Hiding Functions and Computational Security of Image Watermarking Systems *

Nicholas Tran
Department of Mathematics & Computer Science
Santa Clara University
Santa Clara, CA 95053-0290, USA
ntran@math.scu.edu

Abstract

We introduce a complexity-theoretic model for studying computational security of binary image watermarking systems. Our model restricts algorithms used by the sender and the attacker to the class $\mathcal{H}$ of hiding functions. These are efficiently computable functions that preserve visual fidelity of the input image. Security of watermarking systems is to be established with complexity results about hiding functions. We also survey current theories of vision and propose an automata-theoretic model for visual fidelity called c-similarity. Finally we propose a candidate for $\mathcal{H}$ based on c-similarity and show that it is robust and contains infinitely many functions computable in polynomial time.

1 Introduction

The ease in duplicating, manipulating, and transmitting information on the World Wide Web has made protection of digital documents an urgent research topic. The emerging field of digital watermarking seeks to invent methods of embedding into a digital document some unique information, called a watermark, so that it is inconspicuous and yet difficult to fake or remove. The watermark can then be used for a variety of purposes, including copyright protection, fingerprinting, access control, advertising, annotations and captioning. Watermarking systems exist for all types of media (including text), but the majority of them deal with still images [1, 2, 3, 4, 5]. A good reference on digital watermarking and the more general field of steganography (information hiding) can be found in [6].

Security of a watermarking system is measured in terms of the difficulty in destroying its watermarks. Current literature often refers to this measure as the robustness of a watermarking system to distinguish it from the related notion of security of a steganographic system. The latter is measured in terms of the difficulty in detecting the existence of covert communication. Destruction of its watermarks is not the only way to defeat a watermarking system; it is also necessary to impose restrictions on the structure of watermarks to resist interpretation attacks, which seek to confuse true ownership by embedding an equally valid watermark [7].

At present evaluations of security of image watermarking systems are performed via empirical testing. General-purpose testing tools such as StirMark and UnZign are quite effective in disabling most systems; they do so by applying minor random geometric distortions to images, thereby distorting embedded watermarks beyond recognition. Benchmarks based on these tools have been proposed in [8, 9] to allow standardized evaluations of different systems.

Three theoretical models for studying *perfect security* of steganographic systems have been introduced [10, 11, 12]. Following Shannon's work on secrecy of cryptosystems [13], these models define perfect secrecy of steganographic systems in information-theoretic terms, e.g. as the total lack of relative entropy between the covertext and stegotext distributions, or of mutual information between the stegotext and the watermark. None of these models puts any restriction on the computational power of the players (conventionally named as Alice the sender, Bob the receiver, and Wendy the attacker). More importantly, none of the models adequately addresses the issue of modeling vi-

*This research was supported in part by NSF Grant EPS-9874732 with matching support from the state of Kansas, and by AFOSR Grant F49620-00-1-03 with matching support from the Kansas Technology Enterprise Corporation.

Proceedings of the 15th IEEE Computer Security Foundations Workshop (CSFW'02)
1063-6900/02 $17.00 © 2002 IEEE

sual fidelity, which is a fundamental constraint on the set of allowable algorithms by the sender as well as the attacker. It is either ignored or modeled using simple distortion metrics such as Mean Squared Error or Signal-to-Noise Ratio.

Such distortion metrics are known to have poor correlation to human visual perception. More sophisticated metrics based on models of human spatial vision have been proposed [14, 15, 16]. A recent literature review of the state of the art in distortion metrics can be found in [17]. These metrics are well-suited for engineering calculations but are too specific to be used in a *qualitative* model of visual fidelity.

2 Computational Security

In this paper we propose a complexity-theoretic model for studying *computational security* of binary image watermarking systems. Unlike previous work, our model reflects two practical constraints on algorithms for embedding and destroying watermarks, namely:

1. they must be computable with a reasonable amount of resources, and
2. they must preserve visual fidelity of the input images.

A well-accepted formal model for efficient computation is the class PF of polynomial-time Turing computable functions [18]. Our model requires that operations performed by both the sender and the attacker be restricted to a natural subclass $\mathcal{H}$ of PF which contains *hiding functions*, i.e. those preserving visual fidelity of their input binary images. A system is defined to be computationally secure if its embedded watermarks are detectable/extractable even after distortion by any hiding function. Security of watermarking systems is to be established with complexity results about hiding functions.

