A Secure, Robust Watermark for Multimedia

Ingemar J. Cox†, Joe Kilian†, Tom Leighton‡ and Talal Shamoon†*

Abstract

We describe a digital watermarking method for use in audio, image, video and multimedia data. We argue that a watermark must be placed in perceptually significant components of a signal if it is to be robust to common signal distortions and malicious attack. However, it is well known that modification of these components can lead to perceptual degradation of the signal. To avoid this, we propose to insert a watermark into the spectral components of the data using techniques analogous to spread sprectrum communications, hiding a narrow band signal in a wideband channel that is the data. The watermark is difficult for an attacker to remove, even when several individuals conspire together with independently watermarked copies of the data. It is also robust to common signal and geometric distortions such as digital-to-analog and analog-to-digital conversion, resampling, and requantization, including dithering and recompression and rotation, translation, cropping and scaling. The same digital watermarking algorithm can be applied to all three media under consideration with only minor modifications, making it especially appropriate for multimedia products. Retrieval of the watermark unambiguously identifies the owner, and the watermark can be constructed to make counterfeiting almost impossible. Experimental results are presented to support these claims.

1 Introduction

The proliferation of digitized media (audio, image and video) is creating a pressing need for copyright enforcement schemes that protect copyright ownership. Conventional cryptographic systems permit only valid keyholders access to encrypted data, but once such data is decrypted there is no way to track its reproduction or retransmission. Conventional cryptography therefore provides little protection against data piracy, in which a publisher is confronted with unauthorized reproduction of information. A digital watermark is intended to complement cryptographic processes. It is a visible, or preferably invisible, identification code that is permanently embedded in the data, that is, it remains present within the data after any decryption process. In the context of this work, data refers to audio (speech and music), images (photographs and graphics), and video (movies). It does not include ASCII representations of text, but does include text represented as an image. A simple example of a digital watermark would be a visible "seal" placed over an image to identify the copyright owner. However, the watermark might contain additional information, including the identity of the purchaser of a particular copy of the material.

In order to be effective, a watermark should be:

Unobtrusive The watermark should be perceptually invisible, or its presence should not interfere with the work being protected.

Robust The watermark must be difficult (hopefully impossible) to remove. Of course, in theory, any watermark may be removed with sufficient knowledge of the process of insertion. However, if only partial knowledge is available, for example, the exact location of the watermark within an image is unknown, then attempts to remove or destroy a watermark by say, adding noise, should result in severe degradation in data fidelity before the watermark is lost. In particular, the watermark should be robust to

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

[0] ‡Post:Mathematics Department and Laboratory for Computer Science, MIT, Cambridge, MA 02139.
Email: ftl@math.mit.edu

[*] Authors appear in alphabetical order.

[0] Portions of this paper are based on an abridged version "A Secure, Imperceptable yet Perceptually Salient, Spread Spectrum Watermark for Multimedia" which appeared in the Proc. of Southcon'96, June 1996 ©IEEE.

Common signal processing The watermark should still be retrievable even if common signal processing operations are applied to the data. These include, digital-to-analog and analog-to-digital conversion, resampling, requantization (including dithering and recompression), and common signal enhancements to image contrast and color, or audio bass and treble, for example.

Common geometric distortions (image and video data) Watermarks in image and video data should also be immune from geometric image operations such as rotation, translation, cropping and scaling.

Subterfuge Attacks: Collusion and Forgery In addition, the watermark should be robust to collusion by multiple individuals who each possess a watermarked copy of the data. That is, the watermark should be robust to combining copies of the same data set to destroy the watermarks. Further, if a digital watermark is to be used as evidence in a court of law, it must not be possible for colluders to combine their images to generate a different valid watermark with the intention of framing a third-party.

Universal The same digital watermark algorithm should apply to all three media under consideration. This is potentially helpful in the watermarking of multimedia products. Also, this feature is conducive to implementation of audio and image/video watermarking algorithms on common hardware.

Unambiguous Retrieval of the watermark should unambiguously identify the owner. Further, the accuracy of owner identification should degrade gracefully in the face of attack.

Previous digital watermarking techniques [1, 2, 4, 5, 12, 13, 16, 17, 19–22] are not robust, and the watermark is easy to remove. In addition, it is unlikely that any of the earlier watermarking methods would survive common signal and geometric distortions. The principal reason for these weaknesses is that previous methods have not explicitly identified the perceptually most significant components of a signal as the destination for the watermark. In fact, it is often the case that the perceptually significant regions are explicitly avoided. The reason for this is obvious – modification of perceptually significant components of a signal results in perceptual distortions much earlier than if the modifications are applied to perceptually insignificant regions. Hence, for example, the common stategy of placing a watermark in the high frequency components of a signal's spectrum.

The key insight of this paper is that in order for it to be robust, the watermark *must* be placed in perceptually significant regions of the data despite the risk of potential fidelity distortions. Conversely, if the watermark is placed in perceptually insignificant regions, it is easily removed, either intentionally or unintentionally by, for example, signal compression techniques that implicitly recognize that perceptually weak components of a signal need not be represented.

The perceptually significant regions of a signal may vary depending on the particular media (audio, image or video) at hand, and even within a given media. For example, it is well known that the human visual system is tuned to certain spatial frequencies and to particular spatial characteristics such as line and corner features. Consequently, many watermarking schemes that focus on different phenomena that are perceptually significant are potentially possible. In this paper, we focus on perceptually significant *spectral* components of a signal.

Section 2 begins with a discussion of how common signal transformations, such as compression, quantization and manipulation, affect the frequency spectrum of a signal. This motivates why we believe that a watermark should be embedded in the data's perceptually significant frequency components. Of course, the major problem then becomes how to insert a watermark into perceptually significant components of the frequency spectrum without introducing visible or audible distortions. Section 2.2 proposes a solution based on ideas from spread spectrum communications.

The structure of a watermark may be arbitrary. However, Section 3 provides an analysis based on possible collusion attacks that indicates that a binary watermark is not as robust as a continuous one. Furthermore, we show that a watermark structure based on sampling drawn from multiple i.i.d Gaussian random variables offers good protection against collusion.

Of course, no watermarking system can be made perfect. For example, a watermark placed in a textual image may be eliminated by using optical character recognition technology. However, for common signal and geometric distortions, the experimental results of Section 4 strongly suggest that our system satisfies *all* of the properties discussed in the introduction, and displays strong immunity to a wide variety of attacks, though

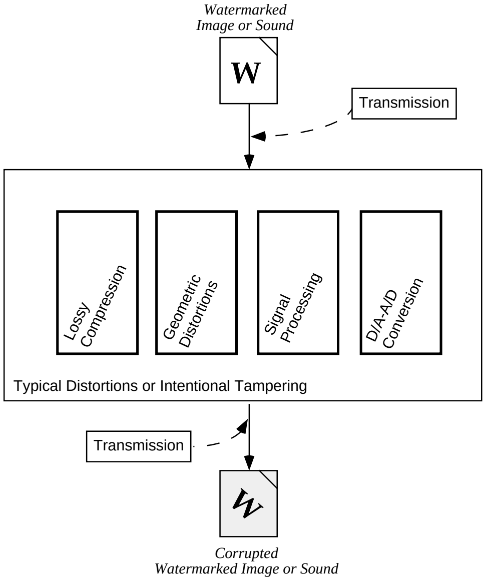

Figure 1: Common processing operations that a media document could undergo

more extensive experiments are needed to confirm this. Finally, Section 5 discusses possible weaknesses and enhancements to the system.

