Resolving Rightful Ownerships with Invisible Watermarking Techniques: Limitations, Attacks, and Implications

Scott Craver, Nasir Memon, Boon-Lock Yeo, *Member, IEEE*, and Minerva M. Yeung, *Member, IEEE*

***Abstract*— Digital watermarks have been proposed in recent literature as a means for copyright protection of multimedia data. In this paper we address the capability of invisible watermarking schemes to resolve copyright ownership. We show that, in certain applications, rightful ownership cannot be resolved by current watermarking schemes alone. Specifically, we attack existing techniques by providing counterfeit watermarking schemes that can be performed on a watermarked image to allow multiple claims of rightful ownership. In the absence of standardization and specific requirements imposed on watermarking procedures, anyone can claim ownership of any watermarked image.**

In order to protect against the counterfeiting techniques that we develop, we examine the properties necessary for resolving ownership via invisible watermarking. We introduce and study *invertibility* and *quasi-invertibility* of invisible watermarking techniques. We propose noninvertible watermarking schemes, and subsequently give examples of techniques that we believe to be nonquasi-invertible and hence invulnerable against more sophisticated attacks proposed in the paper. The attacks and results presented in the paper, and the remedies proposed, further imply that we have to carefully reevaluate the current approaches and techniques in invisible watermarking of digital images based on application domains, and rethink the promises, applications and implications of such digital means of copyright protection.

***Index Terms*—Attacks on digital watermarks, copyright protection, counterfeit watermarks, cryptography, invertible and noninvertible watermarking, invisible watermarks, quasi-invertible watermarking.**

I. Introduction

THE rapid growth of digital imagery coupled with the ease by which digital information can be duplicated and distributed has led to the need for effective copyright protection tools. Various watermarking schemes and software products have been recently introduced in attempt to address this growing concern. Given the flurry of activity that has resulted, it is natural to ask a few questions regarding all these efforts: What is a digital watermark? Why are digital watermarks necessary, or in other words, what can digital watermarks achieve, or fail to achieve? What can digital watermarks do for copyright protection in addition to current copyright laws and avenues for resolving copyright grievances?

In general, there are two types of digital watermarks addressed in existing literature, visible and invisible watermarks.[1] These watermarking schemes are designed mainly for two purposes—copyright protection and data authentication. In this paper we shall focus on the applicability of invisible watermarking techniques for one instance of copyright protection—that is, identification of an image's rightful owner(s). In this case, the watermarks embedded in an image have to be recoverable, despite intentional or unintentional modification of the image. They should be robust against innocent image processing operations like filtering, requantization, dithering, scaling, cropping, etc., and common image compression techniques. They must also be invulnerable to deliberate attempts to forge, remove, or invalidate watermarks.

A variety of invisible watermarking schemes have been reported in recent years (for example, [2]–[10] and commercial system like Digimarc's [11], [12]; see also [13] and references therein). Such techniques can be broadly classified in two categories: spatial-domain and transform-domain based. The earlier watermarking techniques reported were spatial in nature, the simplest being the ones that modified the least significant bits (LSB) of an image's pixel data [4]. Improvement and variants of these techniques are proposed in [10], [3], [14], [7], [8]. These techniques have been shown to be quite robust against lossy image compression, filtering, and scanning. As opposed to spatial-domain-based techniques, that have relatively low-bit capacity, transform-domain-based techniques can embed a large number of bits without incurring noticeable visual artifacts. Such techniques can be employed with common image transforms like discrete cosine transforms (DCT), wavelets, Fourier transforms such as the FFT, and Hadamard transforms. One of the earlier transform-domain-based techniques tailored to JPEG lossy image compression [15] was reported by Zhao and Koch in [5]. Other techniques include the works report in [9] and [6]. A more robust technique based on spread spectrum principles is given by Cox

Manuscript received March 1997; revised August 1997. This paper was presented in part at IS&T/SPIE Electronic Imaging 1997: Storage and Retrieval for Image and Video Databases V, San Jose, CA in February 1997 and at the 1997 International Conference on Image Processing, Santa Barbara, CA in October 1997.

S. Craver is with the Department of Mathematics, Northern Illinois University, DeKalb, IL 60115 USA.

N. Memon is with the Imaging Technology Department, Hewlett Packard Laboratories, Palo Alto, CA 94304 USA, on leave from the Department of Computer Science, Northern Illinois University, DeKalb, IL 60115 USA.

B.-L. Yeo and M. M. Yeung were with IBM T. J. Watson Research Center, Yorktown Heights, NY 10598 USA. They are now with Intel Research Labs, Santa Clara, CA 95052 USA.

Publisher Item Identifier S 0733-8716(98)01105-6.

[1] Some papers, such as [2], discuss watermarking other forms of multimedia data such as sound clips. Our research has focused on image data, and hence we say "invisible" when in a wider sense we mean "imperceptible." The ideas presented in this paper also apply to other forms of multimedia data.

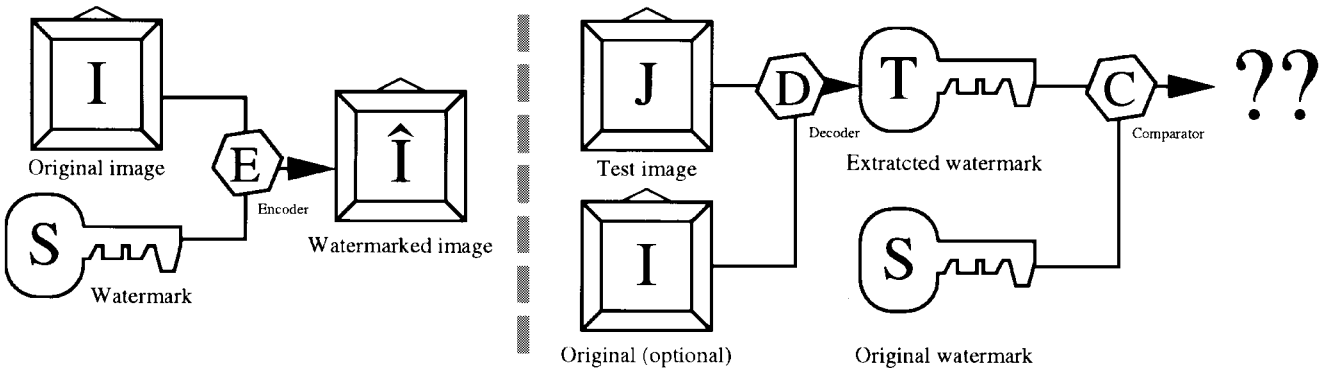

Fig. 1. Encoding, decoding, and comparing embedded watermarks in an image.

et al. [2]. They embed a set of independent and identically distributed samples drawn from a Gaussian distribution into the perceptually most significant frequency components of the data. Results reported with the largest 1000 DCT coefficients show the technique to be remarkably robust against various image processing operations, and after printing and rescanning.

Unfortunately, many of these existing schemes have not addressed the *ends* of invisible watermarking schemes. They instead focused on the *means* to label an image invisibly and the *robustness* of the inserted labels against malicious attacks. As a result, the concerns regarding what watermarks can achieve or fail to achieve may not have been properly addressed. While it is of course important to address the technical capabilities of watermarking techniques, equally important is the ability to know how and when said watermarking techniques can be used to protect data, and if such protection is based on sound legal justification. For example, consider a distributed system similar to the World Wide Web. In this model, we have a number of users who create digital images and make them available to many other users to view and potentially copy. Suppose that in this system we have no central authority actively monitoring, maintaining, or enforcing ownership rights to information. If a user wants to be able to retain ownership rights to an image (and all its modified forms obtained by geometric and other common "content preserving" transformations), is it possible to do so by using one of the robust watermarking techniques reported in the literature? That is, can these watermarking techniques really be used to prove beyond a reasonable doubt that an image in dispute has actually been derived from a user's original?

We will show in this paper that the answer to the above question is no, at least not with some current invisible watermarking schemes which we shall show to be unable to resolve rightful ownership of an image watermarked with multiple ownership labels. In addition, without any standardization of watermarking techniques or specification of certain requirements for them (that is, without properly answering the question "What is a digital watermark?"), we shall show that anyone can claim ownership of any image by the methods described in later sections. The results, coupled with recent attacks on watermarking schemes reported in [16], further suggest that we have to carefully rethink our approaches to invisible watermarking of images, and reevaluate the promises of such digital means of copyright protection. In other words, it is crucial that any watermarking scheme proposed for copyright protection be able to answer the last two questions: "Why is it necessary?" and "How useful is it?"

The paper is organized as follows. In Section II, we present general definitions and notations used to describe digital watermarking schemes. In Section III, we discuss how digital watermarking can be used to resolve rightful ownership, and depict a scenario in which there may be more than one "rightful" owner of an image. We then show in Section IV that such a scenario can actually be created by developing counterfeit watermarking schemes that can be performed on a watermarked image to allow multiple claims of rightful ownership. An implementation of such a scheme, which is used to invalidate the watermarking method proposed by Cox *et al.* [2] is also described. In Section V we define the term *invertibility* of invisible watermarking schemes, and present *noninvertible* watermarking schemes as a method of preventing the type of attack described in Section IV. Unfortunately, noninvertibility by itself does not prove to be enough to prevent a more powerful attack presented in Section VI. This leads us to the definition of *quasi-invertible* watermarking schemes. We then present a noninvertible watermarking technique that, to the best of our knowledge, is also not quasi-invertible. In Section VII, we address invertibility issues for invisible watermarking schemes that do not use the original image in the watermark extraction process, and provide an example of a counterfeitattack on the scheme proposed by Pitas [14]. We conclude in Section VIII with a discussion.

II. Definitions and Notation

In this section, we give a generalized formulation of invisible watermarking schemes. We define in general terms the process of watermark insertion into an image and the use of invisible watermarks to determine the ownership of a watermarked image. Fig. 1 illustrates the encoding process by which a watermark is inserted into an image, and the decoding process by which a watermark is recovered and then compared to the inserted watermark.

Here we use I to denote an image, S a watermark consisting of a sequence of *ownership labels* $S = \{s_1, s_2, \cdots\}$ and $\hat{I}$ the watermarked image. $\mathcal{E}$ is an encoder function if it takes an

image I and a watermark S, and generates a new image which is called the *watermarked* image $\hat{I}$, i.e.,

$$\mathcal{E}(I,S)=\hat{I}. \tag{1}$$

It should be noted that we do not exclude the possibility that the watermark S is dependent upon the image I. In such cases, the encoding process described by (1) still holds.