We explore variations of our definition to accommodate various types of existing watermarking systems: whether the watermark to be embedded is automatically generated or chosen arbitrarily; whether the original image and/or the watermark is available to the receiver; and whether decoding means detecting or extracting the watermark. We identify some constraints implied by the definition on the encoding function of public watermarking systems which i) generate watermarks automatically; ii) provide neither the original image nor the watermark to the receiver; and iii) require extraction of the watermark.

Next we briefly review current theories of vision and identify three approaches to modeling hiding functions: with automata, with frequency-domain transforms, or with sets of lines and edges. We also propose a candidate for $\mathcal{H}$. To measure similarity between binary images, we use a generalized version of the two-dimensional finite automaton, introduced by Blum and Hewitt [19] and studied extensively in the context of computer vision by Rosenfeld and others [20, 21]. Two binary images are said to be *c-similar* if they are found indistinguishable by these automata. As evidence for suitability of our choice of model, we show that these machines can detect qualitative differences in contrast, parity, handedness, and proportions. We also show that testing for c-similarity can be performed efficiently in polynomial time.

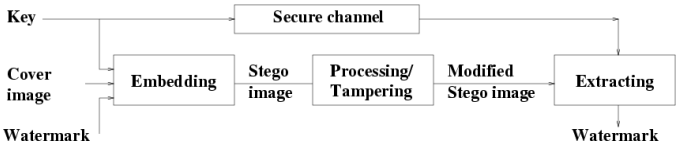

Figure 1. Model of a binary image watermarking system.

The class $\mathcal{H}$ of *hiding functions* can now be defined as functions whose outputs are c-similar to their inputs. We show that $\mathcal{H}$ contains an infinite number of nontrivial functions computable in polynomial time. We also show that $\mathcal{H}$ is closed under composition and polynomial-time inverse.

The rest of this paper is organized as follows. Section 3 describes our model of binary image watermarking systems and defines the notion of computational security for such systems. Section 4 explores variations of our definition and its implication on a certain type of watermarking systems. Section 5 reviews current theories of vision on the stages of image processing and how such processing is carried out. Section 6 introduces the formal definitions of 2FA, c-similarity, and the class $\mathcal{H}$ of hiding functions. Section 7 presents various examples of 2FA and c-(non)similarity to motivate the concepts. Section 8 establishes the properties of $\mathcal{H}$, and Section 9 discusses our results and directions for future research.

3 Model and Definition of Security

In the following, let Σ be the set $\{0,1\}$, Σ^* be the set of all binary strings, $\Sigma_{2d}^{m,n}$ be the set of all $m \times n$ rectangular binary images, and Σ_{2d}^{*} be the set of all rectangular binary images. Also let PF be the set of polynomial-time computable functions, and $\mathcal{H}$ be an appropriately defined subclass of PF containing hiding

Proceedings of the 15th IEEE Computer Security Foundations Workshop (CSFW'02)
1063-6900/02 $17.00 © 2002 IEEE

functions.

Figure 1 shows our model of a watermarking system. Alice, the sender, begins with a cover image $I \in \Sigma_{2d}^{m,n}$. She then uses a key K to select an embedding function $e_K : \Sigma_{2d}^{m,n} \mapsto \Sigma_{2d}^{m,n} \times \Sigma^*$, which maps I to the ordered pair (J, W), where J is the stego image and W is the embedded watermark. In the following, $e_K^i(I)$ will be used to denote the stego image J and $e_K^w(I)$ to denote the watermark W; an important restriction on the function $e_K^i()$ is that it must be a hiding function.

Alice sends J to Bob, the receiver. She also sends over a secure channel to Bob the key K and the original I. Bob uses K to obtain the corresponding detection function $d_K : \Sigma_{2d}^{m,n} \times \Sigma_{2d}^{m,n} \mapsto \{0, 1\}$, which belongs to PF. Alternatively, Alice may obtain the key K directly from Bob, in which case only the original needs to be sent over the secure channel.

On transit, the stego image J may be subjected to normal processing (such as compression) or malicious attacks by Wendy, the attacker; again, all such manipulations are functions $f : \Sigma_{2d}^{m,n} \mapsto \Sigma_{2d}^{m,n}$ in $\mathcal{H}$. Upon receiving J', the modified version of J, Bob computes $d_K(J', I)$ to detect the watermark; d_K outputs 1 (yes) or 0 (no).

Computational security of a watermarking system can now be defined as follows:

Definition 1 *A watermarking system is* computationally secure *if for any key K, the pair of embedding function $e_K \in \mathcal{H}$ and detection function $d_K \in PF$ satisfy the condition:*

$$\forall h \in \mathcal{H} : d_K(I, h(e_K^i(I))) = 1.$$

In other words, a system is computationally secure if its detection functions always detect embedded watermarks from stego images even after being altered by a hiding function.