2 Watermarking in the Frequency Domain

In this section, we first discuss how common signal distortion affect the frequency spectrum of a signal. This analysis supports our contention that a watermark must be placed in perceptually significant regions of a signal if it is to be robust. Section 2.2 proposes inserting a watermark into the perceptually most significant components of the spectrum using spread spectrum techniques.

2.1 Common signal distortions and their effect on the frequency spectrum of a signal

In order to understand the advantages of a frequency-based method, it is instructive to examine the processing stages that an image (or sound) may undergo in the process of copying, and to study the effect that these stages could have on the data, as illustrated in Figure 1. In the figure, “transmission” refers to the application of any source or channel code, and/or standard encryption technique to the data. While most of these steps are information lossless, many compression schemes (JPEG, MPEG etc.) can potentially degrade the data’s quality, through *irretrievable* loss of data. In general, a watermarking scheme should be resilient to the distortions introduced by such algorithms.

Lossy compression is an operation that usually eliminates perceptually non-salient components of an image or sound. If one wishes to preserve a watermark in the face of such an operation, the watermark must be placed in the perceptually significant regions of the data. Most processing of this sort takes place in the frequency domain. In fact, data loss usually occurs among the high frequency components. Hence, the watermark must be placed in the *significant* frequency components of the image (or sound) spectrum.

After receipt, an image may endure many common transformations that are broadly categorized as geometric distortions or signal distortions. Geometric distortions are specific to images and video, and include such operations as rotation, translation, scaling and cropping. By manually determining a minimum of four or nine corresponding points between the original and the distorted watermark, it is possible to remove any two or three dimensional affine transformation [7]. However, an affine scaling (shrinking) of the image leads to a loss of data in the high frequency spectral regions of the image. Cropping, or the cutting out and removal of portions of an image, also leads to irretrievable loss of data. Cropping may be a serious threat to any spatially based watermark such as [5] but is less likely to affect a frequency-based scheme, as shown in Section 4.5.

Common signal distortions include digital-to-analog and analog-to-digital conversion, resampling, requantization, including dithering and recompression, and common signal enhancements to image contrast and/or color, and audio frequency equalization. Many of these distortions are non-linear, and it is difficult to analyze their effect in either a spatial or frequency based method. However, the fact that the original image is known allows many signal transformations to be undone, at least approximately. For example, histogram equalization, a common non-linear contrast enhancement method, may be removed substantially by histogram specification [9] or dynamic histogram warping [6] techniques.

Finally, the copied image may not remain in digital form. Instead, it is likely to be printed, or an analog recording made (onto analog audio or video tape). These reproductions introduce additional degradation into the image that a watermarking scheme must be robust to.

The watermark must not only be resistant to the inadvertant application of the aforementioned distortions. It must also be immune to intentional manipulation by malicious parties. These manipulations can include combinations of the above distortions, and can also include collusion and forgery attacks.

2.2 Spread spectrum coding of a watermark

The above discussion makes it clear that the watermark should *not* be placed in perceptually insignificant regions of the image or its spectrum since many common signal and geometric processes affect these components. For example, a watermark placed in the high frequency spectrum of an image can be easily eliminated with little degradation to the image by any process that directly or indirectly performs low pass filtering. The problem then becomes how to insert a watermark into the most perceptually significant regions of an spectrum without such alterations becoming noticeable. Clearly, any spectral coefficient may be altered, provided such modification is small. However, very small changes are very susceptible to noise.

To solve this problem, the frequency domain of the image or sound at hand is viewed as a *communication channel*, and correspondingly, the watermark is viewed as a signal that is transmitted through it. Attacks and unintentional signal distortions are thus treated as noise that the immersed signal must be immune to. While we use this methodology to hide watermarks in data, the same rationale can be applied to sending any type of message through media data.

Rather than encode the watermark into the least significant components of the data, we originally conceived our approach by analogy to spread spectrum communications [18]. In spread spectrum communications, one transmits a narrowband signal over a much larger bandwidth such that the signal energy present in any single frequency is imperceptible. Similarly, the watermark is spread over very many frequency bins so that the energy in any one bin is very small and certainly undetectable. Nevertheless, because the watermark verification process knows of the location and content of the watermark, it is possible to concentrate these many weak signals into a single signal with high signal-to-noise ratio. However, to confidently destroy such a watermark would require noise of high amplitude to be added to *all* frequency bins.

Spreading the watermark throughout the spectrum of an image ensures a large measure of security against unintentional or intentional attack: First, the spatial location of the watermark is not obvious. Furthermore, frequency regions should be selected in a fashion that ensures severe degradation of the original data following any attack on the watermark.

A watermark that is well placed in the frequency domain of an image or a sound track will be practically impossible to see or hear. This will always be the case if the energy in the watermark is sufficiently small in any single frequency coefficient. Moreover, it is possible to increase the energy present in particular frequencies by exploiting knowledge of masking phenomena in the human auditory and visual systems. Perceptual masking refers to any situation where information in certain regions of an image or a sound is occluded by perceptually more prominent information in another part of the scene. In digital waveform coding, this frequency domain (and, in some cases, time/pixel domain) masking is exploited extensively to achieve low bit rate encoding of data [8, 11]. It is clear that both the auditory and visual systems attach more resolution to the high energy, low frequency, spectral regions of an auditory or visual scene [11]. Further, spectrum analysis of images and sounds reveals that most of the information in such data is located in the low frequency regions.

