In the Proceedings of the 1998 IEEE International Conference on Image Processing.

CAPACITY ISSUES IN DIGITAL IMAGE WATERMARKING

Sergio D. Servetto[†*] Christine I. Podilchuk[§] Kannan Ramchandran[†]

http://www.ifp.uiuc.edu/~servetto/ chrisp@research.bell-labs.com kannan@ifp.uiuc.edu

[†]Beckman Institute, [*]Dept. of Computer Science,
Univ. of Illinois at Urbana-Champaign.
405 N. Matthews St., Urbana, IL, 61801.

[§]Multimedia Communications Research Dept.,
Bell Labs - Lucent Technologies.
600 Mountain Av., Murray Hill, NJ, 07974.

ABSTRACT

In this work we study certain aspects of the problem of watermarking images, not for the classical example of ownership identification, but instead for embedding a unique identifier that can act as a "serial number" for the image in which it is embedded. In this context, two fundamental questions are: (a) what is the maximum number of different watermarks that can be distinguished reliably? (b) what are good design techniques for watermarking methods to approximate that maximum? Under the assumption that attacks can be modeled as additive noise, we provide answers to both questions. To answer (a), we first show how the process of inserting a watermark is analogous to that of storing bits in certain devices, and then we compute the storage capacity of these devices. To answer (b), we present a specific design of a modulator to store information in those devices. Numerical simulations reveal that although strongly suboptimal in the number of bits they can effectively store in an image, simple and very low complexity modulator designs are able to pack enough bits in an image to be useful in practice.

1. INTRODUCTION

1.1. Problem Description

Popular communication systems like the Internet allow for the wide distribution of electronic data. Content providers are faced with the challenge of how to protect their electronic data. This problem has generated a flurry of recent research activity in the area of digital watermarking of electronic content for copyright protection. Unlike the traditional visible watermark found on paper, the challenge here is to introduce a digital watermark that does not alter the perceived quality of the electronic content, while being robust to attacks. For instance, in the case of image data, typical signal processing operations (such as linear and nonlinear filtering, cropping, rescaling, noise removal, lossy compression, etc.) should ideally be such that if they result in alteration or suppresion of the inserted watermark, then the resulting image must have been so severely degraded to render it worthless. Equally important to this robustness requirement, the watermark should not alter the perceived visual quality of the image. It is clear then that from a signal processing viewpoint, the two basic requirements for an effective watermarking scheme (i.e., robustness and transparency) conflict with each other.

Watermarking applications, at a coarse level, can be grouped into two main categories: *source-based* applications and *destination-based* applications.

- Source-based watermarks are typically used for purposes of ownership identification and tampering detection. A unique watermark signal is hidden in all copies of a particular image, prior to its distribution. Then, examination of the particular signal hidden in a given image can reveal who the originator of the image is, whether parts of the image have been tampered with (e.g., if the picture in a photo id has had the face replaced), etc. Furthermore, watermarks can be used to embed application-dependent information (not necessarily dealing with security issues), that can be kept even when the image is transferred accross different media such as disk, D1 tape, high-quality printouts, etc.
- Destination-based watermarks are typically used for tracing purposes. A distinct watermark signal that uniquely identifies a particular copy of the image is hidden in each copy, prior to its distribution, acting as a "serial number" for the image. Then, in case of detecting multiple unauthorized copies of a given image, retrieval of that serial number can identify the particular user whose image was used to create the illegal copies.

In this work, we are interested in *destination-based* watermarks.

1.2. Watermarks Based on Spread-Spectrum Techniques

Cox et al. [3] introduced the concept of "spread spectrum watermarking". They observed that for a watermark to be robust, it must be somehow inserted into the most perceptually significant image components: otherwise, it could be erased completely simply by suppressing from a given image its least perceptually significant components (e.g., using a good lossy coding algorithm), without altering the perceived image quality. Clearly, in order for the changes introduced to the perceptually significant components of an image to remain invisible, these changes must be small. However, small perturbations are very sensitive to noise.

