GEOMETRIC PROPERTIES OF WATERMARKING SCHEMES

Ee-Chien Chang *

DIMACS, Rutgers University
changec@cz3.nus.edu.sg

Michael Orchard *

Princeton University

ABSTRACT

A variety of image watermarking schemes have been proposed using orthogonal transformations, projections, and coding techniques to embed imperceptible watermarks into images. Analytical studies of the watermarking problem have typically been based on studying the performance limits of these known algorithms. In contrast, this paper formalizes the watermarking problem in an "algorithm independent" framework, representing any watermarking algorithm as a partition of the image space into a collection of sets, and defining requirements of these sets that must be met by any solution to the watermarking problem. Specifically, these requirements define and constrain the false-alarm ratio, distortion, robustness and security of a watermarking system. Using this formalism, we first characterize common features of algorithms that solve the watermarking problem. We show how the requirements defined earlier force important differences between watermarking signal sets and classical communication systems signal sets. Next, we show how common components of existing watermarking algorithms (e.g. transformations, projections, and coding) can be associated with specific requirements of the watermarking definition. Finally, we show how our new formalism of the watermarking problem provides a procedure for optimal design of watermarking systems to target specified false-alarm, distortion, and robustness objectives.

1. FORMULATION

This section offers a theoretical formulation of the problem of hiding a watermark in a collection of N coefficients. Although our discussion centers on images, these N coefficients can be an arbitrary representation of other multimedia source, for example, images, video, objects, etc.

A watermarking scheme has two operations–*engraving* and *detection*. It is convenient to represent these two operations as two subsets in the image space–the *Kernel* K and the set of watermarked images W. Given the original image I, engraving embeds a watermark into I, resulting in another image I_0, which is the image in K closest to I. The detection takes an image I' and declares it to be watermarked if and only if $I' \in W$. A typical example of W is the set of images within a predefined distance from the kernel.

We define the watermarking design problem as the design of the structure of the sets W and K in order to best meet watermarking objectives specified as constrains on *false alarm ratio*, *distortion*, *robustness* and *security*. The false-alarm ratio requires that a randomly selected image falls into W with a specified very small probability. Distortion reflects our desire to minimize the (perceptible) change in the original image caused by the process of engraving a watermark. Robustness requires that a random noise perturbation of the watermarked image (that is, $I_0 \in K$) should has a very small probability of finding a non-watermarked image (that is, an image in $\overline{W}$). Security requires that a smart attacker should have no easy way to find a non-watermarked image near any $I_0 \in K$.

We assume that the image space is a subset $U \subseteq [0,1]^N$ and the usual Euclidean 2-norm $\|\cdot\|$ is the visual distance between images. In the definitions that follow, $|A|$ is the size of the set A and $S_r(I)$ is the sphere of radius r around image I.

1. A watermarking scheme satisfies **False alarm ratio** F if
$$F \geq \frac{|W|}{|U|}.$$
2. A watermarking scheme satisfies **Distortion** ϵ if for any $I \in U$, there exists a $J \in K$ such that
$$\epsilon \geq \|J - I\|.$$
3. A watermarking scheme satisfies **Robustness** r if for any $I \in K$,
$$\frac{|S_r(I) \cap \overline{W}|}{|S_r(I)|} < N^{-1} \tag{1}$$
4. A scheme is s-**Secure** if any attacker can not efficiently find an $I' \in \overline{W} \cap S_s(I)$, for most $I \in K$.

*Work in this paper was partially done at NEC Research Institute.

We can note a natural correspondence between the first two requirements, false alarm ratio and distortion, and the rate and distortion of the uniform source coding problem. In fact, if we consider only these two requirements, good source coding lattices would also represent good candidate for K, permitting low distortion with a small $|W|$. However, for this lattice structure, robustness requires that each lattice point be surrounded by a sphere of watermarked images, and, unless $r \ll \epsilon$, the resulting size $|W|$ fails to meet a reasonable false alarm requirement. Since $r > \epsilon$ in most practical applications (i.e. the attacker is always willing to distort the image more than the engraver) classical coding lattices *do not* give viable watermarking solutions. Specifically, we prove that for a practical watermarking scheme, W cannot not be the disjointed union of spheres where each image in K contributes to the center of a sphere.

Beside providing insights into what structures are not promising watermarking solutions, the above requirements constructively restrict the kernel K and W in ways that hint at their required geometric structure. The false-alarm ratio dictates that the number of watermarked images is a small fraction of the size of the image space. These watermarked images must be highly clustered in order to satisfy the second and third conditions. First, recognize that all watermarked images surround kernel points. But, the discussion of the previous paragraph shows that, given a distortion constraint, even the best packed distribution of kernel points surrounded by isolated spheres of W results in $|W|$ exceeding the false alarm ratio. Thus, for a fixed distortion limit and reasonable $r > \epsilon$, the first and third conditions can only be satisfied if the spheres surrounding kernel points share elements - that is, if kernel points, as well as their corresponding spheres, are clustered. On the other hand, the distortion condition forces the clusters of watermarked images to be distributed across the image space. The condition on security requires the orientation of the boundary of W to be kept secret from the attackers. In the next section, we give a specific form of W and K that satisfies these requirements. However, we conjecture that all reasonable watermarking schemes have similar geometric form locally, if not globally.

