}^{M}((\mathcal{F}_i(\mathbf{C})-\mathcal{F}_i(\mathbf{B}))G_i)^2}$$

where $\mathcal{F}_\rangle(\cdot)$ indicates the i'th term of the frequency transform of its argument, and G_i is a perceptual weight for frequency i.

In watermark space, the perceptual distance metric, $d(\cdot)$, is simple to define. If we use $Y(\cdot)$ to map any two watermark vectors into media space using the same reference content, B, then the two resulting media vectors will differ only in the low frequencies that are used in watermark space. Thus, only the first K terms of the summation for the difference between them will be non-zero. So

$$\begin{aligned} d(\mathbf{r},\mathbf{s}) &= D(Y(\mathbf{r},\mathbf{B}),Y(\mathbf{s},\mathbf{B})) \\ &= \sqrt{\sum_{i}^{K}((r_i-s_i)G_i)^2}. \end{aligned}$$

Finally, the region of acceptable fidelity, $\mathcal{S}(\cdot)$, the detection region, $\mathcal{R}(\cdot)$, and the mixing function, $f(\cdot)$, can all be defined exactly as in the first example, with the only difference being that they are here defined for K-dimensional vectors, rather than M-dimensional vectors.

4 Watermarking as communications

Figure 1 illustrates the basic elements of a classical communications system. The message to be transmitted is first encoded. This step typically takes a binary input stream and translates it into a binary output stream, usually for error correction and/or frequency spreading purposes. The encoded message is then used to modulate a carrier signal in any of a variety of ways, e.g. amplitude, frequency, phase, spread-spectrum, etc. The modulated carrier signal is transmitted via a transmission channel, where it encounters additive noise. The receiver demodulates the noisy signal to a (possibly corrupted) encoded message. Finally, this message is decoded to produce the received message.

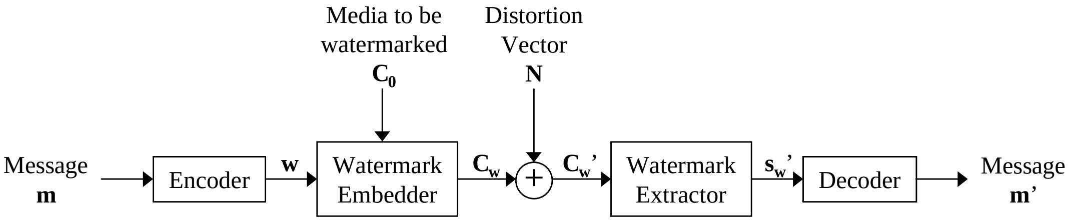


Figure 2: Watermarking as communications.

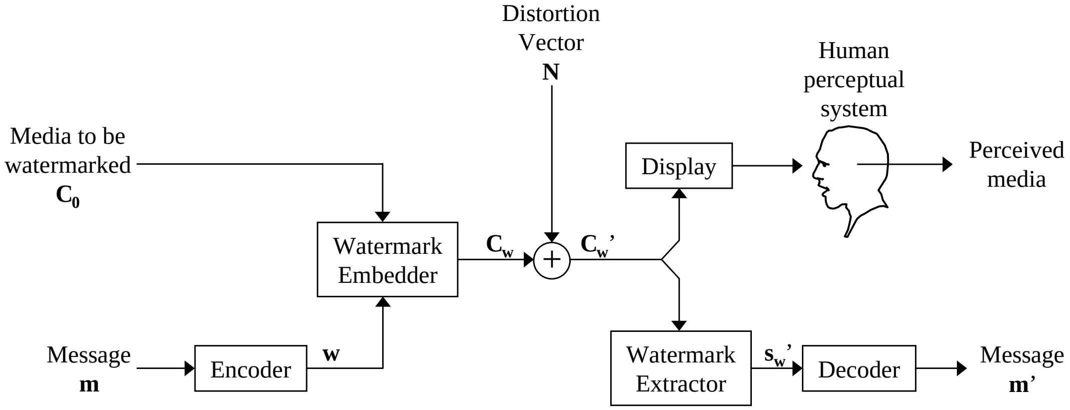


Figure 3: Complete model of watermarking

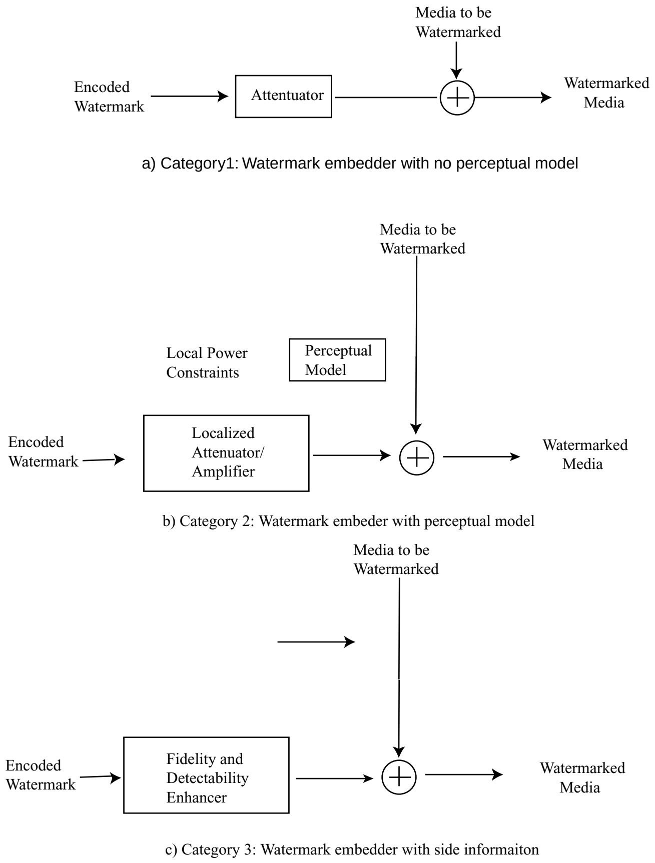
In Figure 2, we show one way in which watermarking can be mapped into this framework. Here, the modulation step is replaced by the step of embedding the encoded message[5] into some media content $\mathbf{C}$ and the demodulation step is replaced by the step of extracting the (possibly corrupted) watermark signal from the received signal[6], $\mathbf{C_w}'$.