Work performed while S. Servetto was a summer intern at Bell Labs, between 5/97–8/97

The approach taken in that work was conceived based on an analogy to the operation of secure communication systems based on the spread spectrum technique [1]. In such systems, an information bearing narrowband signal is converted into a wideband signal prior to transmission, by modulating the information waveform with a wideband noiselike waveform that is not known to the jammer. As a result of this bandwidth expansion, within any narrow spectral band, the total amount of energy from the information signal is small; however, by appropriately combining all these weak narrowband signals at the demodulator, the original information signal is recovered. Hence a jammer, unaware of the shape of the wideband carrier, is forced to spread its available power over a much larger bandwidth, thus reducing its effectiveness. The same idea is applied to insert a watermark. Many small (and secret) changes are introduced into the most significant image components, and because during the watermark extraction process (the receiver, in the secure comm system analogy) the location and value of these changes is known, it is possible to concentrate the information of all these small changes to come up with a robust decision on the presence/abscence of a watermark. Furthermore, to destroy such a watermark noise of high amplitude would be required in all these perceptually significant components, drastically reducing the perceived image quality too.

Some of the most effective watermarking methods developed so far are based on this idea.

1.3. Watermarks Based on Models of the Human Visual System

A significant amount of effort has been invested in understanding properties of the human visual system, in order to apply this knowledge in the development of solutions to image processing problems. Recently, visual models have been developed specifically for the performance evaluation of *lossy* image compression algorithms [8] (i.e., algorithms which degrade the original image quality in their reconstruction, in order to achieve higher compression ratios than would be otherwise possible). One common paradigm for perceptual coding is based on deriving an image dependent mask containing the *just noticeable differences* (JND), a set of thresholds used to compute perceptually-based quantizers. These models, originally built for the perceptual coding application, are ideally suited for watermarking: the JND thresholds provide upper bounds on watermark intensity levels. Hence, a criterion is available to address simultaneously the conflicting goals of robustness and transparency: a watermark can be made maximally strong, subject to the invisibility constraint (where invisibility is determined from the JND thresholds).

An effective watermarking technique based on these principles was presented in [4].

1.4. Main Contributions and Paper Organization

Our work focuses on destination-based watermarks, for tracing purposes. In this context, an attack that does not eliminate the watermark completely, but instead changes it just enough so that when the watermark is read a different identifier is recovered, is still considered successful. Hence, two most important issues that any such watermarking method must deal with are 1.- deciding how many different watermarks can be distinguished reliably, and 2.- how to design watermarking algorithms that can effectively achieve this maximum. The main contribution presented in this work is an answer to both questions in a simple and elegant manner. The problem of inserting watermarks is formulated as one storage of data in an imperfect device, where the imperfections are due to image processing operations that intentionally or unintentionally modify the watermark. In this framework, the number of different watermarks that can be reliably distinguished is found by computing the capacity of that device; and furthermore, this formulation motivates a specific method to effectively store watermarks in that device.

In terms of related work, we found two relevant references:

- Ó Ruanaidh et al [6] present an argument based on the calculation of the capacity of a suitably defined Gaussian channel, to motivate a particular scheme for encoding watermarks; however, they do not address the issue of how many watermarks can be reliably encoded. Furthermore, in their construction of the Gaussian channel, the noise variances involved were picked based on some sound intuition, but not based on a rigorous definition of "visibility", as given by perceptual image models.
- Smith and Comiskey [7] propose to consider the image to be watermarked as noise in a communication system, where the signal to transmit is the watermark. While their approach has the advantage of not requiring the original image to extract a watermark, their main disadvantage is that by not distinguishing between image data (that must be preserved) and impairments introduced by a jammer (that must be defeated), the number of different watermarks that can be reliably distinguished is significantly reduced.

The rest of this paper is organized as follows. In Section 2, we show how images can be modeled as storage devices (for watermarking purposes), and compute the storage capacity of such devices. In Section 3 we present the design of an algorithm to insert watermarks in an image; this algorithm is essentially a modulator for the storage devices. In Section 4, we report on extensive numerical simulations conducted to determine the actual number of bits that can be reliably stored using our modulator, to measure how much this number deviates from the theoretical limit. The paper concludes with a summary and a discussion on future work in Section 5.