A decoder function $\mathcal{D}$ takes an image J (J can be a watermarked or unwatermarked image, and possibly corrupted) whose ownership is to be determined, and recovers a watermark S' or *evidence* of a watermark S' from the image. In this process, an additional *reference image* I can also be included, that is often the original (and unwatermarked) version of J. This is due to the fact that some decoding schemes may make use of the original image in the watermarking process to provide extra robustness against intentional and unintentional corruption of pixel values. If the decoding scheme involves a reference image I, we have

$$\mathcal{D}(J,I)=P(T) \tag{2}$$

where P is a function indicating the presence of watermark T in J. We shall call this type of watermarking scheme a "*private*" watermarking scheme, following the terminology in [13]. Some examples of private watermarking schemes include [2], [9], [6]. When $P(T)=T$, the decoding simply returns the extracted watermark T. P may also be of the form $P(\,)=\mathrm{Evid}(\,)$ that returns a scalar value indicating the evidence of the presence of watermark S in J.

If the decoding does not need I in the decoding process, we write a general decoding function as

$$\mathcal{D}(J)=P(T). \tag{3}$$

This type of watermarking scheme is called a "*public*" watermarking scheme according to [13]. Some examples of public watermarking schemes include [5], [3], [14].

When $P(T)=T$, the extracted watermark T is then compared to the owner's watermark S by a comparator function $\mathcal{C}_\delta$, and a binary output decision is generated indicating a match or otherwise:

$$\mathcal{C}_\delta(T,S)=\begin{cases}1, & c\geq\delta;\\ 0, & \text{otherwise.}\end{cases} \tag{4}$$

Here, c is the correlation of the two watermarks. A diagram of the decoding process is shown in Fig. 1. Without loss of generality, a public watermarking scheme can be treated as a three-tuple $(\mathcal{E},\mathcal{D},\mathcal{C}_\delta)$, such that $\mathcal{D}(\mathcal{E}(I,S))=S$ for any image I and any allowable watermark S.

We will use $\mathcal{D}(X)$ as a generic notation to also denote the decoding of private watermarking scheme, when the reference image is unambiguous. In addition, we shall use $\mathcal{C}_\delta(\mathcal{D}(\hat{I}),S)$ as a generic expression to measure the presence of watermark S in $\hat{I}$. When $\mathcal{D}(\hat{I})=T$, then $\mathcal{C}_\delta(\mathcal{D}(\hat{I}),S)=\mathcal{C}_\delta(T,S)$. When $\mathcal{D}(\hat{I})=\mathrm{Evid}(\hat{I})$, then $\mathcal{C}_\delta(\mathrm{Evid}(\hat{I}),S)=\mathcal{C}_\delta(\mathrm{Evid}(\hat{I}))$, and

$$\mathcal{C}_\delta(\mathrm{Evid}(\hat{I}))=\begin{cases}1, & \mathrm{Evid}\ (\hat{I})\geq\delta;\\ 0, & \text{otherwise.}\end{cases} \tag{5}$$

For the rest of the paper, unless otherwise stated, we primarily focus on private watermarking schemes with $P(T)=T$. Similar issues related to public watermarking are addressed in Section VII.

A variety of encoding and corresponding decoding processes have been proposed in the literature. One common approach is represented by *feature-based* private watermarking schemes that embed a watermark $S=\{s_1,s_2,\cdots\}$ into a set of derived *features* $F(I)=\{f_1(I),f_2(I),\cdots\}$. The embedding process is achieved by an *insertion operation* that we denote by the symbol $\oplus$, i.e., $f_i^*=f_i\oplus s_i$. The insertion operation has an inverse operation, namely the *extraction operation*, that we denote by $\ominus$, i.e., $f_i^*\ominus f_i=s_i$. Note that for notational simplicity we take the insertion (and extraction) process to be binary operators, although in general they could be arbitrary functions of f_i and s_i. Also note here that such class of feature-based watermarking schemes require a reference feature set $\{f_i\}$ that is derived from the image I in the watermark extraction operation.

Usually, the feature set $\{f_1(I),f_2(I),\cdots\}$ is chosen such that slight modification of individual features does not *perceptually* degrade image I. In addition, it is also desirable that each element in this set of features will not be changed significantly when the image is not perceptually degraded. An example of such a set of features would be transform-domain (e.g., DCT, wavelet) coefficients that contain significant energy content. The labels s_i that compose the watermark in this case could be real numbers drawn from a specific distribution and the insertion operation could simply be the addition of s_i to these coefficients.

Example 1: An invisible watermarking scheme as proposed by Cox *et al.* [2].

In this scheme, a two-dimensional (2-D) DCT of the image I is taken and the set $F(I)$ corresponds to the n largest-magnitude AC coefficients, typically the low-frequency ones. The encoder $\mathcal{E}$ takes a watermark S and places it in the set $F(I)$. An inverse 2-D DCT is then taken, yielding the watermarked image $\hat{I}$. To determine if a given image J contains the watermark S, the decoder $\mathcal{D}$ extracts $T=\{t_1,t_2,\cdots\}$ from J, where $t_i=f_i(J)-f_i(I)$. The confidence measure is taken to be the quantity

$$c=\frac{\sum_i t_i\cdot s_i}{\sqrt{\sum_i t_i^2}}. \tag{6}$$

Alternatively, the normalized correlation

$$c=\frac{\sum_i t_i\cdot s_i}{\sqrt{\left(\sum_i t_i^2\sum_i s_i^2\right)}} \tag{7}$$

can be used. In this case, if $J=\hat{I}$, then $c=1$. If J is a modified version of $\hat{I}$, and the changes are not perceptually significant, c will be large value. □

III. Resolving Rightful Ownerships by Invisible Watermarks

It has been generally assumed that invisible watermarking schemes may be used to protect the rights of copyright owners; at the very least, the labels extracted from watermarked images can be used to identify the rightful owner. For example, it is stated in the abstract of [2] that "Retrieval of the watermark unambiguously identifies the owner, and the watermark can be constructed to make counterfeiting almost impossible," or in [14]: "This [watermark] signal completely characterizes the person who applied it and, as a result, proves the origin of the image." But how can we do this? Does it mean that if a person produces a watermark that matches the extracted watermark from an image, then the person could automatically be considered the rightful owner of the image?

Suppose Alice and Bob[2] use the same digital watermarking technique to watermark their images. This means that there is one unique decoding scheme to extract the labels embedded in the images. If the labels extracted from a watermarked image match the particular watermark labels of Alice, then the image is believed to belong to her. Similarly, if the label matches Bob's watermark, then it must be his image. If a watermarked image contains both Alice and Bob's watermarks, whose image is it?

Note that Alice and Bob may be using entirely different watermarking schemes. Given a watermarked image, Alice can take this image and decode the label using her decoding scheme. Similarly Bob can perform the label extraction process with his decoding scheme. If Alice's decoder indicates that the image belongs to her, while Bob's decoder indicates that it is his image, whose image is it? The question of how to determine or resolve rightful ownership of an image in the face of multiple copyright ownership claims has, to our knowledge, not been explicitly raised, or answered. But the scenario is valid, given that an image can be generated and modified digitally, and any image that is watermarked by Alice and in circulation can be watermarked again by Bob. In such cases, Alice and Bob can use the same watermarking techniques, or apply different ones.

Of course, somewhere out in the dark, there are the so-called original images (or, *unwatermarked* images). Without proper copyright registration and the traditional protection of copyright laws, (after all, why are digital watermarks necessary if copyright laws can fully protect the interests of the copyright owners?) one can look to these original images to untangle a case of apparent multiple ownership. Suppose there is a watermarked image $\hat{I}$ in which the watermarks of both Alice and Bob have been detected, and both claim rightful ownership. If Alice keeps her original image (and watermark vector) locked away, she can ask Bob for his original image and check if it contains her watermark. Similarly, Bob can ask Alice for her original image and check for his watermark.

[2]Throughout the rest of this paper, we shall use two fictional characters, Alice and Bob, to illustrate the various scenarios involving the claims of copyright ownership and to bring up the different issues of the application of digital watermarks in resolving rightful ownership. Alice will be used as the originator of an image, and Bob an aspiring forger.

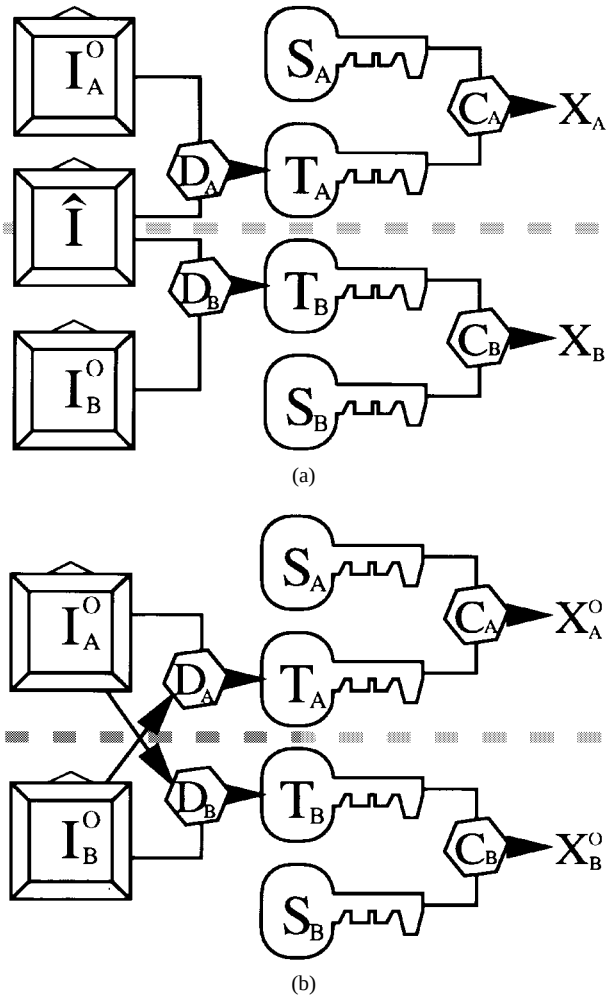

Fig. 2. Decoding tests to extract watermarks for ownership determination: (a) Test I: Alice and Bob both test an image for presence of a watermark. (b) Test II: Alice and Bob test each other's original images for presence of a watermark.