2. REALIZATION

Several known watermarking schemes (for eg. [1, 2]) use a secret transformation, projection, and coding in various ways. A typical detection operation consists of the following steps. The image is first converted to some transform code. The detector then applies a secret orthogonal transformation on the transform code, removes a large number of coefficients, and decodes the message encoded by the remaining few coefficients. We focus on the last three steps, since the actual transform code in use is not an issue here. Geometrically speaking, the last three steps can be viewed in the following way: the set of watermarked images W and kernel K, subjected to the transformation T (not to be confused with the transform for the transform code), is of the form

$$\begin{aligned} T(W) &= \mathbb{R}^{n-k} \times C, \text{ and} \\ T(K) &= \mathbb{R}^{n-k} \times c, \end{aligned} \tag{2}$$

where C is a collection of k-dimensional spheres, and c is the set of center of the spheres in C. Different schemes vary the choices of C, T and k. Here are two examples supporting this generalization.

1. [1] T is a sparse matrix where each column has at most one nonzero entry. The coding C is a collection of cubes, which corresponds to an error-correcting code on the quantized coefficients.
2. [2] The projection and T are realized by a 1 by n matrix containing equal number of 1s and -1s. Thus, it projects onto a 1-dimensional line, and C is an interval of that line.

Besides its generality, (2) associates the three mentioned operations to the watermarking requirements: the secrete transformation provides security, the projection strengthens robustness, and an efficient coding gives low false alarm ratio and distortion. These relationships are further demonstrated in the next two subsections.

2.1. Projection and coding

In this section we examine watermarking schemes in which W and K are formed by projections as given by Eq. (2). It is instructive to visualize a scheme for $n = 3$ and $k = 2$. The set c is a set of points in a two-dimensional plane, the code are discs centered at c and of radius at most ϵ, the kernel are lines through c and orthogonal to the plane, and the watermarked images are 3-dimensional tubes.

With a fixed robustness and k, the selection of C amounts to finding a set of points c in the k-dimensional space such that as few points as possible in that space are far from points in c. This falls into the framework of quantization which has been extensively studied. If we take c as the lattice $A_k^\star$ scaled by δ (see [3]), then (F, ϵ, r) and (t, δ, k) are related as follows:

$$\begin{aligned} \epsilon &\approx \delta\sqrt{\frac{k}{12}} \\ F &\approx V_k \left\{\frac{t}{\delta}\right\}^k, \\ r &\approx \begin{pmatrix} n/2 \\ k/2 \end{pmatrix}^{1/k} t, \end{aligned} \tag{3}$$

where V_k is the volume of a unit k-dimensional sphere. With t fixed, as k increases, robustness decreases but requirements on C relax and a lower false alarm ratio and smaller distortion are achievable. This observation in general holds for any lattices. In our choice of lattice, if $n = 10000$, $\epsilon = 1$, and $r = 20$, then taking $k = 18$ attains false alarm ratio $F \approx 0.00015$ which is near-optimal among all possible k's and δ's.

2.2. Secure schemes and T

The secret transformation T plays a significant role in preventing tampering from smart attackers. Various scenarios and permitted attackers' abilities determine the design of T, which in turn distort the image space in a different ways. A typical attacker tries to determine the local gradient of T by making queries to the detector or consulting a historical list of watermarked images. Using the local gradient, the attacker could easily compute a non-watermarked image by perturbing the watermarked image along that direction.

Although linear transformations are commonly used, the flat linear boundary separating W and $\overline{W}$ is open to a number of efficient attacks[4]. These attackers rely on the assumption that they have access to images on the boundary of W, and the local gradient on such images correlate with the local gradient of the original watermarked image. Introducing nonlinearity could prevent such attempts. On the other hand, linearity is desirable because we want to preserve the local geometry of W and K. Thus, we need a nonlinear T that are smooth but sufficiently distorted so as to avoid attacks.

We give a guideline for selecting such nonlinear T. This guideline is based on the *critical distance*, which is the smallest radius such that the sphere around any image $I \in K$ contains at least half non-watermarked images.

($\star$) The local gradient of T at any image $I \in K$ should not correlate to the local gradient at most images beyond the *critical distance*.

Geometrically, the above criterion enforces nonlinearity in the boundary of W, and gives a guideline on the scale of distortion.

For the form (2), the critical distance $\alpha \approx 2^k r$. Thus, a smaller k weakens protection from smart attackers. It is interesting to recall that a smaller k strengthens protection from random noise.

Example.

The criterion ($\star$) can be achieved by choosing a set of linear transformations obtained from some pseudo-random number generator, together with a set of control points in the image space. Each control point is associated with a linear transformation. The actual secret transformation is an interpolation of these transformations. A candidate of the control points is a lattice whose minimal distance is chosen such that ($\star$) is achieved.