2. IMAGES AS ARRAYS OF STORAGE DEVICES: CAPACITY ISSUES

2.1. Images as Array of Storage Devices

In [4] a technique for inserting watermarks is proposed, in which an image is broken into N components

$$C^{(k)} = [c_1^{(k)}, \ldots, c_n^{(k)}], \qquad k = 1 \ldots N$$

These components can be either blocks of DCT coefficients or sets of subband coefficients. For each component, they define a vector

$$\mathbf{p}^{(k)} = [p_1^{(k)}, \ldots, p_n^{(k)}], \qquad k = 1 \ldots N$$

of positive real numbers, where each $p_i^{(k)}$ denotes the maximum standard deviation of the noise that can be tolerated by $c_i^{(k)}$ and still remain perceptually invisible. The vectors $\mathbf{p}^{(k)}$ are computed based on models of the human visual system thoroughly studied in the context of perceptual coding. A watermark is inserted in an image by generating random vectors

$$W^{(k)} = [w_1^{(k)}, \ldots, w_n^{(k)}] \qquad k = 1 \ldots N$$

with mean zero and identity covariance matrix $\mathbf{I}$; the watermarked image is then defined to be that with components

$$M^{(k)} = [c_1^{(k)} + p_1^{(k)} w_1^{(k)}, \ldots, c_n^{(k)} + p_n^{(k)} w_n^{(k)}], \qquad k = 1 \ldots N$$

Given an arbitrary set of image components $\tilde{C}^{(k)}$, a watermark is retrieved simply by inverting the operations defining $M^{(k)}$:

$$\hat{W}^{(k)} = \left[\frac{\tilde{c}_1^{(k)} - c_1^{(k)}}{p_1^{(k)}}, \ldots, \frac{\tilde{c}_n^{(k)} - c_n^{(k)}}{p_n^{(k)}} \right], \qquad k = 1 \ldots N$$

Our interpretation of an image as representing an array of storage devices is based on the fact that the $C^{(k)}$'s play the role of "boxes" in which the random vectors $W^{(k)}$ are stored.

Consider now the process of retrieving a watermark. In general, either due to intentional attacks or due to normal image processing operations, $W^{(k)} \neq \hat{W}^{(k)}$; this is why we regard our storage devices as being "imperfect". And here is where we make our big modeling assumption: we postulate that a retrieved noise sequence is equal to the original, corrupted by additive noise. That is,

$$\hat{W}^{(k)} = W^{(k)} + J^{(k)}, \qquad k = 1 \ldots N$$

where $J^{(k)} = [j_1^{(k)} \ldots j_n^{(k)}]$ is a zero mean random vector with covariance matrix $\sigma\mathbf{I}$. This is how we arrive at the model shown in Fig. 1.

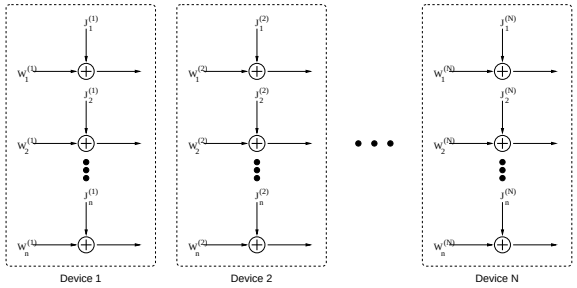

Figure 1: An image, regarded as an array of parallel additive noise channels.

2.2. Storage Capacity of the Array

Our first task was to decide how many different watermarks can be reliably stored in the array. Based on our model above, that problem corresponds to deciding what is the storage capacity of the array, as a function of the noise variance σ^2.

