Are Perfect Image Watermarking Schemes Possible?

C.E. Pizano, G.L. Heileman, C.T. Abdallah, and
Dept of EECE, The University of New Mexico, Albuquerque, NM 87131
cpizano@eece.unm.edu; heileman@eece.unm.edu; chaouki@eece.unm.edu;

M.S. Pattichis
Dept of EECE and Center for High Performance Computing, The University of New Mexico, Albuquerque, NM, 87131
pattichis@eece.unm.edu

***Abstract*—In this paper we propose mathematical performance measures by which the robustness, imperceptibility and information content of various watermarking methods could be measured. These measures rely in the notion of equivalence in the perceptual domain and some basic information-theoretic concepts. Based on these measures we show that the success of image watermarking heavily depend on the functional form of the human visual system perceptual mask. We also analyze the implication for robust watermarking of some commonly used perceptual masks.**

I. Introduction

The global Internet has greatly facilitated the legal and ilegal sharing of information, including digital images.The process of watermarking involves adding an information vector to a digital image in such a way that its ownership or copyright can be enforced. The added information, which must be visually imperceptible, is considered the watermark. The human visual system (HVS) has known characteristics that allow certain changes to the pixel values in an image to go undetected these phenomena are collectibly known as perceptual masking [6], [3]. An attack occurs when a third party modifies the image in order to remove the watermark [2], [4]. As long the image has not suffered too much degradation, the watermark must still be readable in the modified image. Most of the actual methods for watermarking are based on an embedding technique known as baseband pulse modulation. Our reseach has found matematical conditions that indicate when these methods have an inherent weakness against a class of simple additive random noise attacks, assuming that the attacker has the same knwoledge of perceptual mask.

II. The Digital Watermarking model

The classic setup for the watermarking problem is shown in figure 1. In this figure, I is the original image, modeled as the output of a discrete random source. The watermark W is a message from a set $\mathcal{W} = \{\infty, \ldots \mathcal{M}\}$ and K is a secret key. The first part of the process deals with the actual watermarking E of the image. Next, the watermarked version I_w goes through a noisy channel C, and is transformed into I_a. The channel models both intentional and unintentional attacks on the image. The

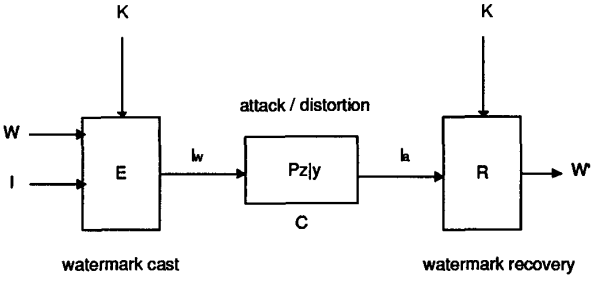

Fig. 1. Classical Watermarking Model.

basic assumption is that the secret key is unknown, therefore the noisy channel can be modeled by the probability distribution $P_{\mathcal{Z}|\mathcal{Y}}(I_a|I_w)$.

The third stage of the process is the receiver side: The function R attempts to recover an approximation of the watermark message W' using the attacked (or noisy) image I_a, and the knowledge of the secret key.

III. Image Space and Perceptual Sets

Digital image I is a vector in $\mathbb{Z}^{3\times d}$ dimensional space where $d = n \times m$ is the size of the image under consideration. We call two images of the same width and height and color space *compatible*. I is a vectorized representation of the more common used image representation as 2D array of pixels, being each pixel a triplet $< r, g, b >$ with $r, g, b \in \mathbb{Z}$.

A perceptual pseudometric φ takes in account the shortcomings of the human visual system, and defines a metric under which two slightly different images (in the MSE sense) I_1 and I_2 are perceptually equivalent (equal). To be precise, this property makes the perceptual metric a pseudo-metric:

$$I_1 \doteq I_2 \iff \varphi(I_1, I_2) = 0 \qquad (1)$$

There is no short-term hope of finding the true φ function, since the perceptibility of distortions in images is a subjective function of the eye and brain and varies from subject to subject, and changes according to how tests are performed.

1) The CPTL set: A standard test involves a subject looking at two images either in sequence or side-by-side for a brief period of time, the subject is directed to indicate when two images are perceptually different. In the image

0-7803-6290-X/00/$10.00 ©2000 IEEE

space, the CPTL defines a set of images (vectors in the space) that are all perceptually equivalent.

Let $\Pi \equiv CPTL(I_o)$ be the CPTL set of the original image. Now define $U(x)$, the vector that upper bounds the entire set as:

$$U(x) \equiv \max_{I \in \Pi} \{I(x)\}$$

and define $L(x)$, the vector that lower bounds the CPTL set as:

$$L(x) \equiv \min_{I \in \Pi} \{I(x)\}$$

These vectors can be constructed plotting all the images that belong to the CPTL set, and then taking the maximum and minimum at every point. Note also that these vectors each define two images that should also belongs to Π.