In watermarking, the post-embedding noise in the transmission channel, $\mathbf{N}$, results from various types of processing that the watermarked media goes through before the watermark is received, e.g. compression and decompression, broadcast over analog channels, image or audio enhancements, etc. It might also result from malicious processing by pirates intent on removing the watermark.

While Figure 2 illustrates a strong relationship between classical communications and watermarking, it is incomplete; it does not illustrate the importance of maintaining the fidelity of the watermarked media. A more complete image of a watermarking system is shown in Figure 3. Here, we have added a second “receiver” in the form of human sensory organs, which should receive a “message” that is essentially the same as the carrier media content. While this represents a deviation from classical communications, in which the carrier signal’s sole function is to carry the encoded message, the resulting fidelity constraint is analagous to a constraint on signal power in a communication channel, albeit with a different metric and motivation.

[5] Note that a binary encoding of the watermark, as would be given by a typical encoder in a classical communications system, is not necessarily desirable. See [CKLS97].

[6] In some watermarking systems, the receiver is provided information about the unwatermarked content before it extracts the watermark. See [CM97] for further discussion of this issue. This is also a form of side information, but now at the receiver.



a) Category1: Watermark embedder with no perceptual model

b) Category 2: Watermark embeder with perceptual model

c) Category 3: Watermark embedder with side informaiton

Figure 4: Three categories to watermark embedding.

Because of the importance of preserving the fidelity of the watermarked media, most designers of watermarks have not viewed the media as a carrier signal to be modulated, but rather as noise that cannot be removed from the system. The systems that have resulted from this view can be divided into two categories, as discussed earlier and illustrated in Figures 4a and 4b. These figures show different schemes for implementing the watermark embedding step of Figures 2 and 3.

In the first category of embedders (Figure 4a), no advantage is taken of the fact that the media content is completely known at the time of embedding. The encoded watermark signal is simply attenuated according to a fixed, global power constraint, and added to the media. To maintain fidelity, the power constraint is kept very low. Spread spectrum techniques are often employed to make the signal detectable in the resulting low signal-to-noise ratio channel[7]. Examples of Category 1 watermarking methods can be found in [Tur89, Car95, TNM90, MT94, HW96, Rho95, BGML96, HMW88, SC96]. Note that this approach to watermarking makes no distinction between the media content, $\mathbf{C_0}$, and the attack noise, $\mathbf{N}$.

[7] Spread spectrum communications was originally developed as a military communications system that would be resistant to enemy jamming, though it is increasingly being used in commercial applications. The basic idea is to take a narrowband signal and spread its energy over a much larger bandwidth. A detailed review of spread spectrum technology can be found in [PSM82].

Cox *et al* [CKLS97] first suggested that the watermark signal should be modified based on the perceptual properties of the content. This led to a second category of watermarking algorithms, depicted in Figure 4b, which make use of a model of human perception. In such algorithms, the watermark signal is locally amplified or attenuated according to the local sensitivity of the model, where locality may be interpreted in either the spatial or frequency domain. The resulting, modified signal is then added to the media content as in a Category 1 embedder. A perceptual model is often used to increase the power of the signal while maintaining fidelity by amplifying the signal wherever the model's sensitivity is low. Alternatively, to improve robustness, especially to compression, the perceptual model may be used to identify perceptually significant regions of the content that must be preserved by a compression algorithm. A watermark can be inserted in these perceptually significant regions using, for example, spread spectrum techniques [CKLS97, RDB96, SZT98, PRH$^+$94, PZ98].

Regardless of whether the goal is to increase fidelity or robustness, category 2 systems employ only the perceptual properties of the media during embedding, i.e. the effect that the watermarked media will have on the human perceptual system. The effect it will have on the watermark detector is not explicitly modeled, and for public watermarking, the content is modeled as noise at the detector.

In the present paper, we present a logical generalization of the perceptually-based methods of Figure 4b. This third category of watermarking algorithms is illustrated in Figure 4c. Here, we have simply removed the stipulation that the media content be used only for local amplification and attenuation of the watermark signal. The result is an example of a communication channel with side information at the transmitter, such as that studied in [Sha58]. Rather than modelling the media content as *unknown* channel noise, knowledge of the content is side information that can be used by the embedder to control both fidelity (by using a perceptual model) and detectability (by using a model of the watermark detector). The embedder can choose any $\mathbf{C_w}$ that satisfies these two constraints according to a desired tradeoff.

Similarly, when the original content, $\mathbf{C_0}$, is present at the detector, such an arrangement can again be considered communication with side information. However, the remaining sections of this paper only examine communication with side information at the embedder and assume that no side information is available at the detector. This is common for "public" watermark systems in which detectors are widespread and not controlled by the content owner.

The systems under consideration in the present article have the following features:

- The watermark embedder takes two inputs: the watermark signal, $\mathbf{w}$, and the media content to be watermarked, $\mathbf{C_0}$. It has one output: the watermarked media, $\mathbf{C_w}$.

- The embedder has complete knowledge of the functionality of the watermark detector, and some level of knowledge of the human perceptual system.

- The embedder has no knowledge of the attack noise that will be applied to the watermarked media before it is input to the watermark detector. However, this is an interesting generalization of the approach which we believe should be studied.

- The watermark detector takes two inputs: the received, possibly watermarked media, $\mathbf{C_u}'$, and the watermark signal to test for, $\mathbf{w}$. It outputs a yes or no answer to the question of whether the given watermark is present in the media.

- If the media input to the detector contains no watermark, then it has not been manipulated by any part of the watermarking system. This means that the detector receives just the media plus the attack noise, $\mathbf{C_0}' = \mathbf{C_0} + \mathbf{N}$.

In order to develop a Category 3 embedder, we must first decide upon a suitable detector. To do so requires knowledge of the statistical characteristics of the content and noise that can be expected. Section 5 provides a discussion of the these charactersistics and Section 6 uses this information to recommend the form of the detector. Finally, Section 7 uses this information to demonstrate how a Category 3 embedder may be designed.

