RC 20509 (July 25, 1996)
Computer Science/Mathematics

IBM Research Report

Can Invisible Watermarks Resolve Rightful Ownerships?

Scott Craver

Department of Mathematics
Northern Illinois University
DeKalb, IL 60115
Email: caj@niu.edu

Nasir Memon

Department of Computer Science
Northern Illinois University
DeKalb, IL 60115
Email: memon@cs.niu.edu

Boon-Lock Yeo

IBM Research Division
T.J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598
Email: yeo@watson.ibm.com

Minerva Yeung

Department of Electrical Engineering
Princeton University
Princeton, NJ 08544
Email: mingy@ee.princeton.edu

IBM Research Division
Almaden · T.J. Watson · Tokyo · Zurich

Can Invisible Watermarks Resolve Rightful Ownerships?

Abstract

Digital watermarks have been proposed in recent literature as the means for copyright protection of multimedia data. In this paper we address the capability of invisible watermarking schemes to resolve copyright ownerships. We will show that rightful ownerships cannot be resolved by current watermarking schemes alone. In addition, in the absence of standardization of watermarking procedures, anyone can claim ownership of any watermarked image. Specifically, we provide counterfeit watermarking schemes that can be performed on a watermarked image to allow multiple claims of rightful ownerships. We also propose *non-invertible* watermarking schemes in this paper and discuss in general the usefulness of digital watermarks in identifying the rightful copyright owners. The results, coupled with the recent attacks on some image watermarks, further imply that we have to carefully re-think our approaches to invisible watermarking of images, and re-evaluate the promises, applications and limitations of such digital means of copyright protection.

Keywords: Authentication of copyright ownerships, invisible watermarks, copyright protection of images, attacks on watermarking schemes, non-invertible watermarking techniques.

1 Introduction

The rapid growth of digital imagery has called upon the needs for effective copyright protection tools. Various watermarking schemes and software products have been introduced recently in an attempt to address this growing concern. It is natural to ask a few questions regarding all these efforts: (1) What is a digital watermark? (2) Why are digital watermarks necessary, or in other words, what can digital watermarks achieve, or fail to achieve? (3) How useful are digital watermarks, or in other words, what can digital watermarks do for copyright protection in addition to current copyright laws, or current avenues of resolving copyright grievances?

In general, there are two types of digital watermarks (signatures) addressed in existing literature: visible and invisible watermarks[1]. These watermarks are developed mainly for two purposes: copyright protection and data authentication. In this paper we shall focus on the large class of invisible watermarks developed for one instance of copyright protection — that is, to identify the rightful owner. In this case the ownership labels which are embedded in an image have to be recoverable despite intentional or unintentional modification of the image. This means that such labels (and thus the corresponding labeling techniques) should ideally be robust against normal image processing operations like filtering, requantization, dithering, scaling, cropping, etc. and common image compression like JPEG image compression standards. They must also be invulnerable to deliberate attempts to forge, remove or invalidate labels.

Existing invisible watermarking schemes for copyright protection have been reported in research literature (for example, see [1, 2, 3, 4]). Unfortunately, many of these schemes did not address the *ends* of invisible watermarking schemes. They instead focused on the robust *means* to mark an image invisibly. In doing so, the concerns brought up by the previous three questions may not be properly and clearly addressed. We will show in this paper that current invisible watermarking schemes cannot resolve rightful ownership of any image watermarked with multiple signatures (labels). In addition, without any standardization of watermarking techniques or specification of certain requirements in the watermarking procedures (that is, without properly answering the question "What is a

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

digital watermark?"), we shall show that anyone can claim ownership of an image by simple methods described in later sections. The results, coupled with the recent attacks on some of the image watermarks reported in [5], further suggest that we have to carefully re-think our approaches to invisible watermarking of images, and re-evaluate 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: we shall give the general definitions and formulations of digital watermarking schemes in Section 2. In Section 3, we discuss how digital watermarking can be used to resolve rightful ownerships, and depict a scenario in which there may be more than one "rightful" owners of an image. We then show in Section 4 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 ownerships. An implementation of such a scheme, which is used to invalidate the watermarking method proposed by [1] is also described. In Section 5 we present the *non-invertible* watermarking schemes as a method of preventing the type of attack described in Section 4. We conclude in Section 6 with a discussion on the use of watermarking schemes for the authentication of rightful ownerships.