Figure 2 illustrates the general procedure for frequency domain watermarking. Upon applying a frequency transformation to the data, a *perceptual mask* is computed that highlights perceptually significant regions

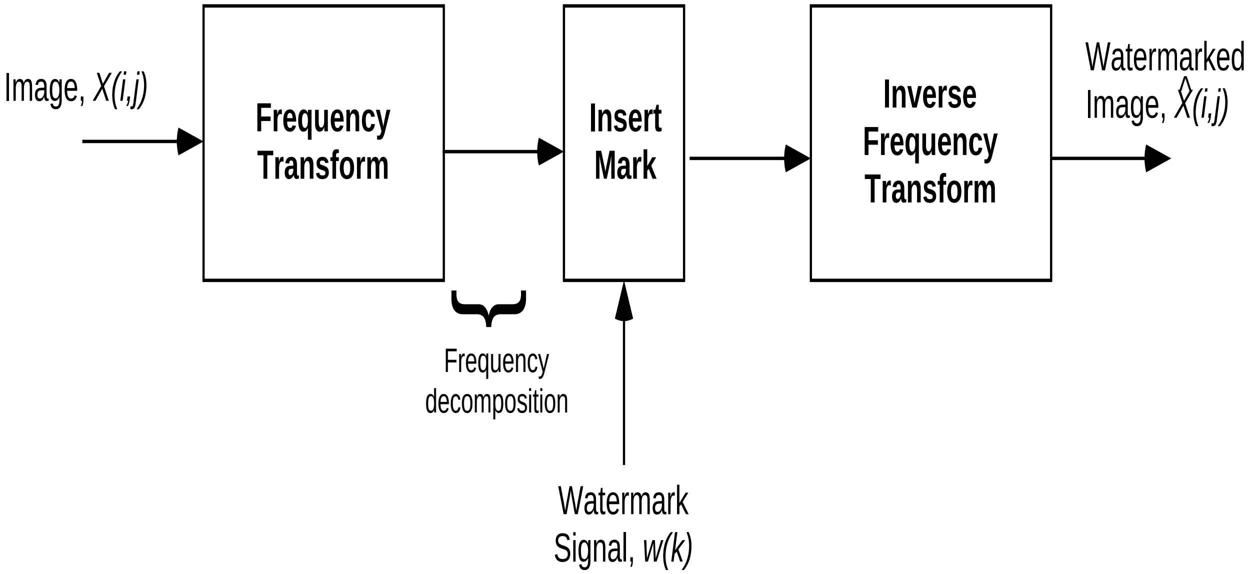

Figure 2: Immersion of the watermark in the frequency domain

in the spectrum that can support the watermark without affecting perceptual fidelity. The watermark signal is then inserted into these regions in a manner described in Section 3.2. The precise magnitude of each modification is only known to the owner. By contrast, an attacker may only have knowledge of the possible range of modification. To be confident of eliminating a watermark, an attacker must assume that each modification was at the limit of this range, despite the fact that few such modifications are typically this large. As a result, an attack creates visible (or audible) defects in the data. Similarly, unintentional signal distortions due to compression or image manipulation, must leave the perceptually significant spectral components intact, otherwise the resulting image will be severely degraded. This is why the watermark is robust.

In principle, any frequency domain transform can be used. However, for the experimental results of Section 4 we use a Fourier domain method based on the discrete cosine transform (DCT) [15], although we are currently exploring the use of wavelet-based schemes as a variation. In our view, each coefficient in the frequency domain has a *perceptual capacity*, that is, a quantity of additional information can be added without any (or with minimal) impact to the perceptual fidelity of the data. To determine the perceptual capacity of each frequency, one can use models for the appropriate perceptual system or simple experimentation.

In practice, in order to place a length n watermark into an $N \times N$ image, we computed the $N \times N$ DCT of the image and placed the watermark into the n highest magnitude coefficients of the transform matrix, excluding the DC component.[1] For most images, these coefficients will be the ones corresponding to the low frequencies. Reiterating, the purpose of placing the watermark in these locations is because significant tampering with these frequency will destroy the image fidelity well before the watermark.

In the next section, we provide a high level discussion of the watermarking procedure, describing the structure of the watermark and its characteristics.

3 Structure of the watermark

We now give a high-level overview of our a basic watermarking scheme; many variations are possible. In its most basic implementation, a watermark consists of a sequence of real numbers $X = x_1, \ldots, x_n$. In practice, we create a watermark where each value x_i is chosen independently according to $N(0, 1)$ (where $N(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2). We assume that numbers are represented by a reasonable but finite precision and ignore these insignificant roundoff errors. Section 3.1 introduces notation

[1] More generally, n randomly chosen coefficients could be chosen from the $M, M \geq n$ most perceptually significant coefficients of the transform.

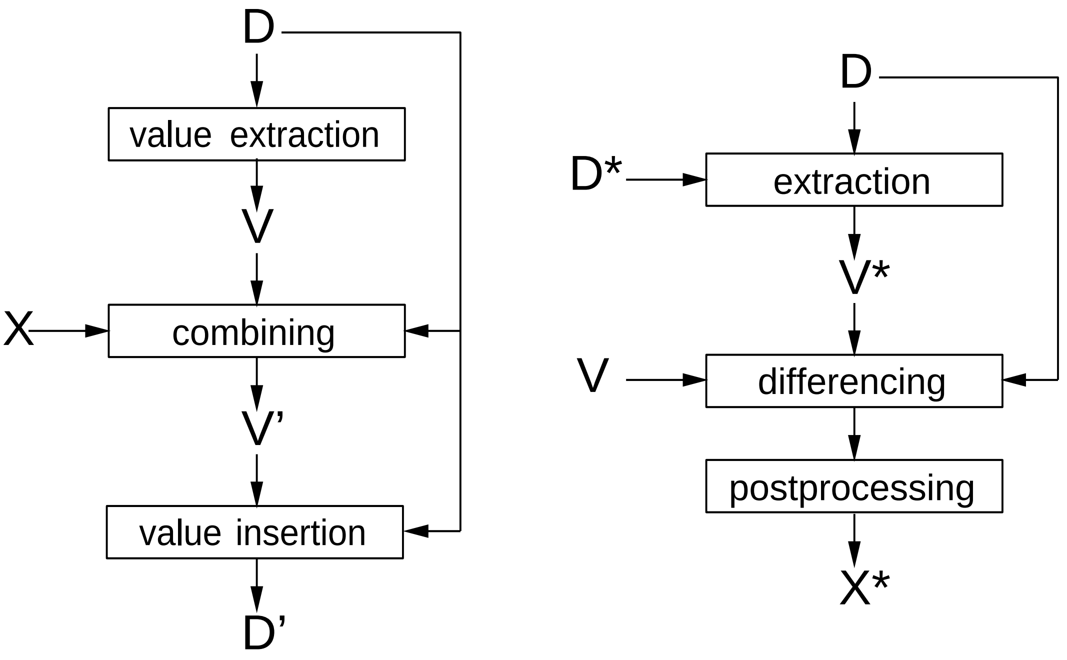

Figure 3: Encoding and decoding of the watermark string

to describe the insertion and extraction of a watermark and Section 3.3 describes how two watermarks (the original one and the recovered, possibly corrupted one) can be compared. This procedure exploits the fact that each component of the watermark is chosen from a normal distribution. Alternative distributions are possible, including choosing x_i uniformly from $\{1, -1\}, \{0, 1\}$ or $[0, 1]$. However, as we discuss in Section 3.5, using such distributions leaves one particularly vulnerable to attacks using multiple watermarked documents.