5 Characteristics of cover data and distortion vectors

In the first two categories of watermarking systems discussed above, little or no distinction is made between the cover data and the distortion vectors. They are both simply considered as noise. However, in the systems under consideration here, this distinction is clear, and it is profitable to consider the different natures of and relationship between these two types of "noise".

We are primarily concerned with the characteristics of these vectors in watermark space. That is, we are interested in the signals extracted from the cover data and the distortion vectors as projected into watermark space.

5.1 Cover data

There are two characteristics of the signals extracted from cover data to consider. The first is the distribution of vectors that can be expected from unwatermarked data that is input to a watermark embedder or detector. The second is the amount by which the embedder may change a vector before causing an unacceptable degradation in fidelity.

The distribution of vectors extracted from unwatermarked data is highly dependent on the extraction function, $X(\cdot)$. For example, if we consider the simplest image watermarking method described in Section 3, where $X(\mathbf{C}) = \mathbf{C}$, the distribution of extracted vectors is the same as the distribution of images. The elements of these vectors are highly correlated, since the pixels in natural images are usually correlated with their neighbors. But sparse sampling, averaging of disparate values, whitening filters, and some image transforms can produce spaces in which the dimensions of naturally occuring images are far less correlated. Furthermore, many of these techniques produce distributions for the individual elements that are zero-mean Gaussians (see, for example, the analysis in [HPGRN98]). All of these techniques can be used in watermark extraction functions.

We will assume that the probability density function of signals extracted from unwatermarked, undistorted cover data is zero-mean, independent Gaussian. The analysis that results applies only to watermarking systems in which the extraction functions are designed to approximately yield such a distribution. It will not apply directly to systems like the simple X(**C**)=**C** system.

The maximum allowable change in an extracted signal before the fidelity of the watermarked data becomes unacceptable is also dependent on the extraction function. Generally, the maximum change will be very small. But, when the dimensionality of the watermark space is significantly lower than that of the media space, the maximum change, relative to the standard deviation of the values in the original extracted signal, may increase.

For example, imagine that our perceptual distance metric is simply the largest distance along any of the N dimensions of media space: $D(\mathbf{C}, \mathbf{B}) = \max_i |C_i - B_i|$. Imagine, further, that if the distance between the watermarked and unwatermarked versions of a piece of cover data is more than $D(\mathbf{C_0}, \mathbf{C_w}) = 1$, then the fidelity of the watermarked version will be unacceptable. So the embedder can change each dimension in media space by up to a value of 1. If the length of a typical cover data vector is significantly larger than 1, then a watermarking system that works directly in media space will have little latitude for embedding watermarks. But, if our extraction function averages groups of k values together, then the situation improves. The expected standard deviations of the signals extracted from unwatermarked cover data will be roughly $1/\sqrt{k}$ times the expected standard deviations of the cover data vectors, while the maximum change in any dimension remains 1. It is thus conceivable that we could effect such a large change in this smaller space that the original media vector is completely cancelled out, while still maintaining fidelity.

In Section 7, when we consider the best strategy for embedding watermarks that are to be detected by thresholding a normalized correlation, we will make no assumptions about the relative amount that we can change the media vector. But, in Section 6, when we consider the best strategy for detecting watermarks, we will assume that the media vector can be completely cancelled out. Thus, the analysis of Section 6 can only be applied directly to systems that work in spaces like the one in which disparate values are averaged together, but applies in principle to other systems as well.

5.2 Distortion Vectors

Unlike the cover data, distortion vectors can never be changed by the watermarking system. Thus, we are interested here only in the distribution of distortion vectors. We argue that an appropriate distortion model for watermarking applications includes a significant correlation between distortion vectors and content vectors to which they are applied. This property will motivate a detection strategy for which side information at the transmitter proves to be particularly exploitable.

The distribution of distortion vectors is often modelled as uncorrelated Gaussian noise. Such a model is convenient, and it is accurate for distortions which are truly random, for example transmission over a noisy analog channel. However, commonly, digital media can fall subject to a randomly selected collection of deterministic, digital processes. For a simple example, consider the process of decreasing the brightness of an image by 10%. Here, the amount that is added to a given pixel value is determined by a simple function of that pixel's original value, not well-modelled by independent

Gaussian noise. More complex examples would consist of any of a variety of compression and decompression algorithms. For most algorithms, the degradation applied to the input image is entirely dependent on the image and the compression rate. The choice of distortions and their parameters may be random, but the result of each distortion is a deterministic function of the media.

We therefore believe that a more realistic model of distortion vectors would specify a correlation between the noise and the media[8]. We model such correlation by defining the distortion vector, **n**, to be a Gaussian random vector with zero mean and covariance matrix

$$\Lambda_{\mathbf{r}} \quad \stackrel{\triangle}{=} \quad \sigma_n^2[\rho\,\mathbf{r}\mathbf{r}^T + (1-\rho)I] \tag{1}$$

where **r** is a vector in watermark space, σ_n^2 represents the overall strength of the distortion and the parameter ρ, $0 \leq \rho \leq 1$, determines the extent of the noted correlation between **n** and **r**.

To examine the accuracy of this model, and to get an idea of suitable values of ρ, we subjected 1800 images to a variety of distortions, and estimated ρ for each type of distortion, measured in RGB pixel space. The images were 720 by 486 pixels in size, and each pixel was represented with three bytes, one each for red, green, and blue. The distortions were:

1. "BLUR 3" filtered the image with a 3x3 box filter.

2. "BLUR 7" filtered the image with a 7x7 box filter.

3. "SHARPEN" filtered the image with the simple, 3x3 sharpening filter shown below:

$$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 9 & -1 \\ -1 & -1 & -1 \end{bmatrix}.$$

4. "PIXELATE 3" divided the image into blocks of 3x3 pixels, and replaced all the pixels in each block by the average of their colors. This is equivalent to shrinking the image down to 1/3 the resolution in each dimension, and then pixel-replicating it back up to the original size.

5. "PIXELATE 7" performed the same process as PIXELATE 3, but with 7x7 blocks.