4 Scope and Consequences of Model

Our model could be modified several ways to accommodate different types of already proposed watermarking systems without affecting our definition of the class $\mathcal{H}$ of hiding functions.

For example, it is assumed that the key K and the cover image I uniquely determine the watermark W to be embedded into I (such as copyright information, serial number). Alternatively, $e_k()$ could be defined as a two-argument function, in which case the encoding function is allowed to embed arbitrary watermark (e.g. secret messages) into the cover image.

Watermarking systems are classified in [22] as *private*, *semi-private*, or *public*, depending on whether the receiver has access to the original image as well as the watermark, only the watermark, or neither. Our model, as defined, covers private systems but could be extended to the other two types by requiring Alice to send over the secure channel only the watermark (which could be generated from the original image) or nothing at all. In the latter case, the detection algorithm could be required to extract the embedded watermark as well.

Such public watermarking systems with automatically generated watermarks and extraction algorithms (as opposed to detection) serve as a good starting point for our study, because they are easier to analyze. We note some of their properties that are direct consequences of our definitions in the following observations. Keep in mind that extracting algorithms return the embedded watermark in the input stego-image.

Observation 1 *$e_K^i()$ cannot be a one-to-one function.*

If $e_K^i()$ is one-to-one, then because it is a permutation of images of the same dimensions, $e_K^i()$ must also be onto. In this case, we show that $e_K^i()$ must use the same watermark for every image of the same dimensions. This is based on the observation that

1. any two images that differ by a single bit are considered visually similar;
2. for any two images I and J of the same dimensions, there is a sequence of images $I_1 = I$, I_2, ..., $I_r = J$, such that I_j and I_{j+1} differ by a single bit.

Suppose $e_K(I) = (I_1, W)$ for some image I. Then $d_K(I_1) = W$, and furthermore $d_K(h_1(I_1)) = W$, where h_1 is a hiding function that flips a single bit of I_1. Now since e_K^i is one-to-one, $h_1(I_1) = e_K^i(I_2)$ for some I_2, and hence if h_2 is a hiding function that flips a single bit of $h_1(I_1)$, then $d_K(h_2(h_1(I_1))) = W$ also. By flipping arbitrary bits and continuing this argument, it's clear that any image of the same dimensions as I must contain the same watermark W.

Observation 2 *$e_K^w()$ cannot be a one-to-one function.*

If different images are assigned different watermarks, then they must be mapped to different stego images, i.e. if $e_K^w(I) \neq e_K^w(J)$ whenever $I \neq J$, then $e_K^i(I) \neq e_K^i(J)$, directly contradicting Observation 1.

Observation 3 *Composition should be not commutative for encoding functions, i.e. $e_K^i \circ e_L^i \neq e_L^i \circ e_K^i$.*

Allowing encoding functions to commute under composition would lead to dispute about ownership of the doubly stego image $e_K^i(e_L^i(I))$:

$$d_K(e_K^i(e_L^i(I))) = d_K(e_L^i(e_K^i(I))) = d_K(e_K^i(I)) = e_K^w(I).$$

Proceedings of the 15th IEEE Computer Security Foundations Workshop (CSFW'02)
1063-6900/02 $17.00 © 2002 IEEE

$$d_L(e^i_K(e^i_L(I))) = d_L(e^i_L(I)) = e^w_L(I).$$

This observation was first made in [7].

Observation 4 *Encoding functions cannot mark already marked images.*

$$d_K(e^i_K(e^i_K(I))) = e^w_K(I) = e^w_K(e^i_K(I)).$$

The next observation applies to private and semi-private watermarking systems.

Observation 5 *$e^i_K()$ is a hiding function. Its inverse (if exists) cannot be a hiding function also.*

If the inverse h of $e^i_K()$ is also a hiding function, we have

$$d_K(I) = d_K(h(e^i_K(I))) = e^w_K(I).$$

In other words, every image I already contains the watermark that would be embedded by $e_K()$, making the encoding process unnecessary.

5 Current Theories of Vision

An understanding of the human vision system would be helpful in searching for a formalization of the class of hiding functions. Recent advances in vision science have helped clarify the various stages of human vision processing and the mechanisms underlying it. A comprehensive treatment of the science of vision can be found in [23].