3.1 Description of the watermarking procedure

We extract from each document D a sequence of values $V = v_1, \ldots, v_n$, into which we insert a watermark $X = x_1, \ldots, x_n$ to obtain an adjusted sequence of values $V' = v'_1, \ldots, v'_n$. V' is then inserted back into the document in place of V to obtain a watermarked document D'. One or more attackers may then alter D', producing a new document D^*. Given D and D^*, a possibly corrupted watermark X^* is extracted and is compared to X for statistical significance. We extract X^* by first extracting a set of values $V^* = v^*_1, \ldots, v^*_n$ from D^* (using information about D) and then generating X^* from V^* and V.

Frequency-domain based methods for extracting V and V^* and inserting V' are given in Section 2. For the rest of this section we ignore the manipulations of the underlying documents.

3.2 Inserting and extracting the watermark

When we insert X into V to obtain V' we specify a scaling parameter α which determines the extent to which X alters V. Three natural formulae for computing V' are:

$$v'_i = v_i + \alpha x_i \tag{1}$$

$$v'_i = v_i(1 + \alpha x_i) \tag{2}$$

$$v'_i = v_i(e^{\alpha x_i}) \tag{3}$$

Equation 1 is always invertible, and Equations 2 and 3 are invertible if $v_i \neq 0$, which holds in all of our experiments. Given V^* we can therefore compute the inverse function to derive X^* from V^* and V.

Equation 1 may not be appropriate when the v_i values vary widely. If $v_i = 10^6$ then adding 100 may be insufficient for establishing a mark, but if $v_i = 10$ adding 100 will distort this value unacceptably. Insertion based on Equations 2 or 3 are more robust against such differences in scale. We note that Equations 2 and 3 give similar results when αx_i is small. Also, when v_i is positive then Equation 3 is equivalent to $\lg(v_i') = \lg(v_i) + \alpha x_i$, and may be viewed as an application of Equation 1 to the case where the logarithms of the original values are used.

3.2.1 Determining multiple scaling parameters

A single scaling parameter α may not be applicable for perturbing all of the values v_i, since different spectral components may exhibit more or less tolerance to modification. More generally one can have multiple scaling parameters $\alpha_1, \ldots, \alpha_n$ and use update rules such as $v_i' = v_i(1 + \alpha_i x_i)$. We can view α_i as a relative measure of how much one must alter v_i to alter the perceptual quality of the document. A large α_i means that one can perceptually "get away" with altering v_i by a large factor without degrading the document.

There remains the problem of selecting the multiple scaling values. In some cases, the choice of α_i may be based on some general assumption. For example, Equation 2 is a special case of the generalized Equation 1 ($v_i' = v_i + \alpha_i x_i$), for $\alpha_i = \alpha v_i$. Essentially, Equation 2 makes the reasonable assumption that a large value is less sensitive to additive alterations than a small value.

In general, one may have little idea of how sensitive the image is to various values. One way of empirically estimating these sensitivities is to determine the distortion caused by a number of attacks on the original image. For example, one might compute a degraded image D^* from D, extract the corresponding values $v_1^*, \ldots, v_n^*$ and choose α_i to be proportional to the deviation $|v_i^* - v_i|$. For greater robustness, one should try many forms of distortion and make α_i proportional to the average value of $|v_i^* - v_i|$. As alternatives to taking the average deviation one might also take the median or maximum deviation.

One may combine this empirical approach with general global assumptions about the sensitivity of the values. For example, one might require that $\alpha_i \geq \alpha_j$ whenever $v_i \geq v_j$. One way to combine this constraint with the empirical approach would be to set α_i according to

$$\alpha_i \sim \max_{j | v_j \leq v_i} |v_j^* - v_j|.$$

A still more sophisticated approach would be to weaken the monotonicity constraint to be robust against occasional outliers.

In all our experiments we simply use Equation 2 with a single parameter $\alpha = 0.1$. When we computed JPEG-based distortions of the original image we observed that the higher energy frequency components were not altered proportional to their magnitude (the implicit assumption of Equation 2). We suspect that we could make a less obtrusive mark of equal strength by attenuating our alterations of the high-energy components and amplifying our alterations of the lower-energy components. However, we have not yet performed this experiment.

3.3 Evaluating the similarity of watermarks

There are a number of ways that one can evaluate the similarity between two watermarks. A traditional correlation measure can be used, for example, or variants such as the t-distribution and Fisher's z transform. Below, we outline an alternative similarity metric, primarily to establish that a false positive judgement is highly unlikely, i.e. an innocent party is unlikely to be wrongly accused of copying.

It is highly unlikely that the extracted mark X^* will be identical to the original watermark X. Even the act of requantizing the watermarked document for delivery will cause X^* to deviate from X. We measure the similarity of X and X^* by

$$\mathsf{sim}(X, X^*) = \frac{X^* \cdot X}{\sqrt{X^* \cdot X^*}}. \tag{4}$$

We argue that large values of $\mathsf{sim}(X, X^*)$ are significant by the following analysis. Suppose that the creators of document D^* had no access to X (either through the seller or through a watermarked document). Then, even conditioned on any fixed value for X^*, each x_i will be independently distributed according to $N(0, 1)$. The distribution on $X^* \cdot X$ may be computed by first writing it as $\sum_{i=1}^n x_i^* x_i$, where x_i^* is a constant. Using the well-known formula for the distribution of a linear combination of variables that are independent and normally distributed, $X^* \cdot X$ will be distributed according to

$$N(0, \sum_{i=1}^{n} {x_i^*}^2) = N(0, X^* \cdot X^*)$$

Thus, $\mathsf{sim}(X, X^*)$ is distributed according to $N(0, 1)$. We can then apply the standard significance tests for the normal distribution. For example, if X^* is created independently from X then it is extremely unlikely that $\mathsf{sim}(X, X^*) > 6$. Note that slightly higher values of $\mathsf{sim}(X, X^*)$ may be required when a large number of watermarks are on file.

3.3.1 Robust statistics

The above analysis required only the independence of X from X^*, and did not rely on any specific properties of X^* itself. This fact gives us further flexibility when it comes to preprocessing X^*. We can process X^* in a number of ways to potentially enhance our ability to extract a watermark. For example, in our experiments on images we encountered instances where the average value of x_i^*, denoted $E_i(X^*)$, differed substantially from 0, due to the effects of a dithering procedure. While this artifact could be easily eliminated as part of the extraction process, it provides a motivation for postprocessing extracted watermarks. We found that the simple transformation $x_i^* \leftarrow x_i^* - E_i(X^*)$ yielded superior values of $\mathsf{sim}(X, X^*)$. The improved performance resulted from the decreased value of $X^* \cdot X^*$; the value of $X^* \cdot X$ was only slightly affected.

In our experiments we frequently observed that x_i^* could be greatly distorted for some values of i. One postprocessing option is to simply ignore such values, setting them to 0. That is,

$$x_i^* \leftarrow \begin{cases} x_i^* & \text{if } |x_i^*| > \text{tolerance} \\ 0 & \text{Otherwise} \end{cases}$$

Again, the goal of such a transformation is to lower $X^* \cdot X^*$. A less abrupt version of this approach is to normalize the X^* values to be either $-1, 0$ or 1, by

$$x_i^* \leftarrow \mathsf{sign}(x_i^* - E_i(X^*)).$$

This transformation can have a dramatic effect on the statistical significance of the result. Other robust statistical techniques could also be used to suppress outlier effects [10].