At this point we would like to emphasize that we still believe invisible watermarks are important to the information infrastructure, with applications that include determining rightful copyright ownerships. However, resolving rightful ownerships of digital images may require, in addition to invisible watermarks, the inclusion of protocols, formal requirements and standardization similar to traditional legal channels that are currently used to copyright images and photographs. Through this paper, we hope to promote new discussions and interests in the research community on the applications and values, as well as limitations, of digital signatures and the corresponding watermarking techniques.

2 Watermarking of Images: Definitions and Formulations

In this section we shall give a generalized formulation of invisible watermarking schemes. We shall define in general terms the process of signature insertion into an image and the use of invisible watermarks to determine the ownership of a watermarked image. Figure

1 illustrates both the encoding process in which a signature is inserted into an image, and the decoding process in which a signature is recovered and then compared to the inserted signature.

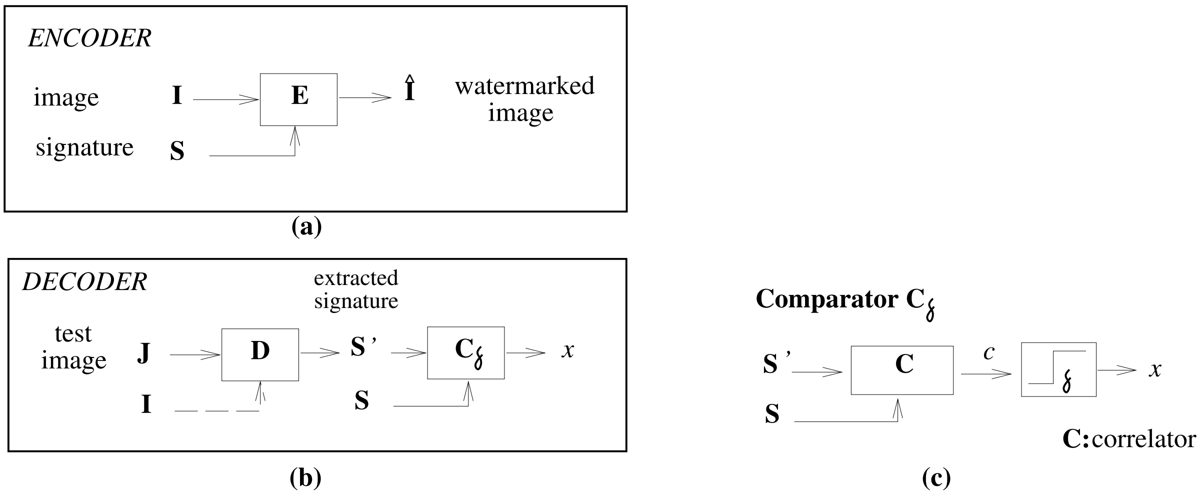

Figure 1: Encoding and Decoding embedded signatures in an image: (a) encoder (b) decoder (c) comparator

Here we denote an image by I, a signature $S = \{s_1, s_2, \cdots\}$ and the watermarked image by $\hat{I}$. $\mathcal{E}$ is an encoder function if it takes an image I and a signature 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 signature S is dependent on the image I. In such cases, the encoding process described by (1) still holds. A diagram of the encoding process is shown in Figure 1(a).