6. "SNOW 20" added normally distributed, random values to each of the R, G, and B values in the image. The values were drawn from a zero-mean distribution with standard deviation 20. The results were clipped to the range of 0 through 255.

7. "DITHER" reduced the image to 8 colors (black, red, green, yellow, blue, magenta, cyan, and white) by means of three independant Bayer ordered dithers [Bay73].

8. "QUANTIZE 5" quantized each R, G, and B value down to 5 bits by zeroing out the three low-order bits.

[8] Note that this form of "correlated noise" is not what is usually meant by "correlated noise". Usually, this refers to auto-correlation which can be overcome by means of whitening filters. However, we are referring to the correlation between the noise *and the transmitted signal*. In this context, the applicability of whitening filters is unclear.

	ρ			
Distortion	Mean	Median	Min	Max
BLUR 3	.196	.192	0	.603
BLUR 7	.218	.224	0	.616
SHARPEN	-.139	-.127	-.831	.483
PIXELATE 3	.168	.160	0	.576
PIXELATE 7	.217	.213	0	.614
SNOW 20	.031	.025	-.094	.628
DITHER	-.006	-.006	-.248	.124
QUANTIZE 5	.763	.763	0	.991
GAMMA 1.18	.898	.931	0	.973

Table 2: Statistics of Content-Distortion Correlation ρ.

9. "GAMMA 1.18" raised each R, G, and B value to the power of 1.18 and scaled back to a range of 0 to 255. This is a standard simulation of what happens in video monitors and other non-linear video equipment (1.18 is a typical number for monitors).

The results are summarized in Table 2. It is clear that ρ cannot be expected to be zero. The significance of this is discussed in Section 6, where, for non-zero ρ, we show that normalized correlation may be preferable to matched filtering.

Note that this model makes no attempt to deal with the issue of malicious attacks, that is distortions performed by would-be pirates with the sole intention of removing the watermark. Distortions of this type, especially those designed using detailed knowledge of the algorithm (for examples, see [CL97, CL98, Kal98, KLvD98, LvD98]), are extremely difficult to analyze, and are not be covered in the present article. However, several malicious attacks are actually pathalogical examples of processes that might otherwise occur with normal processing (for example, see [PAK98]). The analysis based on our model of distortion vectors should have some application to these types of malicious attacks.

The model of distortion vectors given here is used in Section 6 to decide on the best detection region to use when the signal extracted from unwatermarked cover data can be completely cancelled out during watermark embedding. It is not used in Section 7 in order to permit the derivation of sufficiently detailed results.

6 Watermark Detection

We are now interested in developing an effective detection strategy for the watermarking channel. Recalling that the embedder does not have access to content wherein no watermark is to be inserted, we have in this case that the detector observable is given by $\mathbf{r_0}' = \mathbf{r_0} + \mathbf{n}$. If, on the other hand, presence of a watermark is to be conveyed, then ignoring for the moment any distortive constraints, an intuitively pleasing transmission strategy calls for the transmitter to set $\mathbf{r_w} = \mathbf{w}$ for some fixed watermark vector $\mathbf{w}$ in a K-dimensional watermark space; this removes all uncertainty at the detector associated with the content vector $\mathbf{r_0}$, and generates the detector observable $\mathbf{r_w}' = \mathbf{w} + \mathbf{n}$. Although the capability of setting the content vector to a fixed watermark would seem optimistic in a real watermarking application, it affords an analysis which produces an effective watermark detection strategy and is less optimistic in cases where the dimensionality of watermark space is significantly lower than that of content space.

We are led to consider the distribution of the channel output $\mathbf{r_u}'$ conditioned on presence or absence of a watermark, where the subscript $\mathbf{u}$ conveys the notion that such presence or absence is unknown at the detector. These distributions depend on the statistical description of the channel distortion vector $\mathbf{n}$ modelling various effects encountered between watermark insertion and detection such as cropping and D/A followed by A/D conversion. As illustrated in Table 2, an analysis of real images subjected to a variety of corruptive processes demonstrates a significant correlation between the distortion vector $\mathbf{n}$ and the transmitted content vector $\mathbf{r_u}$. We model such correlation by defining $\mathbf{n}$ to be a Gaussian random vector with zero mean and covariance matrix

$$\Lambda_{\mathbf{r_u}} \stackrel{\triangle}{=} \sigma_n^2[\rho\,\mathbf{r_u}\mathbf{r_u}^T + (1-\rho)I]$$

as in Equation 1 where σ_n^2 represents the overall strength of the distortion and the parameter ρ, $0 \leq \rho \leq 1$, determines the extent of the noted correlation between $\mathbf{n}$ and $\mathbf{r_u}$.

When a fixed watermark is transmitted, the detector observable $\mathbf{r_w}' = \mathbf{w} + \mathbf{n}$ is a Gaussian random vector with mean $\mathbf{w}$ and covariance $\Lambda_{\mathbf{w}}$. The probability density function of the output $\mathbf{r_u}'$ conditioned on watermark presence is then given by

$$f(\mathbf{r_u}'|\text{watermark}) = \frac{1}{(2\pi)^{K/2}|\Lambda_{\mathbf{w}}|^{1/2}}\exp\{-(\mathbf{r_u}'-\mathbf{w})^T\Lambda_{\mathbf{w}}^{-1}(\mathbf{r_u}'-\mathbf{w})/2\}$$

where $|\Lambda_{\mathbf{w}}|$ is the determinant of the covariance matrix $\Lambda_{\mathbf{w}}$. In the absence of a watermark, the random vector $\mathbf{n}$ with covariance matrix $\Lambda_{\mathbf{r_0}} = \sigma_n^2[\rho\mathbf{r_0}\mathbf{r_0}^T + (1-\rho)I]$ does not yield a straightforward characterization as $\mathbf{r_0}$ is not deterministic. In this case, we make the simplifying assumption that $\mathbf{r_0} + \mathbf{n}$ is a zero-mean Gaussian random vector with covariance matrix $(\sigma_c^2 + \sigma_n^2)I$, which tends to an exact description as either of σ_n^2/σ_c^2 or ρ tends to zero, leading to the conditional channel output probability density function

$$f(\mathbf{r_u}'|\text{no watermark}) = \frac{1}{(2\pi(\sigma_c^2+\sigma_n^2))^{K/2}}\exp\{-\mathbf{r_u}'^T\mathbf{r_u}'/2(\sigma_c^2+\sigma_n^2)\}.$$