A natural question is whether such postprocessing steps run the risk of generating false positives. Indeed, the same potential risk occurs whenever there is any latitude in the procedure for extracting X^* from D^*. However, as long as the method for generating a set of values for X^* depends solely on D and D^*, our statistical significance calculation is unaffected. The only caveat to be considered is that the bound on the probability that one of $X_1^*, \ldots X_k^*$ generates a false positive is the sum of the individual bounds. Hence, to convince someone that a watermark is valid, it is necessary to have a published and rigid extraction and processing policy that is guaranteed to only generate a small number of candidate X^*.

3.4 Choosing the length, n, of the watermark

The choice of n dictates the degree to which the watermark is spread out among the relevant components of the image. In general, as the number of altered components are increased the extent to which they must be altered decreases. For a more quantitative assessment of this tradeoff, we consider watermarks of the form $v_i' = v_i + \alpha x_i$ and model a white noise attack by $v_i^* = v_i' + r_i$ where r_i are chosen according to independent normal distributions with standard deviation σ. For the watermarking procedure we described below one can recover the watermark when α is proportional to $\sigma/\sqrt{n}$. That is, by quadrupling the number of components used one can halve the magnitude of the watermark placed into each component. Note that the sum of squares of the deviations will be essentially unchanged.

However, when one increases the number of components used there is a point of diminishing returns at which the new components are randomized by trivial alterations in the image. Hence they will not be useful for storing watermark information. Thus the best choice of n is ultimately document-specific.

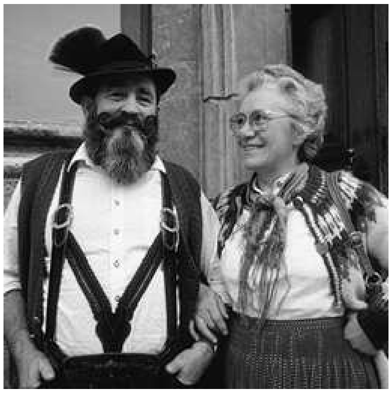

Figure 4: "Bavarian Couple" courtesy of Corel Stock Photo Library.

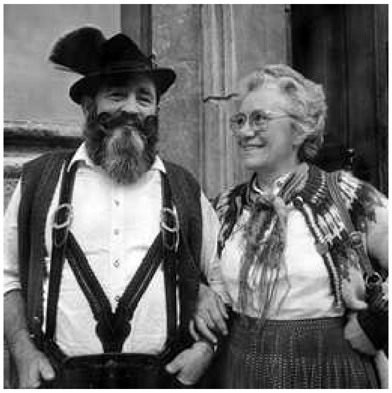

Figure 5: Watermarked version of "Bavarian Couple".

3.5 Resilience to multiple-document (collusion) attacks

The most general attack consists of using t multiple watermarked copies $D'_1, \ldots, D'_t$ of document D to produce an unwatermarked document D^*. We note that most schemes proposed seem quite vulnerable to such attacks. As a theoretical exception, Boneh and Shaw [3] propose a coding scheme for use in situations in which one can insert many relatively weak 0/1 watermarks into a document. They assume that if the ith watermark is the same for all t copies of the document then it cannot be detected, changed or removed. Using their coding scheme the number of weak watermarks to be inserted scales according to t^4, which may limit its usefulness in practice.

To illustrate the power of multiple-document attacks, consider watermarking schemes in which v'_i is generated by either adding 1 or -1 at random to v_i. Then as soon as one finds two documents with unequal values for v'_i one can determine v_i and hence completely eliminate this component of the watermark. With t documents one can, on average, eliminate all but a 2^{1-t} fraction of the components of the watermark. Note that this attack does not assume anything about the distribution on v_i. While a more intelligent allocation of ± 1 values to the watermarks (following [3, 14]) will better resist this simple attack, the discrete nature of the watermark components makes them much easier to completely eliminate. Our use of continuous valued watermarks appears to give greater resilience to such attacks. Interestingly, we have experimentally determined that if one chooses the x_i uniformly over some range, then one can remove the watermark using only 5 documents. We believe a gaussian distribution is somewhat stronger.

Using a probabilistic analysis, it can be shown that any attack on Guassian watermarks must make use of $\Omega(\sqrt{n/\ln n})$ watermarks in order to have any chance of destroying the watermark. (This is provided only that the original image to be protected comes from a Gaussian distribution and that the second moment of the deviation of the new image from the original is small compared to n.) Hence, Gaussian watermarks are better than uniform watermarks, particularly when n is large.

4 Experimental Results

In order to evaluate the proposed digital watermark, we first took the "Bavarian Couple"[2] image of Figure (4) and produced the watermarked version of Figure (5)

4.1 Experiment 1: Uniqueness of watermark

Figure (6) shows the response of the watermark detector to 1000 randomly generated watermarks of which only one matches the watermark present in Figure (5). The positive response due to the correct watermark is very much stronger that the response to incorrect watermarks, suggesting that the algorithm has very low false positives (and false negative) response rates.

[2] The common test image "Lenna" was originally used in our experiments and similar results were obtained. However, questions of taste aside, Playboy Inc. refused to grant copyright permission for electronic distribution.

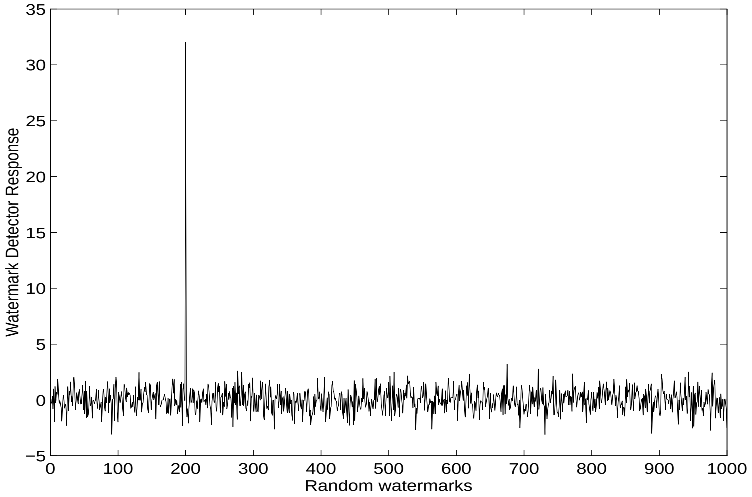

Figure 6: Watermark detector response to 1000 randomly generated watermarks. Only one watermark (the one to which the detector was set to respond) matches that present in Figure (5).