A decoder function $\mathcal{D}$ takes an image J (J can be a watermarked or un-watermarked image, and possibly corrupted) whose ownership is to be determined, and recovers a signature S' from the image. In this process, an additional image I can also be included which is often the original (and un-watermarked) version of J. This is due to the fact that some encoding schemes may make use of the original images in the watermarking process to provide extra robustness against intentional and unintentional corruption of pixels.

$$\mathcal{D}(J, I) = S'. \tag{2}$$

The extracted signature S' will then compared with the owner signature sequence by a comparator function $\mathcal{C}_\delta$, and a binary output decision is generated. It is a 1 if there is a match and 0 otherwise:

$$\mathcal{C}_\delta(S', S) = \begin{cases} 1, & c \geq \delta; \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Here, c is the correlation of the two signatures. A diagram of the decoding process is shown in Figure 1(b), and the comparator is depicted in Figure 1(c). Without loss of generality, watermarking schemes can be treated as a three-tuple $(\mathcal{E}, \mathcal{D}, \mathcal{C}_\delta)$.

The above framework describes what an invisible watermark is and how it can potentially be used to determine ownership. This is a generalized formulation. It does not give any insight into how exactly a watermarking scheme works. In view of this, here we specify some watermarking schemes. In particular, we describe the formulation of a class of invisible watermarking schemes, which we call *feature-based* watermarking schemes, that embed a signature $S = \{s_1, s_2, \cdots\}$ into some set of derived *features* $D(I) = \{f_1(I), f_2(I), \cdots\}$. The embedding process is achieved by an *insertion operation* which we denote by the symbol $\oplus$. That is,

$$f'_i = f_i \oplus s_i.$$

The insertion operation has an inverse operation, namely the *extraction operation*, which we denote by $\ominus$. That is,

$$f'_i \ominus f_i = s_i.$$

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.

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 features would be transformed domain (e.g., DCT, wavelet) coefficients which 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. *[1].*

In this scheme, the 2D DCT is taken of the image I and the set $D(I)$ corresponds to the n AC DCT coefficients of highest magnitude. Such coefficients will typically correspond to low frequency ones. Significant modification made to the image will cause image fidelity to degrade before the watermark does.

The encoder $\mathcal{E}$ takes a signature S and places it in the set $D(I)$. An inverse 2D DCT is taken of this modified matrix, yielding the watermarked image $\hat{I}$. To determine if a given image J contains the signature S, the decoder $\mathcal{D}$ first extract $T = \{t_1, t_2, \cdots\}$ from J as follows:

$$t_i = f_i(J) - f_i(I). \tag{4}$$

The confidence measure c is then taken to be the

$$c = \frac{\sum_i t_i.s_i}{\sqrt{\sum_i t_i^2}} \tag{5}$$

Alternatively, the normalized correlation

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

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 but smaller than 1.

□

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 ownerships and to bring up the different issues of the application of digital watermarks in resolving rightful ownerships.

3 Resolving Rightful Ownerships by Invisible Watermarks

It is a common view that invisible watermarking schemes may be used to protect the rights of copyright owners of images: at the very least, the labels (in other words, the digital signatures) extracted from the watermarked images can be used to identify the rightful owners. But how can we do this? Does it mean straight-forwardly that the one whose signature matches the embedded signature extracted from an image will automatically be the rightful owner of the image?

> *Suppose Alice and Bob 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 a label extracted from a watermarked image matches the particular signature label of Alice, then the image is believed to belong to her. Similarly, if the label matches Bob's signature, then it must be his image. If a watermarked image contains both Alice and Bob's signatures, whose image is it?*
>
> *Suppose now that Alice and Bob use different watermarking techniques. 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 never 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 signatures necessary if copyright laws can fully protect the interests of the copyright owners?) one can always look to these original images for an answer.

Now there is one watermarked image from which the digital signatures of both Alice and Bob have been extracted and both of them are claiming to be its rightful owner. Alice can ask Bob for his original image and check if it contains her signature. Similarly, Bob can ask Alice for her original image and check for his signature. If Bob took 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 easily foil any such ex post facto *watermarking of her image. This is because Bob does not have the access to Alice's original image, even if he has access to the watermarked version of the image.*

Or can she? If Alice's original contains Bob's signature and vice versa, *who owns this image: Alice or Bob?*

In such a case rightful ownership cannot be resolved by invisible watermarks alone. We shall show in the following section that this scenario is not hypothetical, but can be engineered with current watermarking schemes. We will present in detail a counterfeit watermarking scheme that allows multiple claims of ownerships. More precisely, the true owner of an image can no longer argue his or her claim based only on the digital signatures that invisibly embedded in it, as others can engineer an equal amount of evidence that they too own the image. The situation will not happen with visible watermarks such as the one proposed in [6].