2) The AMO Set: Another important set is the set of images that are generated by processing of the original image in manners that are unrelated to watermarking. Mirroring, filtering and JPEG compression are some examples of this type of processing. The AMO set is defined as the set of images that are generated by minor modifications to the original image not including those required to insert a watermark. All the images are perceptually equivalent to the original.

3) The PTL Set: The CPTL set comes from side-by-side comparisons of images, but the original versus the modified image comparison is only available to the image creator. All other observers will not see both images side-by-side, and actually they only see I_w, the watermarked image. The PTL set is defined as the set of all perceptually equivalent images from memory recall. Therefore, in the case of a watermarked image, the level to which an image appears distorted is more subjective and more relaxed than the threshold for the image creator. This leads to a bigger set of equivalent images called the PTL set. The following relation is assumed to hold among these sets:

$$I \in AMO(I) \subseteq CPTL(I) \subseteq PTL(I) \qquad (2)$$

A typical interplay between these different sets is shown in figure 2 with I being the original image vector.

The PLT set is important because an attacker can tolerate more distortion than the image creator; and therefore, an attack can insert more distortion than the watermark process itself. The PTL set can be conveniently approximated as the CPTL set over I_w:

$$PTL(I) \approx CPTL(I_w) \qquad (3)$$

However, by approximating PTL in this manner, equation (2) may be violated.

IV. The Optimal Watermarking Problem

Given the set definitions, we can state the optimal watermarking problem as follows:

Given I and W, generate I_w such that the watermarked image lies in the $CPTL$ envelope but no further from I.

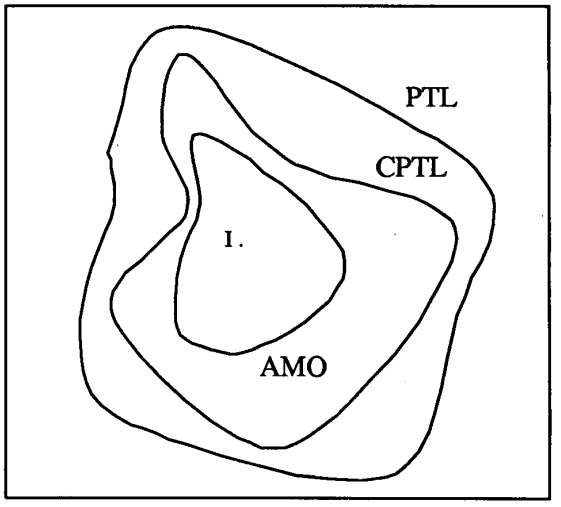

Fig. 2. The Different Perceptual Sets.

Such an image I_w is optimal, in the sense that it introduces as much distortion (power) as possible without making a big perceptual impact on the image (to the author's eye).

Maximum capacity comes by using the maximum power allowable given the channel characteristics. The maximum capacity of the channel must always be used even if the amount of information (i.e., watermark) is less than C. The extra capacity should always be used to embed error detection/correction information or to add redundancy to the watermark.

For the recovery process we have two possible pitfalls: false positives (type I errors) and false negatives (type II errors). False positives occur when the R process outputs a valid watermark message (non-zero) when the input was an unrelated image, or an image from the AMO set. False negatives occur when the decoder is presented with an image that has been watermarked and yet it fails to output the true watermark message or outputs zero (no watermark detected).

We will call a *reasonable* decoder process R one that can detect watermarks and also minimizes both the probability of false positives and false negatives.

V. The Optimal Watermark Attack Problem

From the point of view of the attacker, a optimization problem can be formulated as follows: Given a watermarked image I_w, limited knowledge about the E process, and access as a black box to the R process, design a process A that has the following properties

$$R(A(I_w)) = \{0\} \qquad (4)$$

and,

$$A(I_w) \in PTL(I) \qquad (5)$$

Recall that the best watermarks are those that are closest to the CPTL boundary, and that the only way to