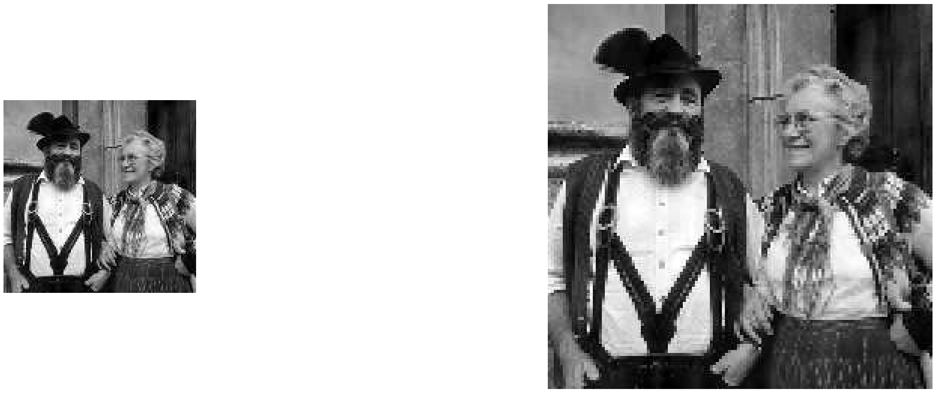

Figure 7: (a) Low pass filtered, 0.5 scaled image of "Bavarian Couple", (b) re-scaled image showing noticable loss of fine detail.

4.2 Experiment 2: Image Scaling

The watermarked image was scaled to half its orginal size, Figure (7a). In order to recover the watermark, the quarter-sized image was re-scaled to its original dimensions, as shown in Figure (7b), in which it is clear that considerable fine detail has been lost in the scaling process. This is to be expected since subsampling of the image requires a low pass spatial filtering operation. The response of the watermark detector to the original watermarked image of Figure (5) was 32.0 which compares to a response of 13.4 for the re-scaled version of Figure (7b). While the detector response is down by over 50%, the response is still well above random chance levels suggesting that the watermark is robust to geometric distortions. Moreover, it should be noted that 75% of the original data is missing from the scaled down image of Figure 7.

4.3 Experiment 3: JPEG coding distortion

Figure (8) shows a JPEG encoded version of "Bavarian Couple" with parameters of 10% quality and 0% smoothing, which results in clearly visible distortions of the image. The response of the watermark detector is 22.8, again suggesting that the algorithm is robust to common encoding distortions. Figure (9) shows a JPEG encoded version of "Bavarian Couple" with parameters of 5% quality and 0% smoothing, which

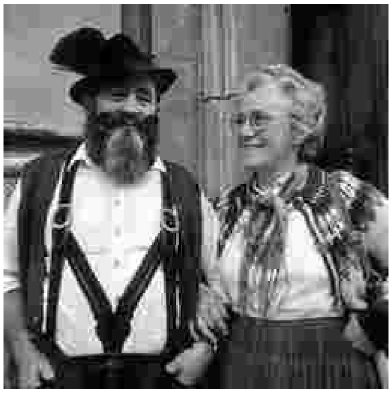

Figure 8: JPEG encoded version of “Bavarian Couple” with 10% quality and 0% smoothing.

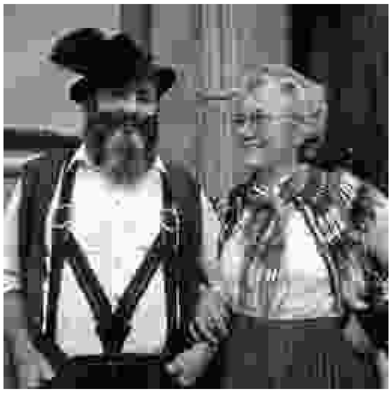

Figure 9: JPEG encoded version of “Bavarian Couple” with 5% quality and 0% smoothing.

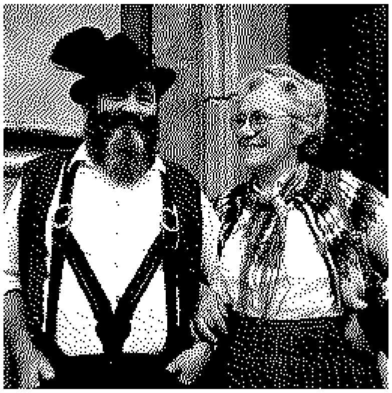

Figure 10: Dithered version of “Bavarian Couple”.

results is very significant distortions of the image. The response of the watermark detector in this case is 13.9, which is still well above random.

4.4 Experiment 4: Dithering Distortion

Figure (10) shows a dithered version of “Bavarian Couple”. The response of the watermark detector is 5.2 again suggesting that the algorithm is robust to common encoding distortions. In fact, more reliable detection can be achieved simply by removing any non-zero mean from the extracted watermark, as discussed in Section 3.3.1. In this case the detection value is 10.5.

4.5 Experiment 5: Clipping

Figure (11a) shows a clipped version of the watermarked image of Figure (5) in which only the central quarter of the image remains. In order to extract the watermark from this image, the missing portions of the image were replaced with portions from the original **unwatermarked** image of Figure (4), as shown in Figure (11b). In this case, the response of the watermark is 14.6. Once again, this is well above random even though 75% of the data has been removed.

Figure (12a) shows a clipped version of the JPEG encoded image of Figure (8) in which only the central quarter of the image remains. As before, the missing portions of the image were replaced with portions from the original **unwatermarked** image of Figure (4), as shown in Figure (12b). In this case, the response of the watermark is 10.6. Once more, this is well above random even though 75% of the data has been removed and distortion is present in the clipped portion of the image.

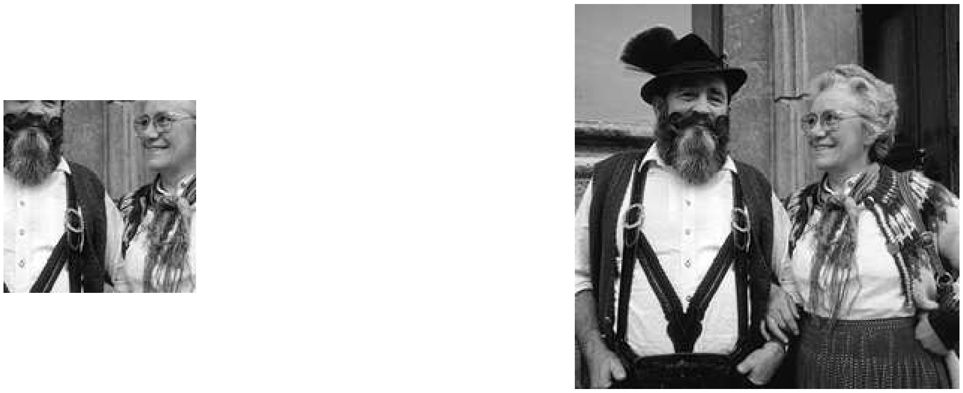

Figure 11: (a) Clipped version of watermarked “Bavarian Couple”, (b) Restored version of “Bavarian Couple” in which missing portions have been replaced with imagery from the original unwatermarked image of Figure (4).