From the work by David Marr and others [24], vision processing can be classified into four stages: imaged-based, surface-based, object-based, and category-based. Light reflecting or emitting from an object strikes the retinas to form two two-dimensional patterns, from which features such as lines, edges and surfaces are deduced. A full three-dimensional representation is then constructed, and the resulting object is classified according to its functions. Much more is known about the first two stages than the latter two, mainly because they have a biological basis and can be studied experimentally.

Two competing theories exist to explain how object representations are constructed from from the two-dimensional retinal patterns. The line-and-edge detector theory (due to Nobel laureates Hubel and Wiesel) attributes to special cell groups (simple, complex, and hypercomplex) in the striate cortex the ability to detect lines, edges, or other basic structures in images. On the other hand, the spatial frequency theory proposes that images are decomposed into sinusoidal gratings, which are images whose luminance varies according to a sine wave over one spatial dimension and is constant over the other one. This process is very similar to Fourier analysis, which is a common technique in digital image processing. There are physical evidences to support both views, and it is possible that human vision processing is based on a combination of both systems.

It seems reasonable to assume that visual fidelity can be determined from retinal images especially since, for the purpose of watermarking, the images to be compared are two-dimensional and may not be easily perceivable as collections of coherent objects. Discussions in the previous paragraph suggest three models for formalizing the class of hiding functions: automata on two-dimensional inputs (image-based); frequency domain transforms, such as Fourier, discrete cosine, wavelet, etc. (surface-based); and sets of lines, edges, and blobs (surface-based).

6 c-similarity and Hiding Functions

To define hiding functions, we need to formalize the concept of image similarity with respect to the human visual system. In this section we propose a model for determining visual fidelity using a variant of the two-dimensional finite automaton. This device was first introduced by Blum and Hewitt [19] and studied by Rosenfeld and others [20, 21] in the context of computer vision. The only difference between the two-dimensional finite automaton and its one-dimensional counterpart is the input tape, which is rectangular and surrounded on all four sides by a boundary of special symbols \$. The variant we will be using adds a second head to the two-dimensional finite automaton, as illustrated in Figure 2. Its formal definition is given below.

Definition 2 *A two-dimensional two-head finite automaton (2FA) is a six-tuple $M = (Q, q_0, F, \Sigma, \$, \delta)$, where*

- *Q is a finite set of state;*
- *$q_0 \in Q$ is the initial state;*
- *$F \subseteq Q$ is the set of accepting states;*
- *Σ is the input alphabet*
- *$\$ \notin \Sigma$ is the boundary symbol;*
- *$\delta : Q \times (\Sigma \cup \{\$\}) \times (\Sigma \cup \{\$\}) \mapsto Q \times \{\leftarrow, \rightarrow, \uparrow, \downarrow\} \times \{\leftarrow, \rightarrow, \uparrow, \downarrow\}$ is the transition function.*

A 2FA begins in state q_0 at the upper left square of a rectangular $m \times n$ tape (which is surrounded by boundary symbols), and makes a series of moves according to the transition function δ. The machine halts if it

Proceedings of the 15th IEEE Computer Security Foundations Workshop (CSFW'02)
1063-6900/02 $17.00 © 2002 IEEE

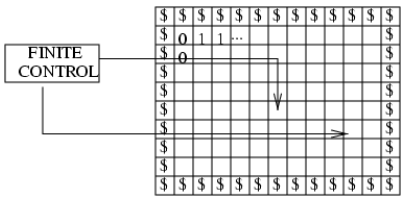

Figure 2. The two-head two-dimensional finite automaton.

enters an accepting state; in this case, we say the input on the tape is accepted. If the machine never halts, then we say the input is rejected.

We can now define two images to be similar if they are indistinguishable by 2FA that are quantified in size.

Definition 3 *Let I and J be two $m \times n$ rectangular images in Σ_{2d}^* and $c \geq 2$ be some constant. I and J are said to be* c-similar *($I \sim_c J$) if every 2FA with at most $\lceil (\log\log mn)^{1-1/c} \rceil$ states either accepts or rejects both I and J.*

The class $\mathcal{H}$ of hiding functions can now be defined as consisting of functions whose outputs are c-similar to their inputs.

Definition 4 *A function $h : \Sigma_{2d}^* \mapsto \Sigma_{2d}^*$ is a* hiding function *if*

1. *h is computable in polynomial time;*
2. *$h(I)$ has the same dimensions as I;*
3. *there exists a constant $c \geq 2$ such that $I \sim_c h(I)$ for all I.*

The class $\mathcal{H}$ consists of all hiding functions.