Fig. 2 illustrates how the tests of ownership by means of invisible watermarks can be implemented. $\hat{I}$ represents a watermarked image in circulation, I_A^0 is Alice's claimed original image, and I_B^0 is the claimed original of Bob. $\mathcal{D}_A, \mathcal{C}_{\delta_A}, \mathcal{D}_B$, and $\mathcal{C}_{\delta_B}$ represents the decoding and watermark comparator functions used by Alice and Bob respectively. To check the presence of Alice's and Bob's watermarks, the watermarks from the watermarked image are first extracted by Test I, as in Fig. 2(a), and compared with their respective watermarks, and similarly the decoding tests can be achieved using the two "original" images by Test II as in Fig. 2(b). The results of the tests illustrated in Fig. 2, together with the logical determination of ownership via the watermark tests, are tabulated in Table I.

If Bob obtained Alice's watermarked image and introduced his own watermark into it, then both Bob's "original" and watermarked images contain Alice's mark. Alice's original does not contain Bob's. Thus, by keeping her original image locked away with the details of the watermark label, Alice can ensure that any copy of $\hat{I}$ that Bob obtains will contain her watermark, easily foiling any such *ex post facto* watermarking of her image.

TABLE I
DETERMINATION OF OWNERSHIP FROM WATERMARK PRESENCE TESTS. "1" INDICATES THE PRESENCE OF WATERMARK, "0" INDICATES THE ABSENCE, AND "*d*" REPRESENTS *don't care*'S

Scenario	Test I		Test II		Derived
	x_A	x_B	x_A^0	x_B^0	Ownership
Case 1	1	0	d	d	Alice
Case 2	0	1	d	d	Bob
Case 3	1	1	1	0	Alice
Case 4	1	1	0	1	Bob
Case 5	1	1	1	1	???

Or can she? If Alice's original contains Bob's watermark *and vice versa*, who owns this image: Alice or Bob? In such a case rightful ownership cannot be resolved by invisible watermarks alone. We show in the following section that this scenario is not hypothetical, but can be engineered with current watermarking schemes. We present in detail a counterfeit watermarking scheme that allows multiple claims of ownership. Such counterfeit schemes can be successfully engineered from a class of invisible watermarking schemes which we shall call *invertible watermarking schemes*. We shall show that invertibility in a watermark insertion scheme renders it useless for establishing ownership.

IV. INVALIDATING CLAIMS OF OWNERSHIPS

To invalidate claims of ownership of an image, it is necessary to generate the confusion illustrated in the case of Alice and Bob as in Case 5 of Table I: 1) that both the watermarks of Alice and Bob are present in the watermarked image in circulation and 2) that there are two original images, each containing the watermark of the other party. Clearly only one of these can be the true original, so it is the attacker's goal to generate a counterfeit image he or she can claim to be the original with as much confidence as can the image's true owner. We show in this section how to create another "original" image $\hat{I}'$ (the counterfeit original) from a watermarked image $\hat{I}$, without the access to the true original image I. We shall show that, despite the fact that Bob does not have any knowledge of Alice's watermark, Bob's counterfeit watermark will be present in Alice's original I as well as her watermarked $\hat{I}$. This means that the criteria 1) and 2) are both satisfied.

More formally, given $\hat{I}$ which is watermarked by some watermarking scheme $(\mathcal{E}, \mathcal{D}, \mathcal{C}_\delta)$, we will reverse-engineer an image $\hat{I}'$, watermark S' and a decoding function $\mathcal{D}'$ that show the following properties:

$$C_{\delta'}\left(\mathcal{D}'\left(\hat{I}, \hat{I}'\right), S'\right) = 1 \tag{8}$$

$$C_{\delta'}\left(\mathcal{D}'\left(I, \hat{I}'\right), S'\right) = 1 \tag{9}$$

where δ' and δ are sufficiently large thresholds. $\mathcal{D}'$ can be the same as, or different from, the decoding function $\mathcal{D}$. On the other hand, the following properties will hold for Alice's case:

$$C_{\delta}\left(\mathcal{D}\left(\hat{I}, I\right), S\right) = 1 \tag{10}$$

$$C_{\delta}\left(\mathcal{D}\left(\hat{I}', I\right), S\right) = 1. \tag{11}$$

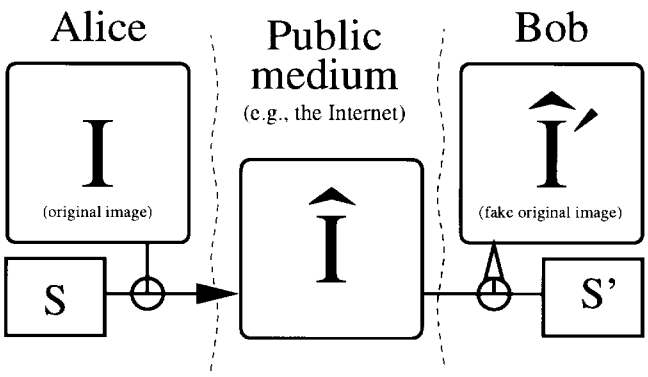

Fig. 3. Forging a watermark. Alice watermarks image I to get $\hat{I}$, which she makes public. Bob computes an image $\hat{I}'$ and watermark S', such that watermarking $\hat{I}'$ with S' yields $\hat{I}$ (SWICO-attack).

For the public watermarking scheme, we can simply replace $\mathcal{D}(X, Y)$ with $\mathcal{D}(X)$ in the above four properties.

Note that (10) is always true because the watermarked image is generated by Alice, while (11) states that Bob's fabricated "original" $\hat{I}'$ contains Alice's watermark S. This is to be expected if the watermarking technique employed by Alice is robust. However, (8) implies that Bob's watermark is also present in the watermarked image of Alice, and (9) states that Alice's original image I contains Bob's watermark S'! Bob can claim by virtue of properties listed in (8) and (9) that both the watermarked image $\hat{I}$ (Alice's watermarked image) and the image I (Alice's original) are but watermarked versions of his original $\hat{I}'$. Of course, Alice, by virtue of properties in (10) and (11), also claims Bob's images to be watermarked versions of hers. Note that Bob has not removed Alice's watermark. However, he has removed her claim of ownership, because every piece of evidence indicating that Alice is the originator of the image in question is matched by an equal piece of evidence that Bob is the originator.

Given only $\hat{I}$, we want to construct $C_{\delta'}, \mathcal{D}', \hat{I}'$ and S' such that the properties in (8) and (9) are satisfied. We shall call an instance of such attack which involves only one watermarked image, a **SWICO** (**S**ingle**W**atermarked**I**mage**C**ounterfeit**O**riginal) attack. It means that, in addition to (8) and (9), the counterfeit original $\hat{I}'$, will give back the same watermarked image $\hat{I}$ after the fake watermark S' is embedded, formally, as $\mathcal{E}(\hat{I}', S') = \hat{I}$. In principle, this can be achieved in a straightforward way, by *removing* a randomly selected watermark S' instead of embedding one. In other words, we identify some features in a watermarked image $\hat{I}$ and claim them to be our watermark S', which we remove from $\hat{I}$ to get our fake original image $\hat{I}'$. This scenario is depicted in Fig. 3.

More precisely, in the context of the feature based watermarking schemes described in Section II, the attacker constructs a counterfeit "original" image by extracting a chosen (possibly random) watermark S' from some feature set $D'(\hat{I}) = \{f_i'(\hat{I})\}$ to generate an image $\hat{I}'$ such that

$$f_i'\left(\hat{I}'\right) = f_i'\left(\hat{I}\right) \ominus s_i'. \tag{12}$$

The set $D'(\hat{I})$ of derived coefficients is assumed to remain more or less the same when the image is not perceptually degraded by an attacker.[3] The decoding scheme, operating on the counterfeit "original" $\hat{I}'$ and the true original I, first extracts $T' = \{t'_1, t'_2, \cdots\}$ as follows:

$$t'_i = f'_i(I) \ominus f'_i\left(\hat{I}'\right). \tag{13}$$

The confidence measure, taken to be the normalized correlation between T' and S', defined in (7), is then compared to the threshold δ'. Because of the robustness of the set $D'(\hat{I})$ against perceptually insignificant modification, we can expect that

$$f'_i(I) \approx f'_i\left(\hat{I}\right). \tag{14}$$

Combining (12)–(14), we have $t'_i \approx s'_i$, so that the correlation between T' and S' is large and implies that $C_{\delta'}(T', S')$ will most likely be equal to one. The attacker (Bob) can thus claim that the true original I contains his watermark S' and that I is a modified version of $\hat{I}'$. Conversely, the robustness of watermarking scheme[4] used to embed S onto I allows the true owner (Alice) to also argue that $\hat{I}'$ contains the watermark S. In other words, properties listed in (11) and (9) are satisfied. For the class of feature-based watermarking schemes that use a reference feature set derived from the original image I for decoding purpose, we shall show as follows this directly implies that (8) is also satisfied.

Let $W' = \{w'_1, w'_2, \cdots\}$ be the watermark extracted by the decoding scheme operating on the watermarked image $\hat{I}$ and the counterfeit "original" $\hat{I}'$: $w'_i = f'_i(\hat{I}) \ominus f'_i(\hat{I}')$. From (12), we have

$$w'_i = s'_i. \tag{15}$$

Thus (8) holds. This means that Bob's watermark will always be present in the watermarked image $\hat{I}$ provided by Alice, in addition to its presence in Alice's original image. We now have a scenario whereby rightful ownership cannot be resolved through an invisible watermarking scheme.

The key step in the counterfeiting technique is described in (12). By "subtracting off" a watermark S' in $\hat{I}$, we are essentially causing a watermark to be present in I, even when we do not have access to I.

We now show an example of how to achieve a counterfeit attack on the class of watermarking schemes whose encoding and decoding process rely on the set of derived coefficients $\{f_1(I), f_2(I), \cdots\}$. In terms of the *insertion* and *extraction* operators, we use simple *addition* "$+$" in place of $\oplus$ and *subtraction* "$-$" in place of $\ominus$.

[3] It is very important to note the difference between "perceptually similar" and "not perceptually degraded by an attacker." It is a mistake to conclude that any image "perceptually similar" to Alice's must contain her watermark, even if the scheme is robust. Alice can, for instance, watermark her image I with one mark to get J, and watermark I with a different mark to get K. J and K are perceptually similar, but neither image's watermark is present in the other. A robust scheme can only guarantee that an original watermark survives in a "perceptually similar" image if the relation between that image and the original does not rely on any information about the watermark itself.