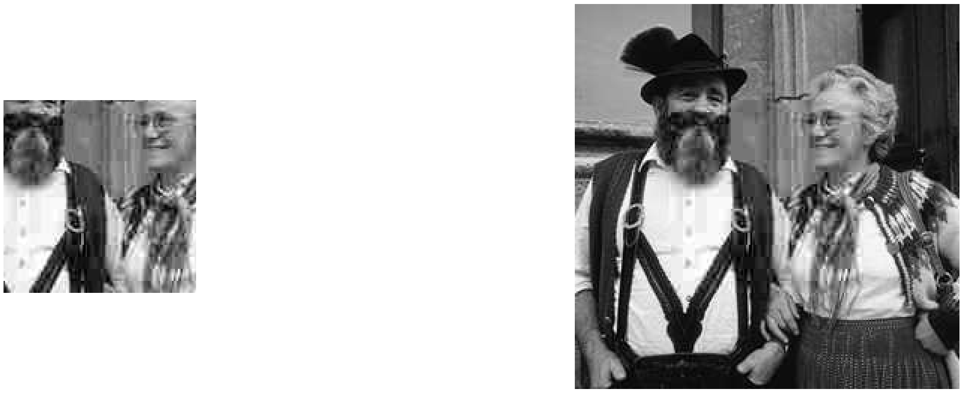

Figure 12: (a) Clipped version of JPEG encoded (10% quality, 0% smoothing) “Bavarian Couple”, (b) Restored version of “Bavarian Couple” in which missing portions have been replaced with imagery from the original unwatermarked image of Figure (4).

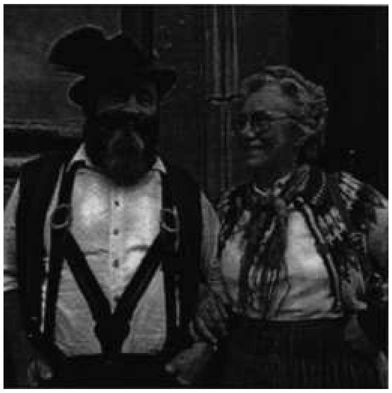

Figure 13: Printed, xeroxed, scanned and rescaled image of "Bavarian Couple".

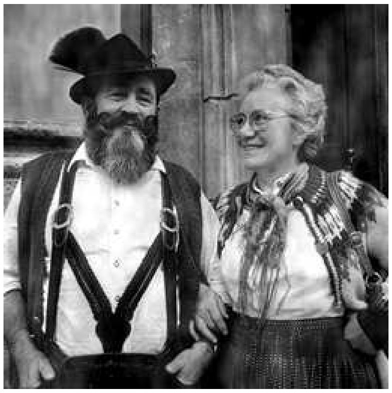

Figure 14: Image of "Bavarian Couple" after five successive watermarks have been added.

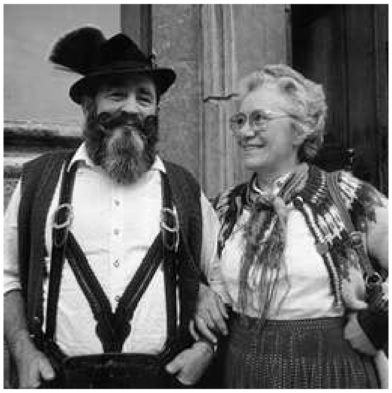

Figure 15: Image of "Bavarian Couple" after averaging together five independently watermarks versions of the "Bavarian Couple" image.

4.6 Experiment 6: Print, xerox and scan

Figure (13) shows an image of Lenna after (1) printing, (2) xeroxing, then (3) scanning at 300 dpi using UMAX PS-2400X scanner, and finally (4) rescaled to a size of 256 × 256. Clearly, this image suffers from several levels of distortion that accompany each of the four stages. High frequency pattern noise is especially noticeable. The detector response to the watermark is 4.0. However, if the non-zero mean is removed and only the sign of the elements of the watermark are used, then the detector response is 7.0, which is well above random.

4.7 Experiment 7: Attack by watermarking watermarked images

Figure (14) shows an image of "Bavarian Couple" after five successive watermarking operations, i.e. the original image is watermarked, the watermarked image is watermarked, etc. This may be considered another form of attack in which it is clear that significant image degradation eventually occurs as the process is repeated. This attack is equivalent to adding noise to the frequency bins containing the watermark. Interestingly, Figure (16) shows the response of the detector to 1000 randomly generated watermarks, which include the five watermarks present in the image. Five spikes clearly indicate the presence of the five watermarks and demonstrate that successive watermarking does not interfere with the process.

4.8 Experiment 8: Attack by collusion

In a similar experiment, we took five separately watermarked images and averaged them to form Figure (15) in order to simulate a simple collusion attack. As before, Figure (17) shows the response of the detector to 1000 randomly generated watermarks, which include the five watermarks present in the image. Once again,

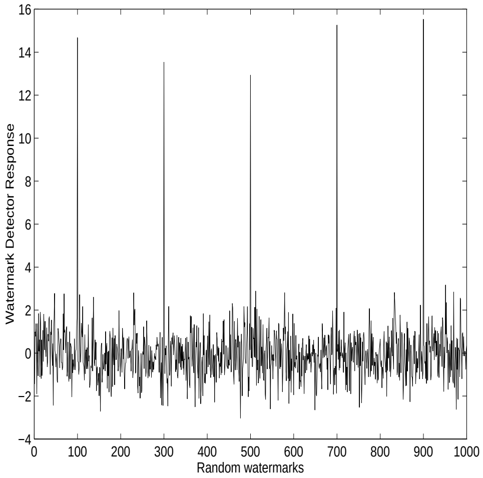

Figure 16: Watermark detector response to 1000 randomly generated watermarks (including the 5 specific watermarks) for the watermarked image of Figure (14). Each of the five watermarks is clearly indicated.

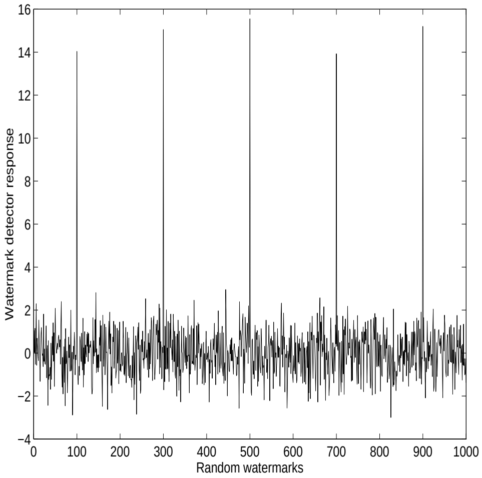

Figure 17: Watermark detector response to 1000 randomly generated watermarks (including the 5 specific watermarks) for the watermarked image of Figure (15). Each of the five watermarks is clearly detected, indicating that collusion by averaging is ineffective.

five spikes clearly indicate the presence of the five watermarks and demonstrate that simple collusion based on averaging a few images is ineffective.