In typical watermarking applications, the probability of a false positive, which occurs whenever the detector concludes presence of an absent watermark, is a primary concern. This property leads to a natural characterization of the detection process as a Neyman-Pearson hypothesis test (see for instance [Poo94]), wherein the probability of proper watermark detection is maximized subject to a prescribed limit on the probability of a false positive. In our case, the Neyman-Pearson test is achieved by a likelihood-ratio test, which takes the form of a comparison of the likelihood ratio

$$L(\mathbf{r_u}') = f(\mathbf{r_u}'|\text{watermark})/f(\mathbf{r_u}'|\text{no watermark})$$

with a threshold chosen to satisfy the prescribed false-positive tolerance with equality; if the threshold is surpassed (or met with equality), watermark presence is concluded, and watermark absence is concluded otherwise.

Monotonicity of the natural logarithm $\log(\cdot)$ allows us to consider the equivalent hypothesis test generated by the log-likelihood ratio

$$l(\mathbf{r_u}') \quad = \quad \log f(\mathbf{r_u}'|\text{watermark}) - \log f(\mathbf{r_u}'|\text{no watermark})$$

which is proportional to each of the quantities (assuming $\sigma_c^2 > 0$ and $\sigma_n^2 > 0$)

$$\frac{\mathbf{r_u}'^T\mathbf{r_u}'}{2(\sigma_c^2+\sigma_n^2)} - \frac{(\mathbf{r_u}'-\mathbf{w})^T\Lambda_{\mathbf{w}}^{-1}(\mathbf{r_u}'-\mathbf{w})}{2},$$

$$\frac{\mathbf{r_u}'^T\mathbf{r_u}'}{2(\sigma_c^2+\sigma_n^2)} - \frac{1}{2(1-\rho)\sigma_n^2}(\mathbf{r_u}'-\mathbf{w})^T\left[I - \frac{\rho\mathbf{w}\mathbf{w}^T}{1+\rho(\mathbf{w}^T\mathbf{w}-1)}\right](\mathbf{r_u}'-\mathbf{w}) \tag{2}$$

and

$$t(\mathbf{r_u}') \triangleq \rho(\mathbf{r_u}'^T\mathbf{w})^2 - \frac{(\sigma_c^2+\rho\sigma_n^2)(1+\rho(\mathbf{w}^T\mathbf{w}-1))}{\sigma_c^2+\sigma_n^2}\mathbf{r_u}'^T\mathbf{r_u}' + 2(1-\rho)\mathbf{r_u}'^T\mathbf{w}, \tag{3}$$

where (2) follows from the matrix inverse identity

$$\left[I+\mathbf{u}\mathbf{u}^T\right]^{-1} = I - \frac{\mathbf{u}\mathbf{u}^T}{1+||\mathbf{u}||^2}$$

for a general Kx1 vector $\mathbf{u}$. Hence, an optimal detector concludes watermark presence if and only if $t(\mathbf{r_u}') \geq \tau$ where τ is chosen so that

$$P(t(\mathbf{r_u}') \geq \tau|\text{no watermark}) = P(t(\mathbf{r_0}') \geq \tau) = \alpha$$

where $P(A|B)$ denotes the probability of an event A conditioned on an event B, and where α is the prescribed limit on false positive probability. If we fix $\rho > 0$, the detector tends with growing K to conclude watermark presence if and only if

$$\frac{(\mathbf{r_u}'^T\mathbf{w})^2}{||\mathbf{r_u}'||^2||\mathbf{w}||^2} \geq \frac{\tau'}{||\mathbf{r_u}'||^2} + \frac{\sigma_c^2+\rho\sigma_n^2}{\sigma_c^2+\sigma_n^2} \tag{4}$$

where we have assumed that $||\mathbf{w}||^2$ grows at least linearly in K in order that the watermark signal-to-noise ratio

$$||\mathbf{w}||^2/E[||\mathbf{n}||^2] = (\sigma_n^2[\rho + K(1-\rho)/||\mathbf{w}||^2])^{-1}$$

remain bounded away from zero, and where τ' is again chosen to meet the false positive requirement with equality. The detection region corresponding to (4) is depicted in Figure 5 for the values $\sigma_c^2 = 10$, $\sigma_n^2 = 1$, $\rho = .5$ and varying values of τ'. When $\tau' = 0$, the detector test statistic is given by the normalized correlation $|\mathbf{r_u}'^T\mathbf{w}|/(||\mathbf{r_u}'||\,||\mathbf{w}||)$. We are led in this case to a detection scheme in which the linear relationship between the observed vector $\mathbf{r_u}'$ and the watermark vector $\mathbf{w}$, as measured by their described angle, is compared with a prescribed threshold.

Although threshold values other than $\tau' = 0$ describe hyperbolic detection regions according to (4), we propose detectors which in general conclude watermark presence if and only if

$$\frac{|\mathbf{r_u}'^T\mathbf{w}|}{||\mathbf{r_u}'||\,||\mathbf{w}||} \geq \tau''(\alpha) \tag{5}$$