7 Some Examples

In this section we provide some examples of 2FA to familiarize the reader with our model and to argue that c-similarity agrees in some respects with the human eye's notion of image similarity. The examples also serve to motivate our choice of model, e.g. the need for an extra head, and the restriction on the number of states to $\lceil (\log\log mn)^{1-1/c} \rceil$.

Example 1 (Contrast)

Our first example illustrates the ability of the 2FA to detect gross but not fine differences in contrast. Let $I_0^{m\times n}$ ($I_1^{m\times n}$) be the $m \times n$ images consisting of only 0 (1) bits. It is easy to construct a 2FA M whose number of states is independent of mn to witness $I_0^{m\times n} \not\sim_c I_1^{m\times n}$: M scans its input line by line; it accepts when reaching the bottom boundary and rejects when reading a 1. In fact, one can construct a 2FA to determine whether an image has more black than white pixels. On the other hand, it doesn't seem possible to construct a 2FA to determine whether every black pixel lies on some prime number row.

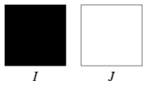

Figure 3. $I_1^{m\times n} \not\sim_c I_0^{m\times n}$.

Example 2 (Parity)

A 2FA M can be constructed to determine whether the number of black pixels of an image equals to a fixed value mod k, where k is at most $\lceil (\log\log mn)^{1-1/c} \rceil$ (which is a small number for all practical purposes): M scans the input image line by line and keeps track of the number of black pixels mod k in its state.

In particular, no two images are similar if they have different parities. On the other hand, it is clear that the human visual system is not sensitive to the parity of the number of black pixels in a newspaper image, so our model is not very realistic in this respect. However, the human eye does check for parity on a global level; e.g. it is easy to tell (if one pays attention) whether there is an odd or even number of people in an image.

Example 3 (Hot spots)

The human eye can better detect irregularities at the corners or center of an image. It is easy to construct a 2FA with $O(\lceil (\log\log mn)^{1-1/c} \rceil)$ states to discriminate images on the value of a bit at a position near a corner such as $(\lceil (\log\log mn)^{1-1/c} \rceil, \lceil (\log\log mn)^{1-1/c} \rceil)$ by counting with the finite control.

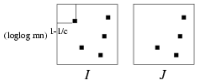

Figure 4. $I \not\sim J$ at position $(\lceil (\log\log mn)^{1-1/c} \rceil, \lceil (\log\log mn)^{1-1/c} \rceil)$.

Similarly, one can construct a 2FA with a constant number of states to find the center of the input image:

Proceedings of the 15th IEEE Computer Security Foundations Workshop (CSFW'02)
1063-6900/02 $17.00 © 2002 IEEE

the 2FA moves one head twice as fast as the other to find the middle column and then the middle row.

A more elaborate construction shows how a 2FA can reach any position on a $2^x \times 2^x$ image using only $O(x)$ states. We show how to construct a 2FA M to locate any row R by computing R as a sum of powers of 2. Let the binary representation of R be $b_{x-1}b_{x-2}\ldots b_0$. First M locates column 2^{x-1} by moving one head along the first row twice as fast as the other head, which moves diagonally if b_{x-1} is 1 and horizontally otherwise. At the end of this stage, the second head is on column 2^{x-1} and row $b_{x-1}2^{x-1}$. M then positions the first head on column 2^{x-1} of the first row by moving both heads to the left until the second head reaches the left boundary (which is still on row $b_{x-1}2^{x-1}$.)

Next, M locates column 2^{x-2} by moving one head along the first row twice as fast as the other head, which moves diagonally if b_{x-2} is 1 and horizontally otherwise. At the end of this stage, the second head is on column 2^{x-2} and row $b_{x-1}2^{x-1} + b_{x-2}2^{x-2}$, etc.

After repeating this process x times, the second head of M will be positioned on row $\sum_{i=0}^{x-1} b_{x-1-i}2^{x-1-i} = R$. The number of states required is $O(x)$.

The above construction shows why it is desirable to restrict the number of states of a 2FA. No two different images would be similar if the number of allowable states is $O(\log m + \log n)$. The bound of $\lceil(\log\log mn)^{1-1/c}\rceil$ is selected because it is a straightforward choice that gives rise to a nontrivial class $\mathcal{H}$.