[4] This is why any watermarking scheme has to be practically robust. Otherwise the attacker, armed with a more robust invisible watermarking scheme, will be able to substantiate his claim of ownership, while the true owner may totally lose his claim because his watermark may be virtually gone after the attack.

TABLE II
A SUMMARY OF THE CONFIDENCE MEASUREMENTS RECORDED FROM A TEST RUN ON SOME TYPICAL IMAGES. THE NUMBERS INDICATE THAT CONFIDENCE MEASUREMENTS ON EXTRACTED WATERMARKS MAY NOT PRODUCE SUFFICIENT EVIDENCE OF OWNERSHIP

Image	Test I		Test II	
	$S : \hat{I}/I$ (c_A)	$S' : \hat{I}/\hat{I}'$ (c_B)	$S : \hat{I}'/I$ (c_A^0)	$S' : I/\hat{I}'$ (c_B^0)
Baboon	32.76	32.77	22.17	22.78
Lena	32.77	32.76	21.35	21.65
Fighter	32.75	32.76	21.38	21.90
Lake	32.69	32.73	21.58	22.40
Peppers	32.22	32.08	20.60	20.82

Example 2: The special case when $D = D'$ and $\mathcal{D}(\ ) = \mathcal{D}'(\ )$, i.e., when the same decoding functions and set of derived coefficients are used in the original watermarking (from I to $\hat{I}$) and the generation of second "original" (from $\hat{I}$ to $\hat{I}'$).

Here, we have $f_i = f'_i$. The decoder $\mathcal{D}$, after comparing the original I and the fabricated "original" $\hat{I}'$, extracts $T = \{t_1, t_2, \cdots\}$, where $t_i = f_i(\hat{I}') - f_i(I) = f_i(\hat{I}) - s'_i - f_i(I) = s_i - s'_i$. Similarly, the decoder $\mathcal{D}'$, after comparing the fabricated "original" $\hat{I}'$ and the true original I, extracts $T' = \{t'_1, t'_2, \cdots\}$, where $t'_i = f'_i(I) - f'_i(\hat{I}') = f_i(I) - f_i(\hat{I}') = f_i(I) - f_i(\hat{I}) + s'_i = s'_i - s_i$.

Thus, $t_i = -t'_i$ and we can show that two correlations:

$$\frac{\sum_i t_i \cdot s_i}{\sqrt{\left(\sum_i t_i^2 \sum_i s_i^2\right)}} \quad \text{and} \quad \frac{\sum_i t'_i \cdot s'_i}{\sqrt{\left(\sum_i t_i'^2 \sum_i s_i'^2\right)}}$$

are identical.

We have illustrated an extreme case where using the same decoding function and using the same set of derived coefficients actually generate the same correlation values when both parties are trying to establish rightful ownership, which clearly cannot be resolved. We now give a more concrete example of this situation with respect to the watermarking technique proposed by Cox *et al.* [2]. □

Example 3: A successful implementation of the proposed attack on the watermarking scheme proposed by Cox *et al.* [2].

We implemented the algorithm described in [2], and then modified it to perform the inverse operation as described above. We used the same formula that Cox *et al.* used to insert a randomly generated watermark into the 1000 largest magnitude AC coefficients, v_i, of the image, yielding updated coefficients v'_i. To perform the inverse operation of identifying and removing a random watermark, this insertion formula was inverted to compute v_i as a function of v'_i, rather than the other way around[5].

Starting with an already watermarked image $\hat{I}$, and a watermark vector S', we computed a new "original" image

[5] A simple modification—for a 500-line C program, a single '*' was changed to a '/'.

TABLE III
A SUMMARY OF THE CONFIDENCE MEASUREMENTS RECORDED FROM 100 TRIALS ON TWO TEST IMAGES (ACCURACY RECORDED TO TWO DECIMAL PLACES). THE STATISTICS OF THE NUMBERS IN THESE TWO IMAGES ARE ALMOST IDENTICAL

Image	Test I						Test II					
	$S:\hat{I}/I$			$S':\hat{I}/\hat{I}'$			$S:\hat{I}'/I$			$S':I/\hat{I}'$		
	max	min	mean	max	min	mean	max	min	mean	max	min	mean
Baboon	33.72	29.93	31.68	33.70	30.01	31.63	24.24	19.44	21.98	24.37	19.44	21.94
Fighter	33.72	29.93	31.68	33.70	30.01	31.63	24.24	19.44	21.98	24.37	19.44	21.94

TABLE IV
A SUMMARY OF THE CONFIDENCE MEASUREMENTS RECORDED FROM A TEST RUN ON SOME IMAGES USING THE ATTACK DESCRIBED IN EXAMPLE 7

Image	k	k'	q				Quality of $\hat{I}'$	
			$\mathcal{D}(\hat{I},S,k)$	$\mathcal{D}(\hat{I},S',k')$	$\mathcal{D}(\hat{I}',S,k)$	$\mathcal{D}(I,S',k')$	PSNR	SNR
Lena	3	3	14.47	15.29	14.47	15.32	44.6	37.5
Peppers	3	4	10.18	14.25	9.86	14.25	40.3	35.0

$\hat{I}'$ (in reality a fake original) without any visible degradation of image quality. Using (6) as a measure of confidence of a watermark's presence in an image, the fabricated watermark S' is present in the original image I with a confidence value of 23.52, while the original watermark S is present in the fake original $\hat{I}'$ with a confidence value of 23.02. A summary of some other test runs of the attack, and on different images, are given in Table II. The notation is as follows: S: I_α/I_β denotes the confidence measurement of the presence of watermark S in image I_α with a reference image I_β used in the watermark decoding process. We also use the same notation in Fig. 2.

It should be noted that each set of the confidence measurements presented previously were recorded from one test run of the attack on a given set of 1000-element watermark sequences; the values in general vary slightly across different sets of watermark sequences. The aggregate statistics are nevertheless quite consistent. To demonstrate this, we tabulated in Table III the results of 100 random trials to insert "real" and "fake" watermarks into the DCT of the images and computed the maximum (*max*), minimum (*min*) and average (*mean*) of the correlation values recorded from the trials. In each trial, a pair of random 1000-element vectors whose components are normally distributed were inserted. Figs. 4 and 5 show the true original, the watermarked image, and a fake original from a test run on the images "Baboon" and "Peppers." Each set of the images are observed to be "perceptually similar." A similar attack can also be engineered on the scheme in [9] which is based on addition of a signal derived from the DCT of a pseudonoise sequence to the DCT of $k \times k$ image blocks. □

The attack described above is universal in the sense that any image watermarked by *any* scheme can be defeated. In the absence of *standardization* on the invisible watermarking techniques, or any specification of requirements on legitimate watermarking schemes, anyone can claim ownership of any watermarked image to which he or she has access. This is because no matter which scheme Alice uses to watermark her image, Bob can always use an invertible watermarking scheme $(\mathcal{E},\mathcal{D},\mathcal{C}_\delta)$ (such as the one in Example 3) to create a counterfeit original (that is, he uses $\mathcal{E}'$ to create a fake original $\hat{I}'$, and can then show that this image, when watermarked with S' using $\mathcal{E}$, will give the watermarked image $\hat{I}$ as in circulation) and proceed to argue that the unique ownership cannot be determined—thus Alice's claim of ownership is not validated based solely on the test of the presence of her invisible watermarks.

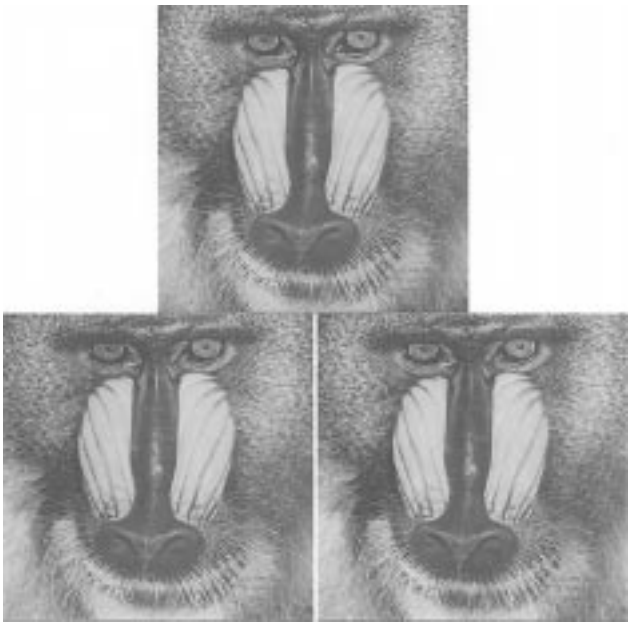

Fig. 4. Three "Baboon" images (from USC database). (Top) the watermarked image ($\hat{I}$) of the original with a 1000-element watermark sequence inserted. (Bottom left) The original image I. (Bottom right) The fabricated "original" image $\hat{I}'$. Measurements of the presence of watermarks in these images are presented in Table II.

V. NONINVERTIBLE WATERMARKING OF IMAGES

In the previous section we demonstrated how one can fabricate an "original" image from a watermarked one such that rightful ownership cannot be resolved. We have accomplished this by inverting a watermarking encoding function $\mathcal{E}$, in order to "remove" a watermark from an image rather than insert one. This is what we mean by an *invertible* watermarking scheme. Clearly, removing this invertibility is a crucial step in foiling our attack. In this section we shall formally define invertibility and noninvertibility of invisible watermarking schemes, discuss noninvertible schemes and provide an example. Noninvertible schemes should not be considered a cure-all, however: We shall demonstrate in Section VI that noninvertibility is a necessary but not sufficient condition for preventing further attacks.

Definition 1—Invertible Watermarking Schemes: A watermarking scheme $(\mathcal{E},\mathcal{D},\mathcal{C}_\delta)$ is **invertible** if, for any image $\hat{I}$,

Fig. 5. Three "Peppers" images (from USC database). (Top) the watermarked image ($\hat{I}$) of the original with a 1000-element watermark sequence inserted. (Bottom left) The original image I. (Bottom right) The fabricated "original" image $\hat{I}'$. Measurements of the presence of watermarks in these images are presented in Table II.

there exists a mapping $\mathcal{E}^{-1}$ such that 1) $\mathcal{E}^{-1}(\hat{I}) = (\hat{I}', S')$, 2) $\mathcal{E}(\hat{I}', S') = \hat{I}$, and 3) $\mathcal{C}_\delta(\mathcal{D}(\hat{I}), S') = 1$, where $\mathcal{E}^{-1}$ is a computationally feasible mapping, S' belongs to the set of allowable watermarks, and the images $\hat{I}$ and $\hat{I}'$ are perceptually similar. Otherwise, $(\mathcal{E}, \mathcal{D}, \mathcal{C}_\delta)$ is **noninvertible**.