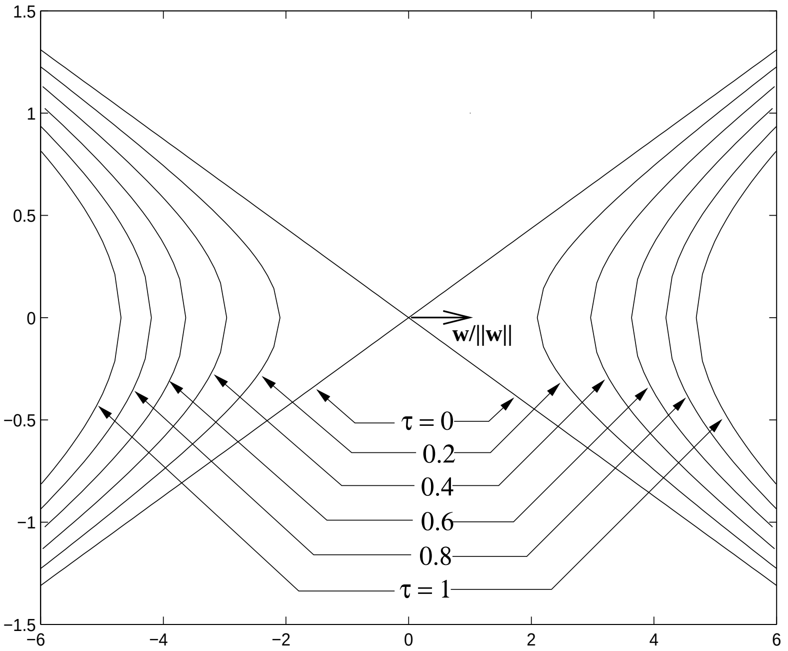


Figure 5: Asymptotic detection regions corresponding to varying thresholds τ for the values $\sigma_c^2 = 10$, $\sigma_n^2 = 1$ and $\rho = .5$. The x-axis represents the directions parallel to the watermark vector and the y-axis is any direction perpandicular to the watermark vector.

for reasons of detector simplicity and robustification of the false positive analysis, where the dependence of the threshold τ'' on the prescribed false positive probability α has been expressed. The robustness follows from the property that, for large K, the random variable

$$Z = \frac{\sqrt{K-3}}{2} \log \frac{1 + \mathbf{r_0}'^T \mathbf{w}/||\mathbf{r_0}'||\,||\mathbf{w}||}{1 - \mathbf{r_0}'^T \mathbf{w}/||\mathbf{r_0}'||\,||\mathbf{w}||}$$

tends to Gaussian distribution with zero mean and unit variance *independent of the values of* σ_c^2 *or* $\mathbf{w}$ [Cox91]. Hence, no statistical description of the content vector covariance parameter σ_c^2 is required in order to develop expressions for false positive probabilities in the determination of an appropriate prescribed threshold $\tau''(\alpha)$. This is a particularly beneficial trait of the the normalized-correlation detector (5) in light of the discrepancy in statistical properties exhibited by different *types* of content vectors.

7 Watermark Insertion

In early watermarking applications, watermark insertion amounted to the addition of a predetermined watermark vector to the original content to be embedded. This approach stemmed from the philosophy that the content vector represented noise from the point of view of the watermark embedder/detector pair. In essence, the addition of a constant watermark vector translated the underlying statistical distribution of content vectors, with the aim that the watermarked content vector lie in an appropriate detection region the majority of the time; a large enough acceptable distortion level would allow the addition of a strong enough watermark vector to achieve such an aim.

In the proposed watermarking strategy, the embedder makes use of the knowledge of the content vector rather than treating it as unknown noise. Thus, while the underlying content vector is still viewed as noise from the point of view of the watermarking detector, recognition of the side information available to the embedder in the form of the content realization leads to a transmission strategy exhibiting a significant performance improvement.

Under the proposed scheme, the embedder uses knowledge of the content vector $\mathbf{r_0}$ to compute a region $\mathcal{S}(\mathbf{r_0})$ within which transformed content vectors satisfy acceptable distortion tolerances based on prescribed content-dependent fidelity measures. The goal then becomes to pick an embedded content vector $\mathbf{r_w}$ from within $\mathcal{S}(\mathbf{r_0})$ which will lead the receiver to conclude presence of a watermark. For a fixed detection strategy and false positive tolerance α, detection is described by a region $\mathcal{R}(\alpha)$ of receptions which are considered to represent watermark presence.

In the absence of channel distortions, it is clear that any embedded vector $\mathbf{r_w}$ selected from the intersection of $\mathcal{S}(\mathbf{r_0})$ and $\mathcal{R}(\alpha)$ will exhibit acceptable distortion *and* lead to correct watermark detection; if the regions $\mathcal{S}(\mathbf{r_0})$ and $\mathcal{R}(\alpha)$ do not overlap for a given content realization $\mathbf{r_0}$ and false positive prescription, then watermark presence can not be conveyed. However, if we consider the potential for channel distortions, determination of an effective transmission vector becomes a more involved task. Even if the intersection of $\mathcal{S}(\mathbf{r_0})$ and $\mathcal{R}(\alpha)$ is non-empty, it is profitable for the embedder to select a transmission from within the intersection that is maximally robust to channel distortions, so that a reasonable probability of correct watermark detection is maintained even following any signal processing applied to the embedded vector.

As an example, consider watermark detection based on the normalized correlation between a given content vector $\mathbf{r_w}$ and a prescribed watermark vector $\mathbf{w}$, characterized by the detection region

$$\mathcal{R}(\alpha) \quad = \quad \left\{ \mathbf{r_w}' : \frac{|\mathbf{r_w}'^T \mathbf{w}|}{||\mathbf{r_w}'|| \, ||\mathbf{w}||} \geq \tau(\alpha) \right\} \ . \tag{6}$$

Such a detection scheme has previously been considered for watermarking applications ([CKLS97, PZ98]), at least in part because of the reduction in statistical modelling of the content vectors required in order to carry through a false positive analysis. Further justification for such a detection scheme was also developed in the previous section.

In order to consider detection robustification based on this strategy, we assume that the reception $\mathbf{r_w}'$, representing the observable upon which the determination of watermark presence is based, is the superposition of the embedded vector $\mathbf{r_w}$ and additive distortion $\mathbf{n}$. We assume that $\mathbf{n}$ is Gaussian distributed with zero mean and covariance matrix $\Lambda_{\mathbf{n}} = \sigma_n^2 I$, where σ_n^2 represents the average strength of encountered distortive effects. Although we have ignored the tendency for a correlation between $\mathbf{r_w}$ and $\mathbf{n}$ addressed in earlier sections, we note that such an assumption will not significantly alter the results to follow, and discuss the generalization of the embedding process to arbitrary models at the end of this section. Then, for a given embedded vector $\mathbf{r_w}$, the probability of correct detection $P_{\mathbf{r_w}}$ is given according to (6) by

$$\begin{aligned} P_{\mathbf{r_w}} &= P(\mathbf{r_w} + \mathbf{n} \in \mathcal{R}(\alpha)) \\ &= P\left(\frac{|(\mathbf{r_w} + \mathbf{n})^T \mathbf{w}|}{||\mathbf{r_w} + \mathbf{n}||\,||\mathbf{w}||} \geq \tau(\alpha)\right) . \end{aligned}$$