4 Invalidating Claims of Ownerships

To invalidate claims of ownerships of an image, it is necessary to generate the confusion illustrated in the case of Alice and Bob — that there are two original images, each contains the watermark of the other party. But in reality there is one and only one original. We shall 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. More formally, given $\hat{I}$ which is watermarked by a watermarking scheme ($\mathcal{E}$,$\mathcal{D}$,C_δ), we have in procession $\hat{I}'$, S' and a decoding function $\mathcal{D}'$ such that the following properties are satisfied:

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

2. $C_{\delta'}(\mathcal{D}'(I, \hat{I}'), S') = 1$.

δ' and δ are sufficiently large thresholds. $\mathcal{D}'$ can be the same as, or different from, the decoding function $\mathcal{D}$. The two parties can use different watermarking schemes in the latter.

> *Alice has an image I. She watermarks it with her watermarking scheme to generate a watermarked image $\hat{I}$ which is then made accessible to the public. Bob takes this watermarked image, and creates a counterfeit original $\hat{I}'$ using our scheme which he then claims to be his original. The first property states that Bob's fabricated "original" $\hat{I}'$ contains Alice's signature S. This is to be expected if the watermarking technique employed by Alice is robust. However, the second property implies that Alice's original image I contains Bob's signature S'!*
>
> *Bob can claim by virtue of property (2) that the image I (Alice's original) actually contains his watermark S'. Of course, Alice, by virtue of property (1), also claims that Bob's "original", $\hat{I}'$, contains her watermark S. Thus, it is not possible to determine the rightful owner of the image.*

Given only $\hat{I}$, we want to construct $C_{\delta'}, \mathcal{D}', \hat{I}'$ and S' such that both properties (1) and (2) are satisfied. This is achieved by *removing* a randomly selected watermark S' instead of embedding the watermark in. The process is illustrated as follows.

> *Bob identifies some features already present in the watermarked image, develops a scheme that removes these features which in return become his signature(s) S', and creates a fake original image $\hat{I}'$ which he then locks away as his "original" along with S', in the same way that Alice would lock away her original I and S. The scenario is shown in Figure 2.*

More precisely, in context of the feature based watermarking schemes described in section 2, the attacker (in this case, Bob) constructs his counterfeit "original" image as follows. He extracts a chosen (possibly random) watermark S' from the set $D'(\hat{I}) = \{f'_i(\hat{I})\}$ to generate an image $\hat{I}'$ such that