$\mathcal{E}^{-1}$ is called the inverse mapping, and $(\hat{I}', S')$ is a member of the inverse image of $\hat{I}$ under $\mathcal{E}$. That is to say, a scheme $(\mathcal{E}, \mathcal{D}, \mathcal{C}_\delta)$ is invertible if a single member of the inverse image of $\hat{I}$ under $\mathcal{E}$ can be feasibly computed.

Lemma 1—Invertibility and SWICO Attack: An image watermarked by an invertible watermarking scheme $(\mathcal{E}, \mathcal{D}, \mathcal{C}_\delta)$ is susceptible to the SWICO attack using the same watermarking scheme.

Proof: Set $\mathcal{D}' = \mathcal{D}$ and $\delta' = \delta$. We then have $\mathcal{C}_{\delta'}(\mathcal{D}'(\hat{I}), S') = 1$ which satisfies (8). Assuming the robustness of the watermarking scheme, we have $\mathcal{D}'(I) \approx \mathcal{D}'(\hat{I})$. Therefore, $\mathcal{C}_{\delta'}(\mathcal{D}(I), S')$ will be sufficiently large and (9) is also satisfied.

Note that Definition 1 does involve the decoding function $\mathcal{D}$. For certain class of watermarking schemes, namely the private schemes, the invertibility definition may not involve the decoding function $\mathcal{D}$—instead, $\mathcal{C}_\delta(\mathcal{D}(I), S') = 1$ is implied by other conditions. The reasons for this have been described in Section IV, (12)–(15). The algorithm proposed by Cox *et al.* [2], for instance, is such an invertible watermarking scheme.

It seems that the SWICO counterfeit attack developed in the previous section can be foiled through more careful requirements for watermarking schemes. In particular, we can require that watermarking schemes used to establish rightful ownership must first satisfy the condition of noninvertibility. There are a number of ways to enforce this new requirement. One approach would be to use one-way functions in the watermark insertion process, so that it is not possible to extract a watermark from an image. The presence of a watermark in an image would have to be inferred by some sort of "trapdoor" function. The authors decided that this particular approach was undesirable, as not only would such an approach run the risk of producing a very complex and inflexible scheme, but it would offer little help in fixing currently existing schemes.

Another approach follows from noting that in order to fabricate a counterfeit original $\hat{I}'$, the attacker may have to choose a watermark S' *before or during* the construction of $\hat{I}'$. If we enforce the extra requirement that any watermark inserted into an image I be dependent upon (that is, a function of) I, we may make it difficult for the attacker to select a false watermark. Simply put, the attacker would have to compute a watermark S' dependent upon the final value of $\hat{I}'$, but that final value cannot be known until S' is used in its computation. This can be achieved by computing a bit sequence $B = \{b_1, b_2, \cdots\}$ from the image I via a one-way hash, that we then use in the process of watermarking I. One way to use this bit sequence is to select the labels s_i that compose the watermark itself. Another way is to use the bits to choose between two different insertion operations $\oplus$ and $\otimes$ for each s_i. If the one-way function is carefully designed such that two perceptually similar images with nonidentical pixel values will be hashed to vastly different bit sequences, the bit sequence obtained from a one-way hash function of the image to be watermarked will make the scheme secure against any trial and error attacks that one can attempt in order to construct a counterfeit original. With this consideration we can adapt some well-known secure one-way hash functions such as MD5 [17], MD4 [18], and SHA [19] for our purposes.

We will provide two such examples, both variants of the Cox *et al.* algorithm illustrated in Example 1. The first we present below, as an example of a noninvertible watermarking scheme. The second will be presented in Section VI, after we show that the first is *still susceptible to a refined version of our attack*, despite noninvertibility!

Example 4: A modified version of the scheme described by Cox *et al.* [2].

We first produce a 1000-bit one-way hash $\{b_1, b_2, \cdots, b_{1000}\}$ of the original image before computing its 2-D DCT. We then use two slightly different equations for inserting the watermark vector elements. For each frequency bin v_i to be modified, we choose one of the two formulas depending on the value of the hash bit b_i. The formulas are chosen to be different enough in their output that a watermark vector consisting of the same vector elements, but using a different 1000-bit hash string, cannot be recovered from the watermarked image.

Specifically, we use two versions of the second update formula in [2] as follows:

$$v_i' = \begin{cases} v_i(1 + \alpha s_i), & b_i = 0 \\ v_i(1 - \alpha s_i), & b_i = 1 \end{cases} \tag{16}$$

where in both cases α was chosen to be 0.1. A 1000-bit hash of the image is computed, and for each of the 1000 largest-magnitude AC coefficients, one of the formula is used depending on the value of the hash bit b_i.

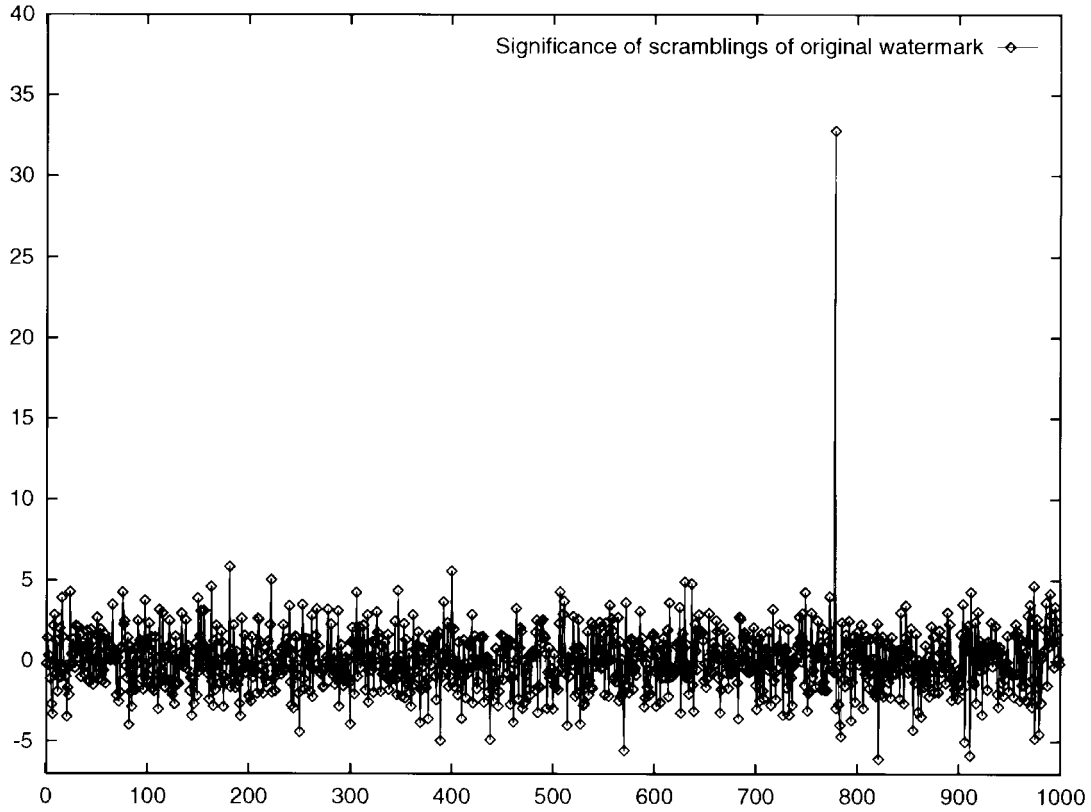

Fig. 6. Results of scrambling hash bits in watermark vectors. The original (as seen by the spike) and 999 copies with scrambled hashes.

Anticipating a possible attack involving rearranging watermark elements to match the required hash values, our scheme requires that the elements be embedded in the high-magnitude matrix elements in a left-to-right, top-to-bottom order. In addition, we impose the requirement that s_i be positive—otherwise, an attacker can simply negate certain watermark vector elements to match the resulting hash bits.

We applied this watermarking scheme to a test image. The original watermark S is applied 1000 times, once with an original 1000-bit hash string, and the other 999 times using randomly selected bit strings. The 1000 different watermarks are tested for presence in the image $\hat{I}$. The results are presented in Fig. 6. As illustrated in the figure, if the 1000-bit hash of the "original" hash string cannot be anticipated, the resulting watermark cannot be expected to have a high degree of presence and is useless for counterfeit attack purposes. □

Example 4 illustrates a scheme that is difficult to invert. Since (using our attack) the fake original image $\hat{I}'$ is not created until after the watermark S' is decided upon—indeed, $\hat{I}'$ depends on S'—the attacker cannot feasibly anticipate the associated hash string $\{b'_1, b'_2, \cdots\}$ that must then be incorporated into the watermark. We believe the proposed scheme is noninvertible although it seems difficult to prove this rigorously.

VI. Attacks with Multiple Watermarked Images

In the previous section we defined what it means for a watermarking scheme to be invertible, and showed an example of what we believe to be a noninvertible scheme that appears to circumvent our attack. In this section, we shall describe a variation of our attack (actually, a more general instance of our attack) that will nevertheless allow us to introduce a counterfeit watermark into an image under this noninvertible scheme. The results obtained in this section suggest that watermarking schemes must meet a stronger requirement than noninvertibility as defined in Section V for resolving ownership.

In the previous sections, we assume the existence of one and only one watermarked version of an image I in an ownership dispute. It is not unreasonable to assume that multiple watermarked versions of the same image I may indeed exist. After all, if we cannot find out which so-called "original" image is the "true original," how can we decide which watermarked version of an image is the "true watermarked" version in circulation? Bob is free to create as many watermarked versions with his own watermarks embedded after he has reverse-engineered a counterfeit original $\hat{I}'$ from Alice's watermarked version $\hat{I}$. Assuming Alice's watermark is robust enough, she can prove that all the watermarked versions Bob creates contains her watermark, as they are derived images of $\hat{I}$. To counter the claims, Bob has to prove the presence of his watermark in these watermarked images and the watermarked version of Alice, in addition to the presence of his watermark in Alice's original.