The distribution of the random variable $\mathbf{n}^T\mathbf{w}$ is Gaussian with zero mean and variance $\sigma_n^2||\mathbf{w}||^2$, while that of $\mathbf{n}^T\mathbf{r_w}$ is zero-mean Gaussian with variance $\sigma_n^2||\mathbf{r_w}||^2$. When the content vectors lie in a watermark vector space of high dimension K, the distribution of $||\mathbf{n}||^2$ tends to a Gaussian with mean $K\sigma_n^2$ and variance $2K\sigma_n^4$. Hence, for large K we have that

$$\begin{aligned} P_{\mathbf{r_w}} &\simeq P\left(\frac{|\mathbf{r_w}^T\mathbf{w}|}{\sqrt{||\mathbf{r_w}||^2 + K\sigma_n^2}||\mathbf{w}||} \geq \tau''\right) \\ &= \begin{cases} 0\ , & \frac{|\mathbf{r_w}^T\mathbf{w}|}{\sqrt{||\mathbf{r_w}||^2+K\sigma_n^2}||\mathbf{w}||} \geq \tau'' \\ 1\ , & \text{otherwise} \end{cases} \end{aligned} \tag{7}$$

Roughly, then, a given embedded content vector $\mathbf{r_w}$ within the detection region can be characterized by its *resistance* to distortive effects, where, with high probability, distortion with strength

$$\frac{K\sigma_n^2 \leq (\mathbf{r_w}^T\mathbf{w})^2}{||\mathbf{w}||^2\tau''(\alpha)^2 - ||\mathbf{r_w}||^2}$$

will leave the watermark unharmed, while stronger distortion will lead the receiver to incorrectly conclude watermark absence. It is evident that an optimal embedder strategy consists of picking from the set $\mathcal{S}(\mathbf{r_0})$ of vectors with acceptable perceptual distortion the embedded vector $\mathbf{r_w}$ which maximizes the quantity

$$\frac{(\mathbf{r_w}^T\mathbf{w})^2}{||\mathbf{w}||^2\tau''(\alpha)^2 - ||\mathbf{r_w}||^2}$$

in order to achieve maximal robustness to subsequent distortive effects. Note that surfaces of equal robustness are rotationally symmetric about the watermark axis according to (7). In Figure 7, a sample of contours of equal robustness are plotted in an arbitrary plane containing the watermark vector, which is assumed to have unit magnitude.

We have shown how an embedder might exploit knowledge of specific content in order o robustify the embedding process to subsequent distortions, including attacks. While different watermark detection strategies and distortion models will require that alternative expressions be optimized in order to achieve maximum robustness to signal processing and other distortions, the basic idea remains that the acceptable set of embedded vectors for a given content realization forms the feasible set for the optimization of a robustness expression, for instance that given in (7). In this manner, the side information provided to the embedder is used to generate an optimal insertion strategy for a given channel model.

8 Conclusions and Future Work

In this paper, we have illustrated some basic similarities and differences between watermarking and traditional communications. Content (cover data) has historically been viewed as a form of noise,

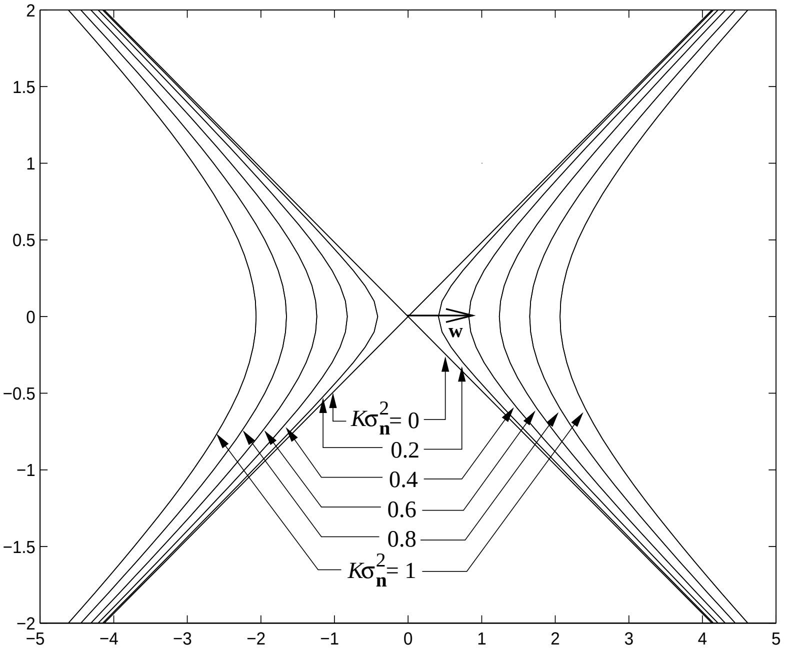


Figure 6: Equal robustness contours in an arbitrary plane containing a watermark vector of unit magnitude for the value $\tau''(\alpha) = 0.9$. Contours are resistant to distortion of strength $K\sigma_n^2$. The x-axis represents the directions parallel to the watermark vector and the y-axis is any direction perpandicular to the watermark vector.

and watermarks treated as transmissions with very low signal-to-noise ratios. However, as we have pointed out here, viewing knowledge of the cover data as side information at the transmitter allows the design of more powerful watermark embedding algorithms. In particular, it becomes possible to calculate the robustness of watermarked data to subsequent attacks, and to maximize robustness within a specified distortive constraint.

We have further argued that an effective watermark detection region is formed by a K dimensional, two-sheet hyperboloid. But, when it is important to place tight upper bounds on false positive rates, the detection region formed by thresholding a normalized correlation is probably preferable.

While these observations promise a significant performance improvement, there is still much room for future work. This might include

- analyzing more accurate models of the distribution of distortion vectors. In particular, our analysis of the embedding process is based on a simplified model of an independent Gaussian distribution. A more accurate model might be considered in order to improve overall performance. One approach might be to make use of the side information available at the embedder to predict the likely distribution of future distortions to the watermarked data.