$$f'_i(\hat{I}') = f'_i(\hat{I}) \ominus s'_i. \tag{7}$$

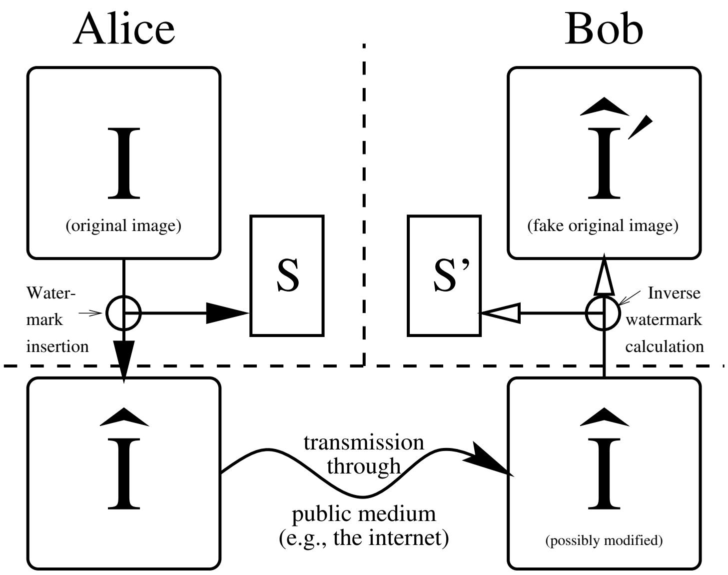

Figure 2: Forging a watermark. Alice watermarks image I to get $\hat{I}$, which she makes public. Bob computes an image $\hat{I}'$ and watermark S', such that watermarking $\hat{I}'$ with S' yields $\hat{I}$.

The set $D'(\hat{I})$ of derived coefficients is assumed to remain more or less the same when the image is not perceptually degraded. 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:

$$t'_i = f'_i(I) \ominus f'_i(\hat{I}'). \tag{8}$$

The confidence measure, taken to be the normalized correlation between T and S', defined in (6), 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

$$f'_i(I) \approx f'_i(\hat{I}). \tag{9}$$

Combining (7), (8) and (9), we have

$$t'_i \approx s'_i, \tag{10}$$

so that the correlation between T' and S' is large and implies that $C_{\delta'}(T', S')$ will most likely be equal to 1. The attacker (Bob) can thus claim that the true original I contains his signature S' and that I is a modified version of $\hat{I}'$. Conversely, the robustness of watermarking scheme[2] used to embed S onto I allows the true owner (Alice) to also argue that $\hat{I}'$ contains the watermark S. We now have a scenario whereby rightful ownership cannot be resolved through invisible watermarking scheme.

The counterfeiting scheme works by inverting the watermarking process. The key step is (7). 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. Notice that removing an existing watermark is not necessary to create this ownership deadlock, and thus the robustness of a watermark to survive image corruption is not enough to guarantee ownership.

We 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 $\ominus$.

Example 2 *A detailed analysis when $\mathcal{D} = \mathcal{D}'$, i.e., same decoding scheme and $D() = D'()$, i.e., when the same 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 $\hat{I}'$).*

[2]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.

Here, we have $f_i = f_i'$. The decoder $\mathcal{D}$, after comparing the original I and the fabricated "original" $\hat{I}'$, extracts $T = \{t_1, t_2, \cdots\}$:

$$\begin{aligned} t_i &= f_i(\hat{I}') - f_i(I) \\ &= f_i(\hat{I}) - s_i' - f_i(I) \\ &= s_i - s_i' \end{aligned}$$

Similarly, the decoder $\mathcal{D}'$, after comparing the fabricated "original" $\hat{I}'$ and the true original I, extracts $T' = \{t_1', t_2', \cdots\}$:

$$\begin{aligned} t_i' &= f_i'(I) - f_i'(\hat{I}') \\ &= f_i(I) - f_i(\hat{I}') \\ &= f_i(I) - f_i(\hat{I}) + s_i' \\ &= s_i' - s_i \end{aligned}$$

Thus, $t_i = -t_i'$ and the two correlations are identical:

$$\frac{\sum_i t_i.s_i}{\sqrt{(\sum_i t_i^2 \sum_i s_i^2)}} = \frac{\sum_i t_i'.s_i'}{\sqrt{(\sum_i t_i'^2 \sum_i s_i'^2)}} \tag{11}$$

$$= \frac{\sum_i s_i^2 - \sum_i s_i.s_i'}{\sqrt{(\sum_i (s_i - s_i')^2 \sum_i s_i^2)}} \tag{12}$$

The second term in (12) is very small if we assume little or no correlation between S and S' (which would be the case in practice) and the whole expression (12) would be large.

□

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.

Example 3 *A successful implementation of the proposed attack on the watermarking scheme proposed by Cox et. al. [1].*

In order to provide a more concrete example of counterfeiting an original image we wrote a program to implement the algorithm described in [1], and then modified it to perform the inverse operation as describe above. We used the same formula that Cox and *et al.* used to insert randomly generated watermark vector elements into an image's 1000 highest AC DCT coefficients v_i, 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[3].