Example 4 (Handedness)

The human eye can easily detect the difference between an asymmetric picture and its mirror image. Similarly, a 2FA can be used to detected handedness as illustrated in the following example. Let I be the blank image except for some black pixel on its left-hand side, and J be the mirror image of I. There is a 2FA M that accepts if and only if the left half of its input is blank; it does so by using the extra head to tell when the first head has crossed into the right half. M distinguishes I from J. It should be noted that the use of the second head is indispensable in this case.

Figure 5. Handedness

Example 5 (Proportionality)

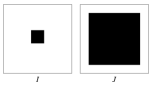

Figure 6. A Difference in Proportion

The human eye can detect the difference between the following two images.

Although both images consist of a black square centered in a bigger white square, the difference lies in the proportion of their sizes (10% vs. 90%). It is easy to construct a 2FA to distinguish the difference: the 2FA locates the center of the upper left quadrant and accepts if and only if the pixel at that location is black. Again, the use of the second head is necessary in this example.

8 Properties of $\mathcal{H}$

In this section we show that the problem of deciding whether two images are c-similar is decidable in polynomial time. We use this result to show that $\mathcal{H}$ is not a trivial class: beside the identity function, it also contains an infinite number of interesting functions computable in polynomial time.

First we show that c-similarity defines an equivalence relation over the set of all $m \times n$ images.

Proposition 1 *$\sim_c$ is an equivalence relation over $\Sigma_{2d}^{m,n}$.*

Proof: Reflexivity and associativity follow directly from the definition of c-similarity. To see transitivity, suppose $I \sim_c J$ and $J \sim_c K$, where I and J are $m \times n$ images. If $I \not\sim_c K$, then there must be a 2FA M with at most $\lceil(\log\log mn)^{1-1/c}\rceil$ states that accepts I and rejects K (or vice-versa). Since $I \sim_c J$, M accepts J and hence M witnesses the fact that $J \not\sim_c K$, a contradiction. ■

The next proposition states that the number of 2FA allowed by the definition of c-similarity is quite small.

Proposition 2 *Let Σ be an alphabet. The number of different 2FA of at most $\lceil(\log\log mn)^{1-1/c}\rceil$ states is $o(\log mn)$.*

Proof: Each 2FA with at most Q states can be described by a table of $|Q| \times (|\Sigma|+1) \times (|\Sigma|+1)$ entries. There are $|Q| \times 4 \times 4 + 1$ choices for each entry, and hence the total number of different 2FA with at most $\lceil(\log\log mn)^{1-1/c}\rceil$ states is

$$N = (16(\lceil(\log\log mn)^{1-1/c}\rceil)+1)^{(\lceil(\log\log mn)^{1-1/c}\rceil)(|\Sigma|+1)^2} \leq$$

Proceedings of the 15th IEEE Computer Security Foundations Workshop (CSFW'02)
1063-6900/02 $17.00 © 2002 **IEEE**

$$(\alpha(\log\log mn)^{1-1/c})^{\beta(\log\log mn)^{1-1/c}} \leq$$

$$2^{\gamma(\log\log mn)^{1-1/c}\log\log\log mn} = o(2^{\log\log mn}) = o(\log mn),$$

for appropriate constants α, β, and γ. ■

Propositions 1 and 2 immediately imply that for any particular value of $c \geq 2$ and large enough values for m and n, there is a large set of $m \times n$ images that are not identical but c-similar to one another.

Proposition 3 *Let $c \geq 2$ be any constant. For sufficiently large values of m and n, there exists a set of at least $2^{mn-\log mn}$ $m \times n$ images that are pairwise c-similar.*

Proof: By Propositions 1 and 2, there are at most $2^{\log mn}$ different c-similar equivalence class of $\Sigma_{2d}^{m,n}$. Since there are 2^{mn} different images, one such equivalence class must have at least $2^{mn-\log mn}$ members. ■

We are now ready to establish the first main result of this section.

Theorem 1 *Given two $m \times n$ images I and J in Σ_{2d}^*, the problem of determining whether $I \sim_c J$ can be solved in polynomial time.*

Proof: The following Turing machine M decides whether $I \sim_c J$: M constructs all possible 2FA with at most $\lceil(\log\log mn)^{1-1/c}\rceil$ states and simulates each over I and J. M answers 'yes' if and only if each of the 2FA either accepts or rejects both I and J. Since there are $o(\log mn)$ different 2FA with at most $\lceil(\log\log mn)^{1-1/c}\rceil$ states, and the simulation of each 2FA over I or J takes at most $(\lceil(\log\log mn)^{1-1/c}\rceil)(mn)^2$ steps (the machine can visit each valid pair of tape positions only a finite number of times), the total time taken is $O((mn)^2 \log mn \log\log mn)$. Hence M runs in polynomial-time. ■

We use Theorem 1 to show that $\mathcal{H}$ contains infinitely many nonidentity functions computable in polynomial-time.