We shall illustrate this form of attack using two watermarked images, one created by Alice, $\hat{I}$, from her original I with her watermark S embedded that is the true watermarked version, and the other created by Bob, $\hat{I}''$, from his counterfeit original $\hat{I}'$ with his watermark S' embedded that is a fake watermarked

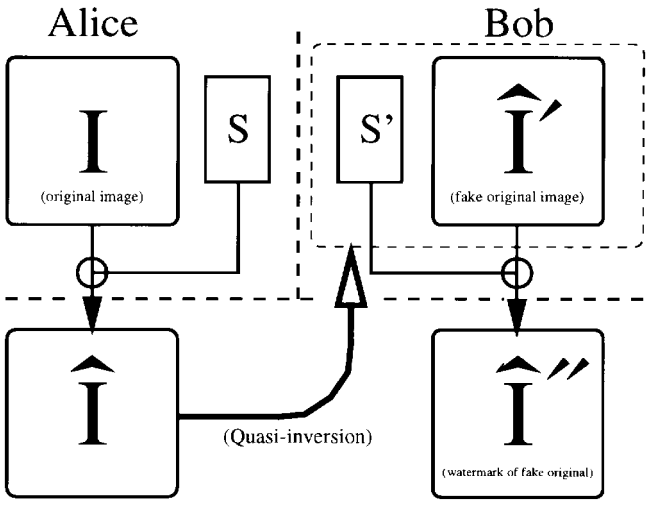

Fig. 7. Forging a watermark: the four-image case. If Bob can compute any image $\hat{I}'$ (of reasonable quality) and watermark S' such that S' is present in $\hat{I}$, even if watermarking $\hat{I}'$ with S' yields an image different from $\hat{I}$, the counterfeit-attack is still successful. (TWICO-attack).

image. Again, as in the SWICO attack, given only $\hat{I}$, we want to construct $C_{\delta'}, \mathcal{D}', \hat{I}'$, and S' such that the properties in (8) and (9) will hold; but unlike the SWICO attack, here we do not require that the insertion of the fake watermark S' onto the counterfeit original $\hat{I}'$ must produce the same watermarked image $\hat{I}$, but instead can be another watermarked image $\hat{I}''$. We call such a counterfeit-attack involving two watermarked versions of an image, a TWICO (**T**win **W**atermarked **I**mages **C**ounterfeit **O**riginal) attack. This is illustrated in Fig. 7. Similarly, we can generalize the attack to involve multiple (more than two) watermarked images.

Notice that the TWICO attack demonstrates an even more dangerous property: A watermarking scheme need not be invertible to be susceptible to such a counterfeit attack. That is, rather than having to compute an image $\hat{I}'$ and watermark vector S' such that marking $\hat{I}'$ with S' yields Alice's watermarked $\hat{I}$, Bob only needs to compute $\hat{I}'$ and S' such that marking $\hat{I}'$ with S' yields an image possibly different from but *similar* enough to $\hat{I}$, in terms of perceptual quality and the features preserved for watermark extraction, that S' can still be expected to lie in I. This leads us to define another type of watermarking scheme:

Definition 2—Quasi-Invertible Watermarking Schemes: A watermarking scheme $(\mathcal{E}, \mathcal{D}, \mathcal{C}_\delta)$ is ***quasi-invertible*** if, for any image $\hat{I}$, there exists a mapping $\mathcal{E}^{-1}$ such that 1) $\mathcal{E}^{-1}(\hat{I}) = (\hat{I}', S')$ and 2) $C_\delta(\mathcal{D}(\hat{I}), S') = 1$, where $\mathcal{E}^{-1}$ is a *computationally feasible* mapping, S' belongs to the set of allowable watermarks, and the images $\hat{I}$ and $\hat{I}'$ are perceptually similar. Otherwise, $(\mathcal{E}, \mathcal{D}, \mathcal{C}_\delta)$ is ***nonquasi-invertible***.

Lemma 2—Quasi-Invertibility and TWICO-Attack: An image watermarked by a quasi-invertible watermarking scheme $(\mathcal{E}, \mathcal{D}, \mathcal{C}_\delta)$ is susceptible to the TWICO attack using the same watermarking scheme.

Proof: Similar to the proof for Corollary 1. □

Notice that the definition of quasi-invertibility is similar to the definition of invertibility, the only difference being that a single constraint (i.e., that $\mathcal{E}(\hat{I}', S') = \hat{I}$) is relaxed. It follows that any invertible watermarking scheme is quasi-invertible.

Watermarking schemes that are noninvertible can still be susceptible to attacks (TWICO-attack) that eventually lead to ownership deadlocks, if they are quasi-invertible. We shall show that the modified scheme presented in Example 4 is an example of a noninvertible scheme that is nonetheless quasi-invertible, and still susceptible to counterfeit attack in a more complex form.

The problem with the scheme described in Example 4 is that the 1000 watermark elements are chosen independently of one another, allowing an attacker to "mix-and-match" elements from a number of watermarks already present in an image to create a new one that will also be present in the image. For instance, imagine that an image is watermarked twice using the Cox *et al.* algorithm, first with a watermark W, and then with a watermark T. Now imagine that the same image is instead watermarked with the following watermarks W' and T', constructed from a 1000-element bit string $a_1, a_2, \cdots, a_{1000}$:

$$W_i' = \begin{cases} W_i, & a_i = 1 \\ T_i, & a_i = 0 \end{cases} \quad T_i' = \begin{cases} W_i, & a_i = 0 \\ T_i, & a_i = 1. \end{cases}$$

Clearly, inserting the watermarks W' and T' into an image I yields the same result as inserting the watermarks W and T, since the same 1000 DCT elements are used for each insertion. In other words, the independence of the vector elements of a pair of watermarks allows us to compose a new watermark vector out of elements of vectors already present in an image: if W and T are already present in I, then arbitrarily chosen vectors W' and T' can be expected to be equally present in I. This is the basis of an attack that defeats the noninvertible scheme presented in Section IV.

In this attack, Bob simply constructs a 1000-element watermark W, extracts W from $\hat{I}$ using the 1000-bit string $1111\cdots1$ (that is, using the second insertion formula for every single watermark vector element) to form an image $\hat{I}_1'$, and similarly, constructs a second watermark T and extracts it using the string $0000\cdots0$ to form an image $\hat{I}_2'$. The two resulting images $\hat{I}_1'$ and $\hat{I}_2'$ are averaged to yield an image $\hat{I}'$, which he claims to be his original. This image is then hashed, and the real watermark S' is computed thus: If the ith hash bit of the average is zero, Bob uses the ith vector element of the second watermark; else he uses the ith element of the first, to build a 1000-element vector S. Finally, he watermarks $\hat{I}'$ with W' to produce a watermarked image $\hat{I}''$, which he claims to be the watermarked version of his "original." W' is then a vector composed of elements from W and T so as to match the 1000-bit hash of I.

Can such an attack work? After all, although watermarks using this spread-spectrum scheme have been shown [2] to be robust to many operations performed upon a marked image, it is assumed that the operations are not based on any specific information about the watermark. Here we are averaging one watermarked image with another whose watermark vector "points" in exactly the opposite direction. Can either watermark survive?

As Fig. 8 shows, they do survive, and Bob's mix-and-match method of constructing any watermark works astonishingly

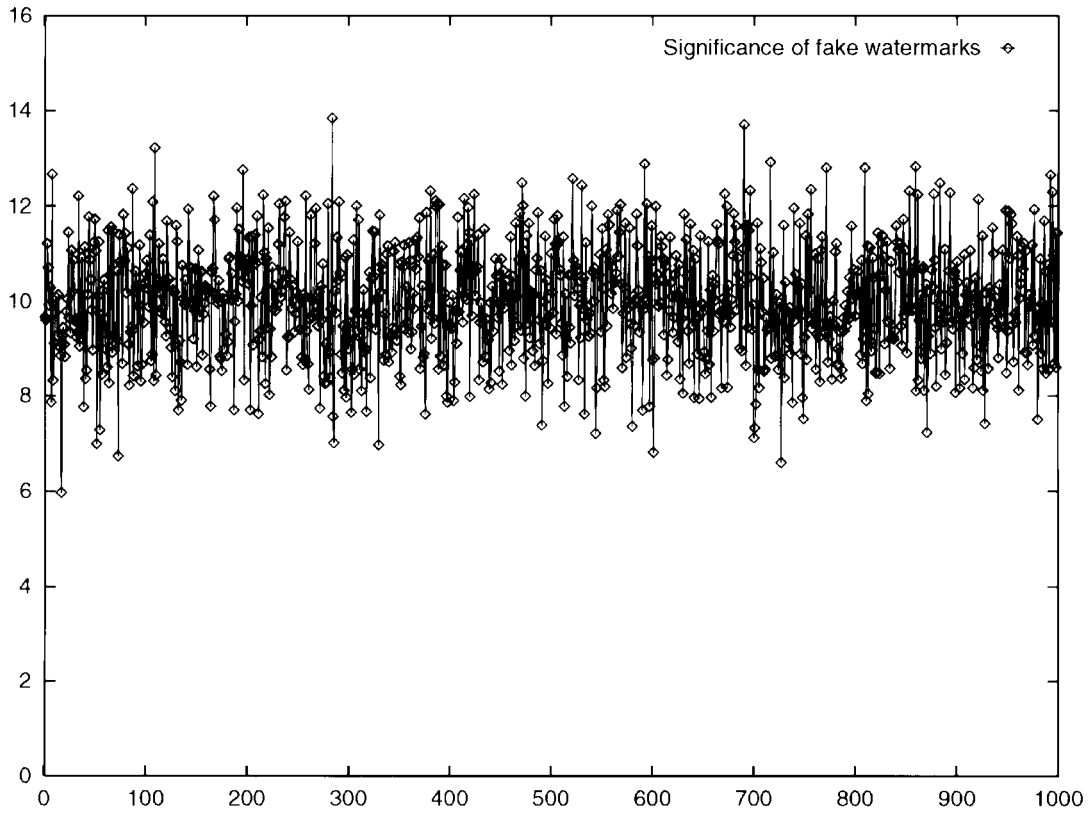

Fig. 8. Results of averaging oppositely watermarked images. Here we see 1000 watermarks picked to fit 1000 randomly chosen hash strings.

well. For this third experiment, a watermarked image was used to fabricate a fake original using this averaging technique, increasing the value of α in the formula from 0.1 to 0.27 so that the watermark exists more prominently, as well as causing a visible change in image quality. One thousand watermarks were created based on random bit strings, and each was tested for presence in the *original* (i.e., Alice's original) image. The mean correlation value was 9.97, well above random, and the highest computed correlation was 13.8. The original (Alice's) watermark is present in the fake original with a significance of 14.2. This is clearly too close for comfort.