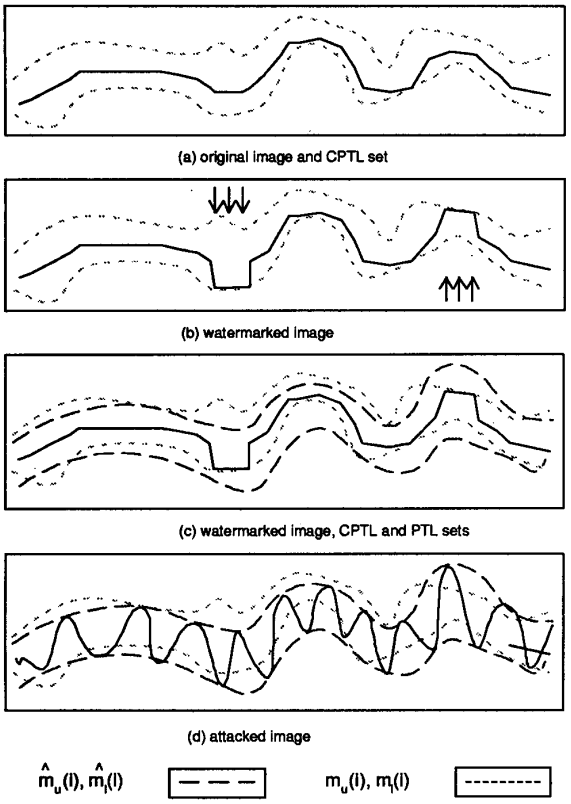

Fig. 3. An optimal attack.

generate them is by having a functional or algorithmic approximation to the CPTL set for a given source image. It is assumed that the attacker also has access to this function, and will use it to compute the PTL set. An intelligent attacker would then apply an attack with maximum power at each pixel as allowed by the PTL set. It can be argued that in this case the attacker will have a good chance of success.

Specifically, Shannon's channel coding theorem states that a binary symmetric memoryless channel having a symbol transition probability of 0.5 has zero capacity, and therefore no reliable communication over this channel is possible [5]. This means, if the attacker can flip half of the watermark bits, no matter what coding process is used, the watermark will be destroyed. Since the attacker does not know which pixels carry the watermark (they are assumed to depend on a secret key or on some property of the image), the best strategy for the attacker is to attack every pixel by inserting as much distortion as possible given the PTL bound. This insertion will be done as a random choice of either addition or substraction with probability 0.5.

If we consider an image a particular example, then the reason why this attack works become clearer by observing figure 3.

Figure 3(a) the hypothetical original image in black, and the CPTL set represented as an upper and lower bound for each pixel.

Figure 3(b) demonstrates how an optimal watermark is placed in the image in six different pixels as two groups of three pixels in a repetition code. Repetition codes are desirable because they lead to a low frequency watermark.

Figure 3(c) shows the attacker-computed PTL set superimposed on the watermarked image and the CPTL set. Note how the PTL set as approximated by equation (3) does not completely bound the CPTL set.

Part (d) of figure 3, shows an optimal attack that randomly pushes every pixel to one or the other side of the PTL envelope. In this particular example, it is easily seen that the watermark will be undetectable in the attacked image.

In the general case, whether or not watermark can be removed depends on the interplay between the CPTL and the PTL. Mathematically this can be stated as follows: Let i be a pixel belonging to the original image at location (x, y), $v(i)$ the value of the pixel at that location, $d(i)$ the amount of change that the pixel i suffered due to process E, $m_u(i)$ and $m_l(i)$ be the the watermark's method aproximation to the U and L evaluated at the location of pixel i.

Then, an optimal watermark will have:

$$v(i) + d(i) = m_u(i) \tag{6}$$

or,

$$v(i) + d(i) = m_l(i) \tag{7}$$

The choice depends on the details of the embedding process E. The decoder process R computes estimates of m_u and m_l based on the watermarked image denoted $\hat{m}_u$ and $\hat{m}_l$. There are two possible sets of these values depending if equation (6) or equation (7) was applied:

$$\hat{m}_u(i) = m_u(m_u(i)) \tag{8}$$

$$\hat{m}_l(i) = m_l(m_u(i)) \tag{9}$$

or,

$$\hat{m}_u(i) = m_u(m_l(i)) \tag{10}$$

$$\hat{m}_l(i) = m_l(m_l(i)) \tag{11}$$

The minimal conditions for an optimal watermark attack on pixel i are:

$$\hat{m}_u(i) \geq v(i) \tag{12}$$

and,

$$\hat{m}_l(i) \leq v(i) \tag{13}$$

This assures that the approximated PTL reaches the original unwatermarked value at pixel i, and therefore the attack will be greater than or equal to $d(i)$. An optimal attack is:

$$a(i) = \begin{cases} \hat{m}_u(i) & \text{with probability 0.5} \\ \hat{m}_l(i) & \text{with probability 0.5} \end{cases} \tag{14}$$

Therefore, with probability 0.5 the watermark information at pixel i will be undetectable. With the same

probability, the watermark information at pixel i will be enhanced. No *reasonable* decoder will recover *on average* meaningful watermark information from pixels that have the same value as the corresponding pixels in the original image.

VI. Analisis of Watermarking Methods

Equations (12) and (13) capture the essence of the optimal watermarking problem. Both of these equations depend on the shape of CPTL, and both conditions can be expressed in terms of the $m_u()$ and $m_l()$ functions:

$$m_u(m_l(i)) \geq v(i) \tag{15}$$

and,

$$m_l(m_u(i)) \leq v(i) \tag{16}$$

We previously showed that if these two conditions are met, there is no possibility of robust watermarking. The following sections analyse two different models of 'masking' based on the fundamentals of color and human vision.