This was then unleashed on an already watermarked image $\hat{I}$, using a watermark vector S' to yield a new "original" image $\hat{I}'$ (in reality a fake original) without any visible degradation of image quality. Using (5) 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 signature S is present in the fake original $\hat{I}'$ with a confidence value of 23.02.

Figure 4 shows the true original, the watermarked image, and a fake original. There are no visible artifacts observed from looking at these three images. They are virtually identical.

□

In the previous example, not only is the presence of our fabricated watermark well above random, it is virtually the same as the presence of the real watermark in the fabricated image! Based on the test results, who is the rightful owner of this image? Which image is the true original, and which is the fake original? Under such circumstances, what can invisible watermarks achieve? There is no additional evidence available to support any answers to these questions, and consequently, invalidating the claims of rightful ownerships.

The method of 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, any one can claim ownership of any watermarked image he or she has access to.

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

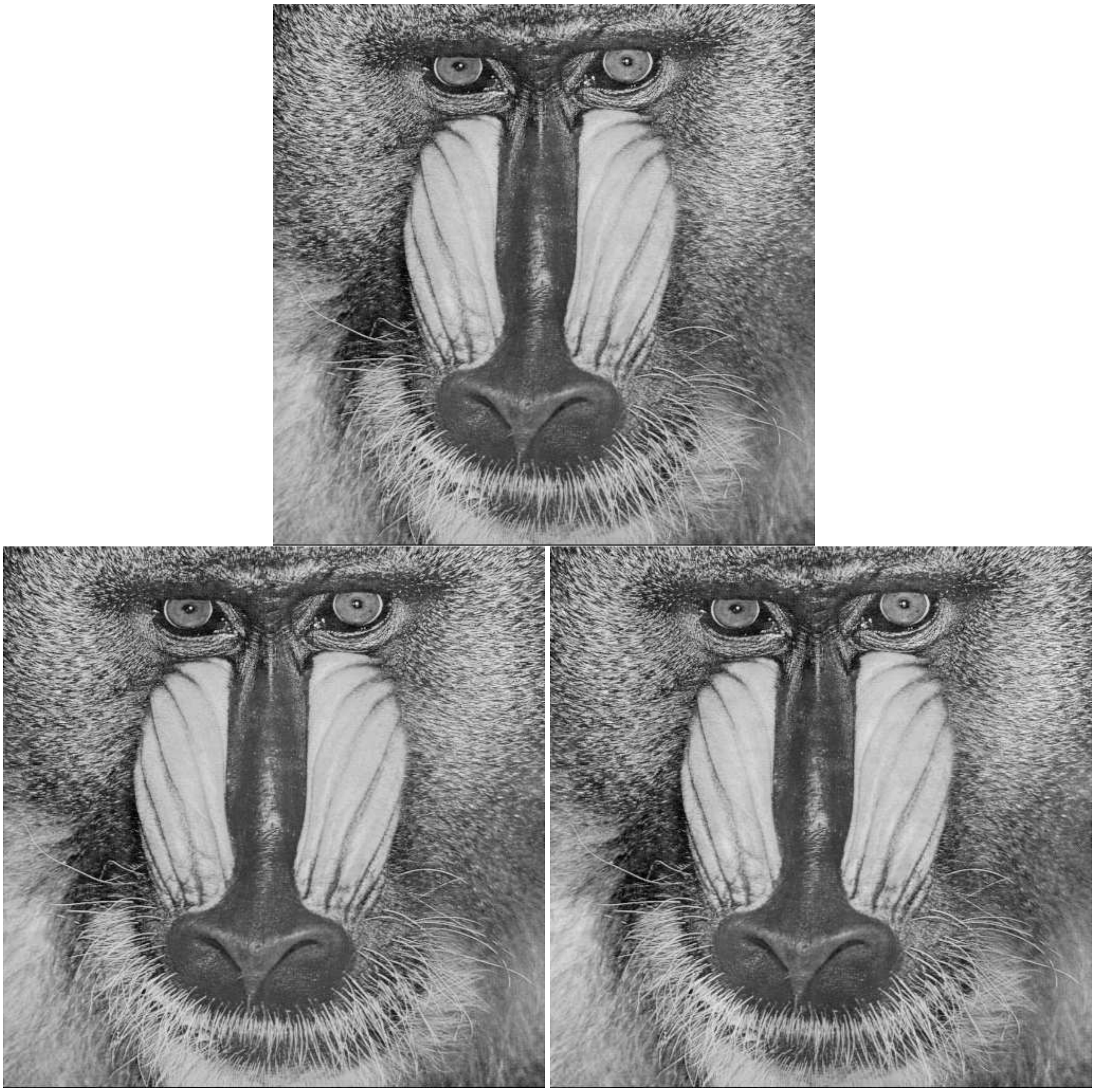

Figure 3: Three "Baboon" images (from USC database). (TOP CENTER) the watermarked image ($\hat{I}$) of the original with 1000-element watermark sequence inserted. (BOTTOM LEFT) The original image I. (BOTTOM RIGHT) The fabricated "original" image $\hat{I}'$. Confidence measure of the original watermark S in image $\hat{I}'$ is 23.02. Confidence measure of the fabricated watermark S' in image I is 23.52.

No matter which scheme Alice uses to invisibly 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.*

5 Non-invertible Watermarking of Images

In the previous section we have shown how one can fabricate an "original" image from a watermarked image such that rightful ownership cannot be resolved. We call the class of watermarking scheme that can be attacked by creating a "counterfeit original" as the *invertible* watermarking schemes. A more precise definition of the class of invertible watermarking schemes is as follows:

Definition: Invertible Watermarking Schemes