What we have shown above is an example of a noninvertible watermarking scheme that, because of its quasi-invertibility, still falls prey to the counterfeit-attacks. Hence we have to require that a good scheme for resolving ownership must be practically nonquasi-invertible. While the previous example illustrates how an algorithm cannot be made secure merely by throwing a one-way function at it, this pitfall can actually be easily avoided by using another, simpler noninvertible watermarking scheme. Like the method posed in Example 4, this technique makes use of a bit string generated by a one-way hash of the image (and perhaps some extra information). However, the string is used in a different way that eliminates the potential for mixing and matching. This scheme also allows us to lighten up some the restrictions we placed on the order and sign of the watermark vector elements.

Example 5: A more secure modified version of the scheme described by Cox *et al.* [2].

We first produce an n-bit $\{b_1, b_2, \cdots b_n\}$ one-way hash of the original image, perhaps combined with a relatively small standard identifier in order to allow multiple distinct watermarks to be inserted into the same image. We then simply use this hash as the seed for the pseudorandom number generator used to generate the normally distributed watermark vector elements. A vector, then, is not considered an allowable watermark unless it could have been generated by a given starting seed for the (possibly standardized) generator.

An attacker would need to be able to compute a fake original $\hat{I}'$ and a vector S' such that a one-way hash of $\hat{I}'$ and some allowable identifier will be a seed which generates S'. Notice that any mix-and-match variant of our attack is rendered completely useless by this technique, since arbitrarily constructing a new watermark S' from already existing watermarks S and T (indeed, using *any* method of constructing a vector other than an allowable pseudorandom number generator) is very unlikely to produce an allowable watermark. Even then, determining the seed that generates an arbitrary allowable watermark is hard.

The originator of the image would, when called upon to prove ownership, present the original image, the identifier if one is used, and the generator if no single standard generator is picked. Clearly, some restriction on the choice of generator would be sensible. Notice that the actual watermark sequence does not need to be stored, since it is entirely dependent upon the image and identifier. We believe this scheme to be nonquasi-invertible, and secure against the different types of attacks presented in this paper. □

In conclusion, noninvertibility of a watermarking scheme is necessary to prevent our counterfeit fabrication attack. It may seem obvious in retrospect that noninvertibility is necessary for a watermarking scheme to be secure, especially when the problem of watermarking is approached from a cryptographic point of view. However, until now the issue has not been addressed, the primary focus of research falling upon engineering *robustness* alone, under the assumption that security would be guaranteed.

But noninvertibility is not sufficient unless "invertibility" is taken in a very general sense: an attacker does not need to be able to compute a fake original $\hat{I}'$ as an unwatermarked version of $\hat{I}$; he or she merely needs to compute an $\hat{I}'$ to be an unwatermarked version of *some* images similar enough to $\hat{I}$ that the fabricated watermark can be found in I. Designers of noninvertible schemes must therefore take extra care to ensure that theirs are nonquasi-invertible as well. This is a surprising development even from a cryptographic perspective—making a watermark encoding scheme one-way is not enough to foil the watermark forging techniques described herein. Moreover, proving that a scheme meets the more general requirement of nonquasi-invertibility may be difficult. We hope to promote discussions of the potential pitfalls of engineering nonquasi-invertibility.

VII. Attacks on Public Watermarking Schemes

We have addressed invertibility and quasi-invertibility in the previous sections. Their definitions, and the definition of invisible watermarking scheme as discussed in Section II are general enough to include a variety of invisible watermarking schemes. However, as we have used as an example of counterfeit attack on the private watermarking scheme proposed by Cox *et al.* [2], in which an original image is necessary for the process of watermark extraction and decoding, some may have the misconception that if the extraction of the embedded watermark does not involve explicit use of an original image, the counterfeit attack may not have been succeeded. Because the extraction operation involves literally a subtraction "$-$" operation with respect to a reference image, a question to ask is if one can take out the need of the reference in the extraction process and how can the counterfeit watermarking scheme be implemented? After all, there are many invisible watermarking techniques that do not involve the use of the original image (for example, [14], [5]) in the extraction of embedded watermarks.

The answer is, the same principle still holds. Indeed, such counterfeit attacks can still be engineered. We shall show an implementation using the technique proposed by Pitas [14].

Example 6: A summary of the public watermarking scheme proposed by Pitas [14] (similar scheme is also described in [7]).

The watermark insertion procedure is as follows: Given an image I, divide the set of pixels into two equal sets A and B (via a random selection of the two sets). The division strategy S (that is, how to divide into the two sets A and B) is also the watermark in this case. To yield the watermarked image, k is added to each pixel in A. Both S and k are needed for proper decoding. In practice $k = 3, 4$ or 5. To detect the watermark from an image $\hat{I}$, based on watermark S, we compute $\overline{w} = \overline{A} - \overline{B}$, where $\overline{A}$ and $\overline{B}$ are the mean pixel values of the pixels in sets A and B respectively.

If $\hat{I}$ is watermarked using S and with value k, then $\overline{w} \approx k$. Otherwise, if S is randomly chosen, $\overline{w} \approx 0$ for a typical image. The test statistics will then indicate the confidence level of the absence or presence of a watermark. Here the test statistics is $q = \overline{w}/s_{\overline{w}}$, where $s_{\overline{w}}$ is the sample standard deviation of $\overline{w}$. It has a distribution with mean $k/s_{\overline{w}}$ or 0, depending on whether there is a watermark presence or not in $\hat{I}$. We shall write $q = \mathcal{D}(\hat{I}, S, k)$ to denote the computation of test statistics q using watermark S and addition k on image $\hat{I}$. □

The above watermarking scheme does not involve an extraction operation with respect to a feature set from a reference image in the decoding of the watermark. In such cases, an inverse function $\mathcal{E}^{-1}$ can be constructed straightforwardly to yield a counterfeit "original" $\hat{I}'$ such that watermarking $\hat{I}'$ with some S' and k returns $\hat{I}$. This, however, does not directly imply the success of counterfeit attacks. A crucial requirement for the watermarking scheme to be invertible, and susceptible to a successful attack, is the presence of the attacker's (Bob's) watermark on the watermarked image $\hat{I}$ which has the watermark of the true owner—Alice—embedded, as well as on the true original I. The challenge for Bob is to reconstruct his watermark directly from, and without changing, the pixel values of Alice's watermarked image; this time, he cannot take advantage of a reference such as the counterfeit original to do a subtraction. In this case, we shall show in the following example that Bob can engineer a division strategy S' which effectively demonstrates the presence of his mark in Alice's watermarked image. The attack is summarized in the following example.

Example 7: An implementation of a counterfeit attack on the watermarking scheme in Example 6.

A successful attack requires the attacker Bob to come up with a division strategy S', and k' such that $q = \mathcal{D}(\hat{I}, S', k') \approx \mathcal{D}(\hat{I}, S, k)$. If S' is randomly selected, then $\mathcal{D}(\hat{I}, S', k') \approx 0$. To overcome this, Bob first chooses the *best* S' that forms two sets, A^* and B^*, such that the difference $\overline{w}$, of the means A^* and B^*, is maximized. This in return will also make q a significantly large number. The optimal strategy is as follows: 1) compute the median m of the pixel values of $\hat{I}$; 2) assign all pixels whose values are greater than m to A^* and those smaller than m to B^*; and 3) randomly assign pixels whose values are equal to m to A^* or B^* until the sizes of A^* and B^* are equal. These steps produce a large $\overline{w}'$ too unrealistic in typical images (for example, $\overline{w}' \approx 87$ for the image Lena—k embedded commonly have values range from 2–5). In addition, the two selected sets A^* and B^* are not *random-looking*, contrary to the common practice that the watermark S' is usually randomly chosen. Bob, however, can introduce randomness into the two sets A^* and B^* by *randomly* swapping some fraction l of pixels between the two sets. An attacker can start by small l and slowly increase l to get the desired $\overline{w}$—this is possible because increasing l decreases $\overline{w}$. In our experiments, $l \approx 0.5$. We denote the resulting sets A' and B' which constitutes the division strategy

S'. The counterfeit original $\hat{I}'$ can then be easily constructed by subtracting out k' from the pixel values on the pixels in A'. Thus watermarking $\hat{I}'$ with k' and set partitioning of A' and B' will give the watermarked image $\hat{I}$.

Experimental results show that the sets partitioned by S' are random without any visible correlation, and the confidence measurements on the presence of Bob's watermark on Alice's original and watermarked image are in line with, or better than, Alice's measurements of her watermark presence in Bob's counterfeit original and her own watermarked image. In other words, $\mathcal{D}(\hat{I}, S', k') \approx \text{or} > \mathcal{D}(\hat{I}, S, k)$ and $\mathcal{D}(I, S', k') \approx \text{or} > \mathcal{D}(\hat{I}', S, k)$ and the attack is successful. Table IV summarizes the confidence measurements in terms of q on the presence of watermarks in two test images with one test run together with image quality measurements of $\hat{I}'$ against I. The numbers vary slightly across different sets of S and S'. There are no observable artifacts in the image quality. Note that Bob can use a k' different from the k Alice uses, or the same. Both cases are illustrated in the table. In our trials, the entire attack on a 512 $\times$ 512 image took up less than 1 s on a 100-MHz PC with 32 MB of memory. □

In other words, the watermarking schemes presented in Example 6 are invertible. There are, of course, remedies that can be deployed to make the attack more difficult, like imposing stringent requirements on how the sets can be partitioned. One method is to use similar solution suggested in Example 5 for private watermarking scheme: Use a one-way hash of the original image (possibly with combinations of the owner's identification) as the seed to generate the set A and B. This will make the attack more difficult, but it has not yet been proven impossible. Nonetheless, what we have illustrated is that even the public watermarking schemes, if not designed carefully and used in the proper context (in our case, using invertible watermarking schemes for applications involving ownership resolution), may not deliver what they have promised. However, we find it difficult to have one general and systematic scheme to attack a wide variety of other public watermarking schemes; rather, a successful attack on any particular scheme or any class of watermarking schemes is a piece of smart engineering work by itself.