4) The LSB Perceptual Mask: The simplest method of watermarking is least significant bit (LSB) coding . The basic idea is to embed the watermark information exclusively in the least significant bits of each pixel in the image.

The lower envelope for the masking function is given by:

$$m_l(i) = v(i) - mod(v(i), 2^{ls}) \tag{17}$$

and the upper envelope function is given by:

$$m_u(i) = v(i) - mod(v(i), 2^{ls}) + (2^{ls} - 1) \tag{18}$$

where ls is the number of least significant bits that are considered perceptually invisible. The value of ls is typically 2 or 3 in computer displays with 8 bits per channel.

Let us now test the robust watermarking inequalities stated in equations (15) and (16). The first condition can be simplified by noting that since $m_u(i) = m_l(i)+(2^{ls}-1)$;

$$m_l(m_l(i)) + (2^{ls} - 1) \geq v(i) \tag{19}$$

It is easy to show that $m_l(m_l(i)) = m_l(i)$. Expanding the definition of $m_l(i)$, equation (19) becomes:

$$v(i) - mod(v(i), 2^{ls}) + (2^{ls} - 1) \geq v(i), \tag{20}$$

which leads to the condition:

$$2^{ls} - 1 \geq mod(v(i), 2^{ls}) \tag{21}$$

Equation (21) is true for all possible values of $v(i)$. For the second condition, equation (16), we can write the following:

$$m_u(m_u(i)) - (2^{ls} - 1) \leq v(i), \tag{22}$$

which can be further simplified nothing that $m_u(m_u(i)) = m_u(i)$. This yields the condition:

$$mod(v(i), 2^{ls}) \geq 0, \tag{23}$$

which also holds for all possible values of $v(i)$.

Since we have shown that both conditions hold, we can conclude that it is impossible to create robust watermarks under the assumption that the perceptual threshold functions are the ones implied by the LSB watermarking method.

5) The Fixed Additive Perceptual Mask: The next simplest perceptual model is to consider a fixed additive (or subtractive) threshold α. The perceptual masking functions are:

$$m_u(i) = \begin{cases} i_{max} & \text{if } v(i) + \alpha > i_{max} \\ v(i) + \alpha & \text{otherwise} \end{cases} \tag{24}$$

$$m_l(i) = \begin{cases} 0 & \text{if } v(i) - \alpha < 0 \\ v(i) - \alpha & \text{otherwise} \end{cases} \tag{25}$$

Where i_{max} is the maximum pixel value. In 8-bit displays i_{max} is 255. The parameter α is the amount that the intensity of a pixel can be changed just before the change becomes perceptually noticeable.

The analysis for this perceptual masking function in terms of the conditions for robust watermarking can be split into two regions: a ***linear*** region where $\alpha < v(i) < i_{max} - \alpha$, and a ***nonlinear*** region where $v(i)$ is close to the maximum or minimum values.

For the ***linear*** region, the masking function is $v(i) \pm \alpha$, and it is trivial to prove that equations (15) and (16) are satisfied. For the ***nonlinear*** region, we can make the observation that the masking functions are equal to the LSB threshold functions of equations (17) and (18), if we let:

$$ls = \log(\alpha + 1)$$

And from the previous section we have proven that the LSB masking function hold the conditions (15) and (16) for any value of $v(i)$. Therefore, for any value of $v(i)$, the fixed additive perceptual mask satisfies equations (15) and (16), implying that robust watermarking with this masking function is impossible.

References

[1] David Aucsmith, editor. ***Information Hiding: Second International Workshop***, volume 1525 of ***Lecture Notes in Computer Science***, Portland, Oregon, U.S.A., 1998. Springer-Verlag, Berlin, Germany.

[2] Jean-Paul M. G. Linnartz and Marten van Dijk. Analysis of the sensitivity attack against electronic watermarks in images. In Aucsmith [1], pages 258–272.

[3] Arun N. Netravali and Birendra Prasada. Adaptative quantization of picture signals using spatial masking. ***Proceedings of the IEEE***, 65(4):536–542, April 1977.

[4] Fabien A. P. Petitcolas, Ross J. Anderson, and Markus G. Kuhn. Attacks on copyright marking systems. In Aucsmith [1], pages 218–238.

[5] C. E. Shannon. A mathematical theory of communication. ***Bell System Tech. Jorunal***, (27):379–423, 1948.

[6] Mitchell D. Swanson, Mei Kobayashi, and Ahmed H. Tewfik. Multimedia data-embedding and watermarking technologies. ***Proceedings of the IEEE***, 86(6):1064–1087, June 1998.