To perform this calculation, we invoke a basic result on optimal jamming strategies presented as a homework problem in [2]. Consider the single-letter channel of Fig. 2 (our array is a collection of these):

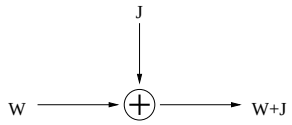

Figure 2: A Discrete Time Channel with Additive Noise.

In that problem, both signal and jammer are assumed of zero mean (i.e., $E(W) = E(J) = 0$), the signal power is constrained by $E(W^2) = P$, and the jammer power is constrained by $E(J^2) = N$; furthermore, it is assumed that W and J are independent. Hence, the channel capacity is given by the mutual information expression $\mathtt{I}(W; W + J)$, and of course is a function of the distributions of W and J. Now, given this setup, the jamming game is defined as one in which the jammer player chooses a distribution on J to minimize $\mathtt{I}(W; W + J)$, while the signal player chooses a distribution on W to maximice $\mathtt{I}(W; W + J)$. For a game so defined, letting $W^* \sim \mathcal{N}(0, P)$ and $J^* \sim \mathcal{N}(0, N)$, it is possible to show that W^* and J^* satisfy the saddlepoint conditions

$$\mathtt{I}(W; W + J^*) \leq \mathtt{I}(W^*; W^* + J^*) \leq \mathtt{I}(W^*; W^* + J)$$

and therefore that

$$\begin{aligned} \min_J \max_W \mathtt{I}(W; W + J) &= \max_W \min_J \mathtt{I}(W; W + J) \\ &= \frac{1}{2} \log \left(1 + \frac{P}{N} \right) \end{aligned}$$

and so the game has a value (the RHS of the last equation is the capacity of a power-constrained gaussian channel). In particular, it follows that a deviation from normality for either player worsens the mutual information from that player's point of view, thus establishing what the optimal transmitter and jammer should be.

This single-letter game formulation precisely answers our question on the capacity of the array:

- Mutual information is the natural cost function to maximize/minimize by the transmitter/jammer of the watermark signal: we are interested in computing the capacity of a certain channel, and channel capacity is defined in terms of $\mathtt{I}(\cdot;\cdot)$ [2].
- Playing this game independently on each single-letter channel in our array is the optimal thing to do: any correlations existing among different $w_i^{(k)}$ or among different $j_i^{(k)}$ could be exploited by the other player in order to increase/decrease mutual information to his advantage.

- From the saddlepoint conditions, it follows that $W^{(k)}$ should be $\mathcal{N}(\vec{0}, \mathbf{I})$, and $J^{(k)}$ should be $\mathcal{N}(\vec{0}, \sigma^2\mathbf{I})$.

This last point deserves special attention. In many applications, Gaussian distributions are used as an idealization of some unknown distribution, and deviations from the gaussian assumption typically result in a degradation of the performance of the algorithms designed for the gaussian case: a typical example is the approximation of a Minimum Mean-Square Error (MMSE) Estimator by its linear version, which in the gaussian case coincide. But we want to emphasize that this is not at all what happens in our case. For us, assuming the noise introduced by the jammer of the watermark signal is gaussian is the *worst*, most conservative assumption we could have made: deviations from gaussianity only help improve the performance of our detector. This is so because what we are trying to do is communicate the watermark signal reliably, even in the presence of jamming noise; but gaussian noise is the hardest noise to penetrate [2].

Finally we have that, as a function of the noise variance σ^2, the capacity of the array is given by:

$$C(\sigma) = \sum_{k=1}^{N}\sum_{i=1}^{n} \frac{1}{2}\log\left(1+\frac{1}{\sigma^2}\right) = \frac{nN}{2}\log\left(1+\frac{1}{\sigma^2}\right) \quad (1)$$

3. EFFECTIVE STORAGE OF INFORMATION IN THE ARRAY

After having established an upper bound on the number of bits that can be stored in the array, in this section we show how to design modulators to effectively do that.