Theorem 2 *$\mathcal{H}$ contains infinitely many functions computable in polynomial time.*

Proof: Let $c \geq 2$ be any constant. We construct a nonidentity function h_c computable in polynomial time and $h_c(I) \sim_c I$ for all binary $m \times n$ image I as follows: given an image I_k, h_c enumerates according to some canonical order the next mn binary images $I_{k+1}, I_{k+2}, \ldots, I_{k+mn}$. The value of h_c is the first such image that is c-similar to I_k, or I_k if none of them is c-similar to I_k. By Proposition 3, for large enough values of m and n there must exist an image I such that $h_c(I) \neq I$, i.e. h_c is not the identity function. It takes $O((mn)^2)$ time to enumerate the mn images following the input in the canonical order, and by Theorem 1 it takes $O((mn)^2 \log mn \log\log mn)$ time to check if each of those is c-similar to the input. Hence h_c is computable in polynomial time. ■

Finally we show that $\mathcal{H}$ has some nice closure properties.

Theorem 3 *$\mathcal{H}$ is closed under composition and polynomial-time computable inverse.*

Proof: First, it is clear from the definition that if $I \sim_c J$ and $d < c$, then $I \sim_d J$. Now suppose $I \sim_c f(I)$ and $I \sim_d g(I)$. Since $I \sim_d g(I) \sim_c f(g(I))$, we have $I \sim_{min(c,d)} f(g(I))$. Thus $\mathcal{H}$ is closed under composition.

Now, let $f \in \mathcal{H}$ be a one-one function and f^{-1} be its *polynomial-time computable* inverse. We have $f^{-1}(I) \sim f(f^{-1}(I)) = I$. Thus $\mathcal{H}$ is closed under polynomial-time computable inverse.

Since the identity function is in $\mathcal{H}$, it follows from Theorem 3 that the restriction of $\mathcal{H}$ to polynomial-time invertible functions is a group. ■

9 Conclusions

We have introduced a complexity-theoretic model for studying computational security of binary image watermarking systems. Unlike previous work, our model restricts algorithms for embedding and destroying watermarks to the class $\mathcal{H}$ of hiding functions. These are efficiently computable functions that preserve visual fidelity of the input images. A system is defined to be computationally secure if its embedded watermarks remain detectable even after distortions by any hiding function. Security of watermarking systems is to be established with complexity results about hiding functions.

We reviewed current theories of vision, which suggest three ways of modeling the class of hiding functions $\mathcal{H}$: with automata, with frequency domain transforms, and with sets of lines and edges.

We also proposed a candidate for the class $\mathcal{H}$, which is based on an automata-theoretic model of visual fidelity called *c-similarity*. We showed that our choice of model agrees in some respects to the human eye's notion of similarity and that the class $\mathcal{H}$ is robust and contains infinitely many functions computable in polynomial time. It should be emphasized that c-similarity

Proceedings of the 15th IEEE Computer Security Foundations Workshop (CSFW'02)
1063-6900/02 $17.00 © 2002 IEEE

is a very unrealistic model for measuring human visual fidelity; its main purpose is to serve as a starting point for a qualitative rather than numerical model for which hardness results can be formulated and proved.

We list below some directions for future research regarding hiding functions and computational security of watermarking systems:

1. We identified some constraints on public watermarking systems as consequences of our definition of security. Do the same for private and semi-private watermarking systems.
2. Enhance two-dimensional finite automata to model frequency-domain transforms and/or sets of lines and edges.
3. Refine the definition *c-similarity* to reflect other properties of the human visual system.
4. Are there functions in our candidate for $\mathcal{H}$ that are not polynomial-time invertible ? Recall that the existence of such functions is required by our model to prove that a system is computationally secure (Section 4);
5. Construct a system of embedding and detection functions which are hiding functions (e.g. obeying c-similarity) or prove that it doesn't exist.

10 Acknowledgement

I thank Casey Carter for helpful discussions on c-similarity and the HVS.

References

[1] R. G. van Schyndel, A. Z. Tirkel, and C. F. Osborne, "A digital watermark," in *Proceedings of International Conference on Image Processing*, 1994, pp. 86–90.

[2] W. Bender, D. Gruhl, N. Morimoto, and A. Lu, "Techniques for data hiding," *IBM Systems Journal*, vol. 35, no. 3-4, pp. 313–336, 1996.