Let $\mathcal{V}$ be the class of invertible (invisible) watermarking schemes. Let S be a digital watermark, and S' be another digital watermark, such that both S and S' belongs to the set of allowable watermarks, then given an image I, $(\mathcal{E}, \mathcal{D}, \mathcal{C}_\delta) \in \mathcal{V}$ if for any S and S', there exists a $(\mathcal{E}', \mathcal{D}, \mathcal{C}_\delta)$ with

$$\mathcal{E}'(\hat{I}, S') = \hat{I}'$$

and

$$\mathcal{E}(\hat{I}', S') = \hat{I},$$

where $\hat{I} = \mathcal{E}(I, S)$, such that $\mathcal{E}'$ is a *computationally feasible* encoding function, and there is no perceptual quality degradation in the derivative images $\hat{I}$ and $\hat{I}'$ from I.
Otherwise, $(\mathcal{E}, \mathcal{D}, \mathcal{C}_\delta)$ is *non-invertible*. □

Note that this definition does not put any requirements on the decoding function $\mathcal{D}$.

While what we have described so far is a wide-sweeping attack, applicable to most current digital watermarking schemes, it seems that it may be foiled through more careful requirements for, or standardization of, the watermarking schemes. In other words, we can require that from legal point of view, to establish rightful ownership through invisible watermarks, the watermarking schemes applied to the images have to satisfy certain requirements — among them, the watermarking schemes cannot be invertible.

There are a number of ways to enforce this new requirement. One could develop a method which can be inverted, but in such a way that image quality is degraded to a high enough degree that the fabricated "original" image is clearly not the real original. In fact, it has been suggested that in the above process of signature fabrication, the image may already be degraded just enough to make it clear to an expert what is the real original and what is the fake one. However, progress in invisible watermarking schemes will most likely yield methods that create even less of a degradation on image quality, to the point that a fake original image as computed above (really, the result of watermarking an image *twice,* using two slightly different schemes), will be visually indistinguishable from the original. In fact, as Figure 4 shows, this is already true — there are no visible artifacts in the fake originals we fabricated. Finally, there are some images (pictures of clouds, say, or abstract art) for which comparisons of quality between original and modified versions will be difficult. In short, judgements based on perceptual "quality" are weak and unreliable for resolving image ownership.