In designing practical modulators, we are confronted with a difficulty: because of the nature of the problem, at the time the watermark is embedded in the image there is no knowledge available on the amount of noise that will be introduced in an attack; however, the storage capacity of the array is a function of that noise variance. Suppose our goal is to reliably distinguish $M = 2^b$ different watermarks: for that purpose, we need to be able to store b *information* bits in the array. Furthermore, suppose that the array consists of N components: there are N storage devices available, each of which has a fixed but unknown capacity (a function of the noise variance). The way in which we deal with this uncertainty is by taking a conservative approach: we store only *one* bit in each device, and then we study how the probability of decoding error is affected as σ^2 changes. The rationale for this is that by carefully designing modulators, we may be able to obtain systems in which, in order to bring the probability of decoding error of a watermark to unacceptable levels, the amount of noise that needs to be introduced is such that the original image is degraded to the point of it becoming completely worthless. And furthermore, since partitions as the ones described in [4] typically yield $N >> b$, we can make our design even more robust to attack, by mapping the b information bits to N channel bits using a good (M, N) code [2], where $M = 2^b$ is the number of distinct messages, and N is the block length of the code.

In order to store one bit in a device $C^{(k)}$ (as those shown in Fig. 1), we pick as our channel waveforms two gaussian vectors $\mathbf{s}^{(k),b} = [s_1^{(k),b} \ldots s_n^{(k),b}]$, with $\mathbf{s}^{(k),b} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $b = 0, 1$. A channel bit $b = 0, 1$ is modulated by storing $W^{(k)} = \mathbf{s}^{(k),b}$ into $C^{(k)}$. To demodulate, given an extracted sequence $\mathbf{s}$, that sequence is correlated against both $\mathbf{s}^{(k),0}$ and $\mathbf{s}^{(k),1}$, and a decision is made based on which correlation is highest. This is the optimal coherent detector for a known signal observed in iid gaussian noise [5].

Note the following important fact about this proposed modulator. The key of its robustness to attack lies in the fact that the jammer does not possess exact knowledge on the specific modulation waveforms used. Although for a random choice of waveforms there is a nonzero probability that both will lie close to each other (resulting in a device with high probability of bit error), this is unlikely to happen: by a straightforward application of the Strong Law of Large Numbers (SLLN), one can show that, for large n, $||\mathbf{s}^0 - \mathbf{s}^1||_2 \approx \sqrt{2n}$ (a.e.). Hence, if n is large enough, we can conclude that with high probability the distance separating our modulation waveforms will be large.

4. EXPERIMENTAL RESULTS

In this section we present some numerical simulations to illustrate the performance of our proposed watermarking technique. We use the standard test image Lena of size 512x512, and we brake it up into 4096 DCT blocks of size 8x8. We fix the number of information bits to store in the image to $b = 32$: if these many bits can be stored in the image reliably, then $2^{32} \approx 4 \cdot 10^9$ distinct watermarks can be distinguished reliably, a number useful in practice. These 32 data bits are mapped to 4096 channel bits using a simple (128,1,128) replication code.

In order to measure the robustness of our technique, we estimate the probability of watermark decoding error for different noise variances. We do this by taking a random sequence of 32 bits, storing them in the image, adding noise to the image, and retrieving the stored bits; we repeat this process 1000 times, and we estimate the probability of error as the ration of the number of incorrectly decoded watermarks to 1000. Also, to measure how image quality degrades due to noise attempting to jam the watermark, we compute the PSNR of the noisy image against the clean original, for the different noise variances. Fig. 3 shows the resulting curves.

We see in Fig. 3 that up to $\sigma = 3.5$, the number of incorrectly retrieved watermarks over 1000 tests is zero, and that for that value of σ, a typical corrupted image achieves a PSNR value of 28.81 dB only. In Fig. 4, we show an enlarged eection of the original image and of the noisy version.

Finally, and not surprisingly given the simplicity of our modulator, we find that for $\sigma = 3.5$, $C(\sigma) = 9970$ bits, significantly more than the 32 bits we can store reliably.