- analyzing potential performance improvement given side information at the detector. Several existing watermarking algorithms assume knowledge of the original, unwatermarked cover data at the detector. Typically, these algorithms subtract the unwatermarked data from the (possibly) watermarked data to reconstruct the watermark. It may be that more sophisticated algorithms can be developed.

- assessing the capacity of the watermarking channel. Shannon's article [Sha58] is devoted to analyzing the capacity of communication channels with side information at the transmitter. It would be interesting to apply a similar analysis to the particular application of watermarking.

Acknowledgements

The authors thank Peter Blicher, Jeff Bloom, Ryoma Oami, Harold Stone and Kazuyoshi Tanaka for their helpful contributions.

References

[Bay73] B. E. Bayer. An optimal method for two-level rendition of continuous-tone pictures. In *International Conference on Communications, Conference Record*, pages (26–11)–(26–15), 1973.

[BCK$^+$99] J. A. Bloom, I. J. Cox, T. Kalker, J-P Linnartz, M. L. Miller, and B. Traw. Copy protection for dvd video. *Proceedings of the IEEE*, 1999.

[BGML96] W. Bender, D. Gruhl, N. Morimoto, and A. Lu. Techniques for data hiding. *IBM Systems Journal*, 35(3/4):313–336, 1996.

[Car95] G. Caronni. Assuring ownership rights for digital images. In *Proc. Reliable IT Systems, VIS'95*. Vieweg Publishing Company, 1995.

[CKLS97] I.J. Cox, J. Kilian, F. T. Leighton, and T. Shamoon. Secure spread spectrum watermarking for images, audio and video. *IEEE Trans. on Image Processing*, 6(12):1673–1687, 1997.

[CL97] I.J. Cox and J.-P. M. G. Linnartz. Public watermarks and resistance to tampering. In *Proceedings of the IEEE International Conference on Image Processing, CDRom*, 1997.

[CL98] I.J. Cox and J.-P. M. G. Linnartz. Some general methods for tampering with watermarks. *IEEE Trans. on Selected Areas in Communications*, 16(4):587–593, 1998.

[CM97] I.J. Cox and M. L. Miller. A review of watermarking and the importance of perceptual modeling. In *Proceedings of SPIE, Human Vision & Electronic Imaging II*, volume 3016, pages 92–99, 1997.

[Cox91] C. P. Cox. *A Handbook of Introductory Statistical Methods.* John Wiley & Sons, 1991.

[HMW88] L. Holt, B. G. Maufe, and A. Wiener. Encoded marking of a recording signal. UK Patent GB 2196167A, 1988.

[HPGRN98] J. J. 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 still images. *IEEE Trans. on Selected Areas of Communications*, 16(4):510–524, 1998.

[HW96] C-T Hsu and J-L Wu. Hidden signatures in images. In *IEEE Int. Conf. on Image Processing*, 1996.

[Kal98] Ton Kalker. Watermark estimation through detector observations. In *Proceedings of the IEEE Benelux Signal Processing Symposium*, pages 119–122, Leuven, Belgium, March 1998.

[KLvD98] T. Kalker, J.P Linnartz, and M. van Dijk. Watermark estimation through detector analysis. In *Proceedings of the ICIP*, Chicago, October 1998.

[LvD98] J.-P. M. G. Linnartz and Marten van Dijk. Analysis of the sensitivity attack against electronic watermarks in images. In *Workshop on Information Hiding, Portland, OR*, 15-17 April, 1998.

[MT94] K. Matsui and K. Tanaka. Video-steganography. In *IMA Intellectual Property Project Proceedings*, volume 1, pages 187–206, 1994.

[PAK98] F.A.P. Petitcolas, R.J. Anderson, and M.G. Kuhn. Attacks on copyright marking systems. In *Workshop on Information Hiding, Portland, OR*, 15-17 April, 1998.

[Pfi96] B. Pfitzman. Information hiding terminology. In *Info-Hiding 96*, pages 347–350, 1996.

[Poo94] H. V. Poor. *An Introduction to Signal Detection and Estimation.* Springer-Verlag, 1994.

[PRH+94] R. D. Preuss, S. E. Roukos, A. W. F. Huggins, H. Gish, M. A. Bergamo, P. M. Peterson, and A. G. Derr. Embedded signalling. US Patent 5,319,735, 1994.

[PSM82] R. L. Pickholtz, D. L. Schilling, and L. B. Millstein. Theory of spread spectrum communications - a tutorial. *IEEE Trans. on Communications*, pages 855–884, 1982.

[PZ98] C. I. Podilchuk and W. Zeng. Image-adaptive watermarking using visual models. *IEEE Trans. on Selected Areas of Communications*, 16(4):525–539, 1998.

[RDB96] J. J. K. O Ruanaidh, W. J. Dowling, and F.M. Boland. Phase watermarking of digital images. In *IEEE Int. Conf. on Image Processing*, 1996.

[Rho95] G. B. Rhoads. Indentification/authentication coding method and apparatus. *World Intellectual Property Organization*, IPO WO 95/14289, 1995.

[SC96] J. R. Smith and B. O. Comiskey. Modulation and information hiding in images. In R. Anderson, editor, *Information Hiding: First Int. Workshop Proc.*, volume 1174 of *Lecture Notes in Computer Science*, pages 207–226. Springer-Verlag, 1996.

[Sha58] C. E. Shannon. Channels with side information at the transmitter. *IBM Journal of Research and Development*, pages 289–293, 1958.

[SZT98] M. D. Swanson, B. Zhu, and A. H. Tewfik. Multiresolution scene-based video watermarking using perceptual models. *IEEE Journal on Selected Areas in Communications*, 16(4):540–550, 1998.

[TNM90] K. Tanaka, Y. Nakamura, and K. Matsui. Embedding secret information into a dithered multi-level image. In *Proc, 1990 IEEE Military Communications Conference*, pages 216–220, 1990.

[Tur89] L. F. Turner. Digital data security system. Patent IPN WO 89/08915, 1989.