VIII. Discussions and Conclusions

We now return to the questions that we posed in the beginning. What is a watermark, why is it needed and how useful is it? Current copyrighting mechanisms for photographs and images involve registration of the item being copyrighted with a centralized authority.[6] All contests of ownership are then resolved by this central authority. It has been recognized for quite some time now that these laws are quite inadequate for dealing with digital data that can be so easily copied and manipulated. This has led to an interest within the research community for developing copyright protection mechanisms. One such effort was aimed at developing watermarking techniques for digital data. A watermark in this context is a signal added to digital data such that it could be used to: 1) identify source of the data or uniquely establish ownership; 2) to identify its intended recipient; and 3) to check if the data has been tampered with. Within each class of applications, there can be variations on the requirements of a watermarking scheme.

It may have appeared from some of the ensuing work that the most important property of such watermarking schemes was their robustness. That is their ability to survive despite malicious attempts at removal. Indeed, in this sense the research efforts have been successful. Watermarking schemes have been proposed and shown to be remarkably robust. However, we have demonstrated in this paper that the ability to embed robust watermarks in digital images does not necessarily imply the ability to establish ownership, unless certain requirements are imposed on the watermarking schemes. In the absence of such requirements, Alice cannot simply lock away an original image that she can use later to establish ownership over a watermarked copy. So what can Alice do? She can still resort to conventional means of registering the image with a central authority and obtaining a copyright.

Contrary to common belief that digital watermarking is a cure-all solution for copyright protection, we have shown that there are certain limitations to what digital watermarking techniques can achieve. Some applications, like resolving rightful ownership, cannot rely on using watermarking techniques that are invertible or quasi-invertible. But then, proving that a technique is not quasi-invertible could turn out to be very difficult.

Despite these limitations, watermarking techniques (invertible or otherwise) can still be useful for protecting Alice's interests. For example, Alice can embed a different watermark in each copy of the image that she sells. This unique watermark will enable her to determine the identity of her specific customer who may be making unauthorized copies and selling them for a profit. A watermark can also be used by Alice to establish her ownership over versions of her image that have been visually modified. For example, Alice can establish that it is her image that is embedded in a larger image. Current copyrighting mechanisms are not well geared for addressing such situations, that in the future can easily arise, given the ease by which data in digital form can be manipulated and the widespread use of the Internet in rapid dissemination of digital information. One can list many more variants of such applications demonstrating the utility of invisible watermarks. However, different applications demand different types of watermarking schemes with different requirements.

In addition to the counterfeit attacks introduced in this paper, we can study watermarking schemes that are robust against general attacks (such as that proposed in [16]) to *remove* or *diminish the presence of* the watermark in the watermarked images. On the other hand, it is also important to investigate cryptographic protocols that can complement watermarking schemes for effective digital copyright protection. Finally, a side note on the types of attacks introduced in this paper:

[6] We quote several sentences in [20] here: In U.S. copyright laws, copyright protection is secured automatically when the work is "created," and a work is "created" when it is fixed in a copy or phonorecord for the first time. No publication or registration or other action in the Copyright Office is required to secure the copyrights. There are, however, certain definite advantages to registration.

These attacks easily allow any one to claim ownership of any images he or she has access to, whether those images have been watermarked or not! The unfortunate fact is that unwatermarked images will fall prey to false ownership claims by someone exploiting the attacks. How to protect these unwatermarked images against deliberate attacks is an issue worthy of further research.

In spite of the promises of digital watermarking, we have to think more carefully at the application end before we propose and adopt yet another watermarking scheme. In other words, the commonly asked questions, such as how to hide marks more invisibly and how to hide marks more robustly must be asked alongside questions like, "for what purposes can this watermarking technique be used?" "in what ways can this watermarking technique be attacked?" and "for what reasons should this watermarking scheme be trusted to deliver on its promises?"

REFERENCES

[1] S. Craver, N. Memon, B. L. Yeo, and M. M. Yeung, "Can invisible watermarks resolve rightful ownerships?," in *Proc., SPIE Storage and Retrieval for Still Image and Video Databases V*, vol. SPIE 3022, Feb. 1997, pp. 310–321; also IBM Tech. Rep. RC 20509, July 25, 1996.

[2] I. J. Cox, J. Kilian, T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for images, audio and video," in *IEEE Int. Conf. Image Processing*, vol. 3, pp. 243–246, 1996.

[3] W. R. Bender, D. Gruhl, and N. Morimoto, "Techniques for data hiding," in *Proc. SPIE: Storage and Retrieval of Image and Video Databases*, vol. 2420, Feb. 1995, pp. 164–173.

[4] R. G. van Schyndel, A. Z. Tirkel, and C. F. Osborne, "A digital watermark," in *Proc., IEEE Int. Conf. Image Processing*, vol. II, 1994, pp. 86–90.

[5] J. Zhao and E. Koch, "Embedding robust labels into images for copyright protection," in *Intellectual Property Rights and New Technologies; Proc. KnowRight'95 Conf.*, pp. 241–251.

[6] F. M. Boland, J. J. K. O'Ruandaidh, and C. Dautzenberg, "Watermarking digital images for copyright protection," in *Proc. IEE Image Processing Applications Conf.*, 1995, pp. 326–330.

[7] N. Nikolaidis and I. Pitas, "Copyright protection of images using robust digital signatures," in *Proc., IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 4, May 1996, pp. 2168–2171.

[8] G. Langolaar, van der Lubbe, and J. Biemond. (1996). Copy protection for multimedia data based on labeling techniques. Available WWW: http://www-it.et.tudelft.nl/pda/smash/public/benelux_cr.html.

[9] M. Swanson, B. Zhu, and A. Tewfik, "Transparent robust image watermarking," in *Proc. Int. Conf. on Image Processing 1996*, vol. 3, pp. 211–214, Sep.

[10] R. Wolfgang and E. Delp, "A watermark for digital images," in *Proc. Int. Conf. on Image Processing 1996*, vol. 3, pp. 219–222.

[11] Digimarc corporation. (1997). Available WWW: http://www.digimarc.com/

[12] G. B. Rhoads, "Steganography methods employing embedded calibration data," U.S. Patent 5 636 292, June 1997.

[13] I. J. Cox and M. L. Miller, "A review of watermarking and the importance of perceptual modeling," in *Proc. SPIE Human Vision and Elect. Imaging II*, vol. SPIE, vol. 3016, Feb. 1997.

[14] I. Pitas, "A method for signature casting on digital images," in *Proc. Int. Conf. on Image Processing 1996*, vol. 3, pp. 215–218.

[15] G. K. Wallace, "The JPEG still picture compression standard," *Comm. ACM*, vol. 34, pp. 30–44, Apr. 1991.

[16] H. S. Stone, "Analysis of attacks on image watermarks with randomized coefficients," Tech. Rep., NEC Res. Inst., May 1996.

[17] R. Rivest, "The MD5 message digest algorithm," Internet Rep., RFC 1321, Apr. 1992.

[18] R. Rivest, "The MD4 message digest algorithm," in *Advances in Cryptology, CRYPTO 92*. New York: Springer-Verlag, 1991, pp. 303–311.

[19] National Institute of Standards and Technology, "Secure Hash Standard," NIST FIPS Pub 180–1, Apr. 1995.

[20] U.S. Copyright Office Circular 1. (Mar. 1996). Available WWW: http://lcweb.loc.gov/copyright; Available Gopher: marvel.loc.gov.

Scott A. Craver received the B.S. and the M.S. degrees in computer science from Northern Illinois University, DeKalb, in 1994 and 1995, respectively, where he is currently working towards the M.S. degree in mathematics at Northern Illinois University.

He was an intern at the IBM Thomas J. Watson Research Center, Yorktown Heights, NY, from May 1997 to December 1997. He is now an intern at Intel Labs, Santa Clara, CA. His research interests include cryptography, natural language processing, and theory of computation.

Nasir Memon received the B.E. degree in chemical engineering and the M.Sc. degree in mathematics from the Birla Institute of Technology, Pilani, India, in 1981 and 1982, respectively. He received the M.S. and Ph.D. degrees in computer science from the University of Nebraska, Lincoln, in 1989 and 1992, respectively.

He is currently on leave of absence as an Assistant Professor in the Computer Science Department at Northern Illinois University, DeKalb, and is visiting the Imaging Technology Department at Hewlett Packard Laboratories, Palo Alto, CA. His research interests include data compression, data encryption, multimedia data security, and communications networks.

Boon-Lock Yeo (S'95–M'96) received the B.S.E.E. degree in electrical engineering from Purdue University, West Lafayette, IN in 1992, and the M.A. and Ph.D. degrees from Princeton University, Princeton, NJ, in 1994 and 1996, respectively.

He was with C-Cube Microsystems, San Jose, CA, in 1990, Siemens Corporate Research, Princeton, NJ, in 1993, and Mathematical Sciences Research Center, AT&T Bell Laboratories, Murray Hill, NJ, in 1995. In December 1995, he joined IBM Thomas J. Watson Research Center, Yorktown Heights, NY, as a Research Staff Member and later became Manager of the Visual Computing and Communications Department. Since January 1998, he has been with Intel Research Labs, Santa Clara, CA, managing the Video Technology Department. His research interests include signal and image processing, data compression and communications, visualization, computer vision, and problems related to multimedia information systems.

Dr. Yeo was the recipient of the Wallace Memorial Fellowship in Engineering 1995–1996, the IBM Graduate Fellowship in 1994–1995, and was a Singapore Technologies Overseas Scholar in 1989–1991. He was the recipient of the 1996 IEEE Circuits and Systems Society Video Transactions Best Paper Award and is currently serving as guest editor of a special issue of Computer Vision and Image Understanding on Computer Vision Applications for Network-Centric Computing.

Minerva M. Yeung (S'90–M'96) received the B.S. degree in electrical engineering from Purdue University, West Lafayette, IN, in 1992 and the M.A. and Ph.D. degrees in electrical engineering from Princeton University, Princeton, NJ, in 1994 and 1996, respectively.

She was with the Imaging Department, Siemens Corporate Research, Princeton, NJ, in 1993 and IBM Research in 1995. From August 1996 to December 1997, she was a Research Staff Member in the Image Applications group at the IBM Thomas J. Watson Research Center, Yorktown Heights, NY. She joined Intel Research Labs, Santa Clara, CA, in January 1998. Her research interests lie in the areas of image processing, video processing and presentation, database retrieval, multimedia data security and watermarking, communications, and digital imaging.

Dr. Yeung was a C. W. Chu Foundation Scholar from 1988 to 1992. In 1994–1995, she was a Sony Graduate Fellow, and was an Intel Foundation Graduate Fellow in 1995–1996.