5 Conclusion

A need for electronic watermarking is developing as electronic distribution of copyright material becomes more prevalent. Above, we outlined the necessary characteristics of such a watermark. These are: fidelity preservation, robustness to common signal and geometric processing operations, robustness to attack, and applicability to audio, image and video data.

To meet these requirements, we proposed a watermark whose structure consisted of 1000 randomly generated numbers with a Normal distribution having zero mean and unity variance. A binary watermark was rejected based on the fact that it is much less robust to attacks based on collusion of several independently watermarked copies of an image. The length of the watermark is variable and can be adjusted to suit the characteristics of the data. For example, longer watermarks might be used for an image that is especially sensitive to large modifications of its spectral coefficients, thus requiring weaker scaling factors for individual components.

The watermark is then placed in the perceptually *most* significant components of the image spectrum. This ensures that the watermark remains with the image even after common signal and geometric distortions. Modification of these spectral components results in severe image degradation long before the watermark itself is destroyed. Of course, to insert the watermark, it is necessary to alter these very same coefficients. However, each modification can be extremely small and, in a manner similar to spread spectrum communication, a strong narrowband watermark may be distributed over a much broader image (channel) spectrum. Conceptually, detection of the watermark then proceeds by adding all of these very small signals, and concentrating them once more into a signal with high signal-to-noise ratio. Because the magnitude of the watermark at each location is only known to the copyright holder, an attacker would have to add much more noise energy to each spectral coefficient in order to be sufficiently confident of removing the watermark. However, this process would destroy the image.

In our experiments, we added the watermark to the image by modifying 1000 of the more perceptually significant components of the image spectrum. More specifically, the 1000 largest coefficients of the DCT (excluding the DC term) were used. Further refinement of the method would identify perceptually significant components based on an analysis of the image and the human perceptual system and might also include additional considerations regarding the relative predictability of a frequency based on its neighbors. The latter property is important to consider in order to minimize any attack based on a statistical analysis of frequency spectra that attempts to replace components with their maximul likelihood estimate, for example. The choice of the DCT is not critical to the algorithm and other spectral transforms, including wavelet type decompositions are also possible. In fact, use of the FFT rather than DCT may prefereble from a computational perspective.

It was shown, using the "Bavarian Couple" image, that the algorithm can extract a reliable copy of the watermark from imagery that has been significantly degraded through several common geometric and signal processing procedures. These include, zooming (low pass filtering), cropping, lossy JPEG encoding, dithering, printing, photocopying and subsequent rescanning.

More experimental work needs to be performed to validate these results over a wide class of data. Application of the method to color images should be straightforward though robustness to certain color image processing procedures should be investigated. Similarly, the system should work well on text images, however, the binary nature of the image together with its much more structured spectral distribution need more work. Furthermore, application of the watermarking method to audio and video data should follow in a straightforward fashion, although, attention must be paid to the time varying nature of these data. A more sophisticated watermark verification process may also be possible using methods developed for spread spectrum communications.

Larger system issues must be also addressed in order for this system to be used in practice. For example, it would be useful to be able to prove in court that a watermark is present without publically revealing the original, unmarked document. This is not hard to accomplish using secure trusted hardware; an efficient purely cryptographic solution seems much more difficult. It should also be noted that current proposal only allows the watermark to be extracted by the owner, since the original unwatermarked image is needed as part of the extraction process. This prohibits potential users from querying the image for ownership and copyright information. This capability may be desirable but appears difficult to achieve with the same level of robustness. However, it is straightforward to provide if a much weaker level of protection is acceptable and might therefore be added as a secondary watermarking procedure. Finally, we note that while the proposed methodology is used to hide watermarks in data, the same process can be applied to sending other forms of message through media data.

Acknowledgements

The authors thank Larry O'Gorman of AT&T Bell Laboratories for bring this problem to our attention, Harold Stone for advice on image transforms and Sebastien Roy for testing the robustness of the watermark.

References

[1] E. H. Adelson. Digital signal encoding and decoding apparatus. Technical Report 4,939,515, United States Patent, 1990.

[2] W. Bender, D. Gruhl, and N. Morimoto. Techniques for data hiding. In *Proc. of SPIE*, volume 2420, page 40, February 1995.

[3] Dan Boneh and James Shaw. Collusion-secure fingerprinting for digital data. In *Advances in Cryptology: Proceedings, CRYPTO '95*. Springer-Verlag, 1995.

[4] J. Brassil, S. Low, N. Maxemchuk, and L. O'Gorman. Electronic marking and identification techniques to discourage document copying. In *Proc. of Infocom'94*, pages 1278–1287, 1994.

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

[6] I. J. Cox, S. Roy, and S. L. Hingorani. Dynamic histogram warping of images pairs for constant image brightness. In *IEEE Int, Conf. on Image Processing*, 1995.

[7] O. Faugeras. *Three Dimensional Computer Vision: A Geometric Viewpoint.* MIT Press, 1993.

[8] Allen Gersho and Robert Gray. *Vector Quantization and Signal Compression.* Kluwer Academic Publishers, Boston, 1992.

[9] R. C. Gonzalez and R. E. Woods. *Digital Image Processing.* Addison-Wesley, 1993.

[10] P. J. Huber. *Robust Statistics.* John Wiley and Sons, 1981.

[11] N. Jayant, J. Johnston, and R. Safranek. Signal compression based on models of human perception. *Proc IEEE*, 81(10), 1993.

[12] E. Koch, J. Rindfrey, and J. Zhao. Copyright protection for multimedia data. In *Proc. of the Int. Conf. on Digital Media and Electronic Publishing*, 1994.

[13] E. Koch and Z. Zhao. Towards robust and hidden image copyright labeling. In *Proceedings of 1995 IEEE Workshop on Nonlinear Signal and Image Processing*, June 1995.

[14] F. T. Leighton and S. Micali. Secret-key agreement without public-key cryptography. In *Proceedings of Crypto*, 1993.

[15] J.S Lim. *Two-Dimensional Signal Processing.* Prentice Hall, Englewood Cliffs, N.J., 1990.

[16] B. M. Macq and J-J Quisquater. Cryptology for digital tv broadcasting. *Proc. of the IEEE*, 83(6):944–957, 1995.

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

[18] 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.

[19] W. F. Schreiber, A. E. Lippman, E. H. Adelson, and A. N. Netravali. Receiver-compatible enhanced definition television system. Technical Report 5,010,405, United States Patent, 1991.

[20] 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.

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

[22] R. G. van Schyndel, A. Z. Tirkel, and C. F. Osborne. A digital watermark. In *Int. Conf. on Image Processing*, volume 2, pages 86–90. IEEE, 1994.