[3] I. J. Cox, J. Kilian, T. Leighton, and T. Shamoon, "A secure, robust watermark for multimedia," in *Information Hiding*, 1996, vol. 1174 of *Springer Lecture notes in Computer Science*, pp. 183–206.

[4] J. Brassil, S. Low, N. Maxemchuk, and L. O'Gorman, "Electronic marking and identification techniques to discourage document copying," *IEEE Journal on Selected Areas of Communications*, vol. 13, no. 8, pp. 1495–1504, Oct 1995.

[5] R. B. Wolfgang, C. I. Podilchuk, and E. J. Delp, "Perceptual watermarks for digital images and video," *Proceedings of the IEEE*, vol. 87, no. 7, pp. 1108–1126, 1999.

[6] S. Katzenbeisser and F. A. P. Petitcolas, Eds., *Information Hiding Techniques for Steganography and Digital Watermarking*, Artech House, 2000.

[7] S. Craver, N. Memon, B.-L. Yeo, and M. M. Yeung, "Resolving rightful ownerships with invisible watermarking techniques: Limitations, attacks, and implications," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 4, pp. 573–586, 1998.

[8] M. Kutter and F. A. P. Petitcolas, "A fair benchmark for image watermarking systems," in *Proceedings of the SPIE, Security and Watermarking of Multimedia Contents*, 1999, pp. 226–239.

[9] F. A. P. Petitcolas and R. Anderson, "Evaluation of copyright marking systems," in *Proceedings of IEEE Multimedia Systems*, 1999, pp. 574–579.

[10] C. Cachin, "An information-theoretic model for steganography," in *Proceedings of the 2nd International Workshop on Information Hiding*, 1998, pp. 306–318.

[11] J. Zöllner, H. Federrath, H. Klimant, A. Pfitzmann, R. Piotraschke, A. Westfeld, G. wick, and G. Wolf, "Modeling the security of steganographic systems," in *Proceedings of the Second Internation Workshop on Information Hiding*, 1998, pp. 344–354.

[12] T. Mittelholzer, "An information-theoretic approach to steganography and watermarking," in *Proceedings of the Third Internation Workshop in Information Hiding*, 1999, pp. 1–16.

[13] C. E. Shannon, "Communication theory of secrecy systems," *Bell Systems Technical Journal*, vol. 28, pp. 656–715, 1949.

[14] J. L. Mannos and D. J. Sakrison, "The effects of a visual fidelity criterion on the encoding of images," *IEEE Transactions on Information Theory*, vol. 20, no. 4, pp. 525–536, July 1974.

[15] A. B. Watson, "The cortex transform: Rapid computation of simulated neural images," *Computer Vision, Graphics and Image Processing*, vol. 39, pp. 311–327, 1987.

Proceedings of the 15th IEEE Computer Security Foundations Workshop (CSFW'02)
1063-6900/02 $17.00 © 2002 **IEEE**

[16] C. Lambrecht and J. Farrell, "Perceptual quality metric for digitally coded color images," in *Proceedings of the European Signal Processing Conference*, 1996, pp. 1175–1178.

[17] C. Lambrecht, *Perceptual models and Architectures for Video Coding Applications*, Ph.D. thesis, École Polytechnique Fédérale de Lausanne, 1996.

[18] C. Papadimitriou, *Computational Complexity*, Addison-Wesley, 1994.

[19] M. Blum and C. Hewitt, "Automata on a 2-dimensional tape," in *Conference Record of 1967 Eighth Annual Symposium on Switching and Automata Theory*, Austin, Texas, 18–20 Oct. 1967, IEEE, pp. 155–160.

[20] A. Rosenfeld, *Picture Languages (Formal Models for Picture Recognition)*, Academic, New York, 1977.

[21] K. Inoue and I. Takanami, "A survey of two-dimensional automata theory," *Information Sciences*, vol. 55, no. 1–3, pp. 99–121, June 1991.

[22] F. A. P. Petitcolas, R. J. Anderson, and M. G. Kuhn, "Information hiding – a survey," *Proceedings of the IEEE*, vol. 87, no. 7, pp. 1062–1078, July 1999.

[23] S. E. Palmer, *Vision Science: Photons to Phenomenology*, MIT Press, 1999.

[24] D. Marr, *Vision*, W. H. Freeman, 1983.

Proceedings of the 15th IEEE Computer Security Foundations Workshop (CSFW'02)
1063-6900/02 $17.00 © 2002 **IEEE**