5. CONCLUSIONS

5.1. Summary

In this work we studied the problem of storing a watermark that can act as a serial number for the digital image in which it was inserted. We argued that in this context, a most important issue is to guarantee that a given watermark will not be mistaken for another one. For this purpose, we developed

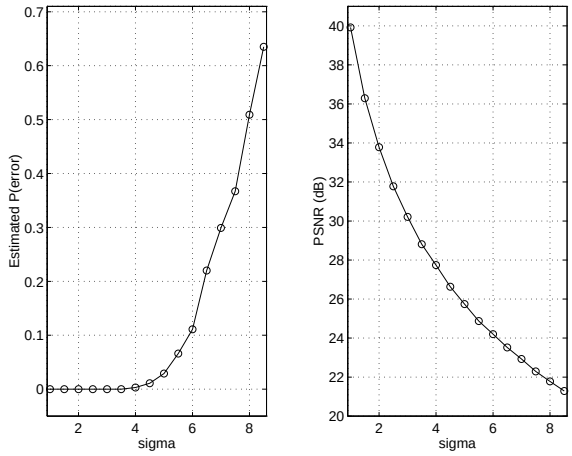

Figure 3: On the robustness of our watermarking technique: left, estimated probability of decoding error; right, PSNR numbers for different amounts of jamming noise.

a model of the watermarking process based on an analogy with a well studied and understood problem in communication theory. Using that intuition, we computed bounds on the performance achievable by watermarking schemes, and we developed a new technique. Experimental results showed that, although our simple modulator is far from being able to pack into an image a number of bits close to the maximum, at very low complexity it can pack enough bits to be useful in practice: by reliably distinguishing over 4 billion watermarks, it would be possible to give one copy of an image to each human being alive, and still be able to trace every single one of them!

5.2. Future Work

Further work is required on the following issues:

- In our experiments we used a very simple replication code. Using better codes (meaning, higher rate codes having the same minimum distance as the replication code), the number of bits that can be reliably stored in the array can be increased: this would come at the expense of added computational complexity, but that is a tradeoff worth exploring.
- We need to perform rigorous experiments to figure out what kind of distortions can be approximated by our additive noise model, and which ones cannot. For example, based on informal tests, we know that additive noise is *not* a good model for geometric distortions.
- Although we have shown our modulator to be very robust to additive noise attacks, the fact that it is based on a *coherent* detector is still a major weakness: further work is required on the use of robust noncoherent detectors in the demodulation process.
- For still images, it seems unlikely that significant performance improvements can be obtained over the methods proposed in [4] to break an image into basic components (DCT blocks, image subbands). However, a straightforward –frame by frame– application of these techniques to video yields extremely poor performance. Further work is required to identify how to construct a suitable storage array for video data.

Figure 4: Visual quality assessment: top, a section of the original Lena; bottom, a section of the jammed Lena. For this much degradation, over 1000 tests, all watermarks were correctly retrieved.

Our next step is to focus on the problem of how to define a storage array for video data.

6. REFERENCES

[1] R. Blahut. *Digital Transmission of Information*. Addison Wesley Publishing Company, 1990.

[2] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley and Sons, Inc., 1991.

[3] I. Cox, J. Killian, T. Leighton, and T. Shamoon. Secure Spread Spectrum Watermarking for Multimedia. Technical Report 95-10, NEC Research Institute, 1995.

[4] C. Podilchuk and W. Zeng. Image Adaptive Watermarking Using Visual Models. *IEEE Journal on Selected Areas in Communications*, 16(4):525–540, May 1998.

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

[6] J. Ó Ruanaidh, W. Dowling, and F. Boland. Watermarking Digital Images for Copyright Protection. *IEE Proceedings on Vision, Image and Signal Processing*, 143(4):250–256, August 1996.

[7] J. Smith and B. Comiskey. Modulation and Information Hiding in Images. In *Lecture Notes in Computer Science (1174)*. Springer-Verlag, 1996.

[8] A. Watson, G. Yang, J. Solomon, and J. Villasenor. Visibility of Wavelet Quantization Noise. *IEEE Transactions on Image Processing*, 6(8):1164–1175, August 1997.