The high dimensionality of the image space induces high complexity in computing the interpolation, and thus the above method is not practical. One direct way to reduce computational complexity is by reducing the dimension. For example, we can use the one-dimensional lattice

$$\langle \dots, -p, 0, p, 2p, 3p, \dots \rangle$$

where $p = [\beta, 0, 0, \dots, 0] \in \mathbb{R}^n$, as the control points. The value β is chosen such that most images in the sphere $S_\alpha(0)$ is contained in $[-\beta/2, \beta/2] \times \mathbb{R}^{n-1}$, where α is the critical distance. Each control point ip is associated with a linear transformation T_i. During the encoding process for an image I, the transformation T to be applied on I is determined by its first coefficient, say a_1. If $i\beta < a_1 \leq (i+1)\beta$, then

$$T = \left(1 - \beta^{-1}(a_1 - i\beta)\right) T_i + \beta^{-1}(a_1 - i\beta) T_{i+1}.$$

3. REMARK ON FALSE ALARM RATIO

It is well-known that we can treats watermarking problem as channel coding with side information [5, 6], where the power, noise and rate correspond to our notions of distortion, robustness and false alarm respectively. By imposing a distribution on the image space and introducing average distortion and robustness to our formulations, it seems that our first three requirements are equivalent to channel coding with side information. However, there is a subtle difference in the notion of rate and false alarm ratio.

Given a scheme that achieves a rate R, we can modify it to one that achieves false alarm ratio 2^{-nR}, where n is the number of coefficients. This is done by designating a particular message as watermark. It is interesting to know whether the converse is possible, that is, for any scheme that achieves false alarm ratio $2^{-nR'}$, whether we can find another scheme that achieves rate R'. We now show that this is not possible.

Consider the setting studied by Costa in [5] where the source is Gaussian and the attack is additive white Gaussian noise. Costa gives the capacity of the channel and thus bounds the achievable rate. For some requirements on power (distortion) and noise (robustness), we can find a scheme in the form of (2), but gives a better false alarm ratio than the channel capacity suggested. Specifically, we look at signal and noise that are drawn i.i.d. from $N(0, 1)$ and $N(0, r)$ respectively, and the watermark satisfies power constraint ϵ where $r \geq \epsilon$ and $r \geq 1$. By [5], the capacity is

$$\kappa = (1/2) \log(1 + \epsilon/r).$$

Now, take a scheme of (2) with $k = 1$, $c = \{-\sqrt{n\epsilon}, \sqrt{n\epsilon}\}$

and

$$\begin{aligned} C &= \{[-\sqrt{n\epsilon} - n^{\frac{1}{4}}\sqrt{r}, -\sqrt{n\epsilon} + n^{\frac{1}{4}}\sqrt{r}], \\ &\quad\ [\ \ \sqrt{n\epsilon} - n^{\frac{1}{4}}\sqrt{r}, \ \ \sqrt{n\epsilon} + n^{\frac{1}{4}}\sqrt{r}]\}, \end{aligned}$$

where n is the dimension of the image space. The expected distortion is $\sqrt{n\epsilon}$ and the average robustness is better than $\sqrt{nr}$. Moreover, the false alarm ratio is

$$F < An^{\frac{1}{4}}\sqrt{r}\exp\left(-n\epsilon/2\right),$$

where A is some constant. Since $r \geq 1$ and $r \geq \epsilon$, we have $F < 2^{-n\kappa}$ for sufficiently large n. Therefore, the constraint on false alarm is weaker than that on rate.

4. SUMMARY AND CONCLUSIONS

We have proposed an algorithm-independent characterization of solutions to the watermarking problem. We have shown that the requirement of robustness disqualifies the use of classical coding lattices for watermarking kernels. Instead, we have shown that watermarking requirements constrain kernel elements to be clustered together in image space. Based on these geometric insights, and motivated by known techniques, we have given a general form for watermarking schemes. Beside finding that known methods fit into this general form, we conjecture that all schemes have this form locally, if not globally. We then associate features of the general structure with various components in typical detection and engraving routines. Using this association, we discuss how parameters of practical watermarking algorithms can be designed to best meet specific requirements.

Acknowledgments

The first author wishes to thank Mohan Kankanhalli for the helpful discussions on watermarking.

5. REFERENCES

[1] B. Chen and G.W. Wornell, "Achievable performance of digital watermarking systems," *IEEE Int. Conf. on Multimedia Computing & Systems*, vol. 1, pp. 13–18, 1999.

[2] I. Pitas, "A method for watermark casting in digital images," *IEEE Trans. on Circuits and Systems on Video Technology*, vol. 8, no. 6, pp. 775–780, 1998.

[3] J.H. Conway and N.J.A. Sloane, *Sphere Packings, Lattices and Groups*, Springer, 1998.

[4] I.J. Cox and J-.P. Linmartz, "Public watermarks and resistance to tampering," *IEEE Int. Conf. on Image Processing*, vol. 3, no. 0_3–0_6, 1997.

[5] M. Costa, "Writing on dirty paper," *IEEE Trans. on Information Theory*, vol. 29, no. 3, pp. 439–441, 1983.

[6] J. Chou, S.S. Pradhan, and K. Ramchandran, "On the duality between distributed source coding and data hiding," *33rd Asilomar conference on Signals, System and Computers*, pp. 1503–1507, 1999.