Another approach would be to use one-way functions in the watermarking process, i.e., it would not possible to extract the watermark once it is inserted. The presence will then have to be inferred. Such a method, hopefully, does not suffer considerably in terms of flexibility or efficiency. With reference to the general formulation for watermarking described in Section 2, making the insertion operation non-invertible is an undesirable move, even though that insertion operation's invertibility is the source of our problems. This is because the confidence measure relies on the fact that the watermark vector can be later extracted, and using a non-invertible insertion formula may make this difficult. If the transform $D()$ that maps the image into feature space is linear, then it may be possible to extract a signature without using the extraction operator $\ominus$. However, in general this may not be always true.

A third approach follows from noting that in order to fabricate a counterfeit original $\hat{I}'$, the attacker (Bob) first chooses a watermark S' that he proceeds to "remove" from the watermarked image $\hat{I}$ that belongs to the owner (Alice). Now, if we enforce the requirement that any watermark S that is inserted into an image I be dependent on I, then we make it difficult for Bob to select S' as it depends on $\hat{I}'$ which has not yet been determined. This can be achieved by computing a bit sequence $B = \{b_1, b_2, \ldots\}$ from the image I that is being watermarked and using these bits in the watermarking process. One way of using the bits 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 $\otimes$. Furthermore, obtaining the bit sequence from a one-way function of the image to be watermarked, would make the scheme secure against any trial and error type of attacks that Bob can attempt in order to construct a counterfeit original. We now provide one such example and discuss possible attacks using the *same* scheme.

Example 4 *A modified version of the scheme described by Cox et. al. [1].*

We first produce a 1000-bit $\{b_1, b_2, \cdots b_{1000}\}$ one-way hash of the original image before computing the 2D DCT of the image. 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, will not show up in the watermarked image.

Specifically, we used two versions of the second update formula in [1] as follows:

$$v_i' = \begin{cases} v_i(1 + \alpha s_i), & b_i = 0; \\ v_i(1 - \alpha s_i), & b_i = 1, \end{cases} \tag{13}$$

where in both cases α was chosen to be 0.1. A 1000-bit hash of the image was computed, and for each of the 1000 highest AC DCT matrix elements, one of the formula was used depending on the value of the hash bit b_i. For convenience, these hash bits were stored in the watermark vector file.

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 can impose the requirement that $s_i > 0$ for otherwise, an attacker can simply negate certain watermark vector elements to match the resulting hash bits, i.e., an attacker chooses an arbitrary binary sequence $a_1, a_2, \cdots, a_{1000}$ and applies the following transformation on the watermarked image $\hat{I}$ to create an "original" $\hat{I}'$:

$$v_i^{''} = \begin{cases} v_i'(1 - \alpha\hat{s}_i), & a_i = 0; \\ v_i'(1 + \alpha\hat{s}_i), & a_i = 1. \end{cases} \tag{14}$$

He then computes the hash of his "original" $\hat{I}'$: $\{\hat{b}_1', \hat{b}_2', \cdots\}$. His watermark is then given as follows:

$$\hat{s}_i' = \begin{cases} \hat{s}_i, & a_i = \hat{b}_i'; \\ -\hat{s}_i, & \text{otherwise.} \end{cases} \tag{15}$$

This could also be avoided by choosing two update formulas that differ in a different way than the two we used. It is also important that the hash string be image-dependent for this scheme to be robust against attacks.

We apply this watermarking scheme on 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 watermarked images are tested for the presence of the signature S. The results are displayed in Figure 4. 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 presence and is thus useless. □

The above example illustrates the difficulty in "inverting" the watermarking scheme. Because the fake "original", $\hat{I}'$, has not been recreated, one cannot know the associated hash string $\{\hat{b}_1', \hat{b}_2', \cdots\}$. This in turn implies that it is not possible to decide which formula to use to "remove" a watermark element from $\hat{I}$. We believe the proposed scheme is non–invertible although it seems difficult to rigorously prove this.

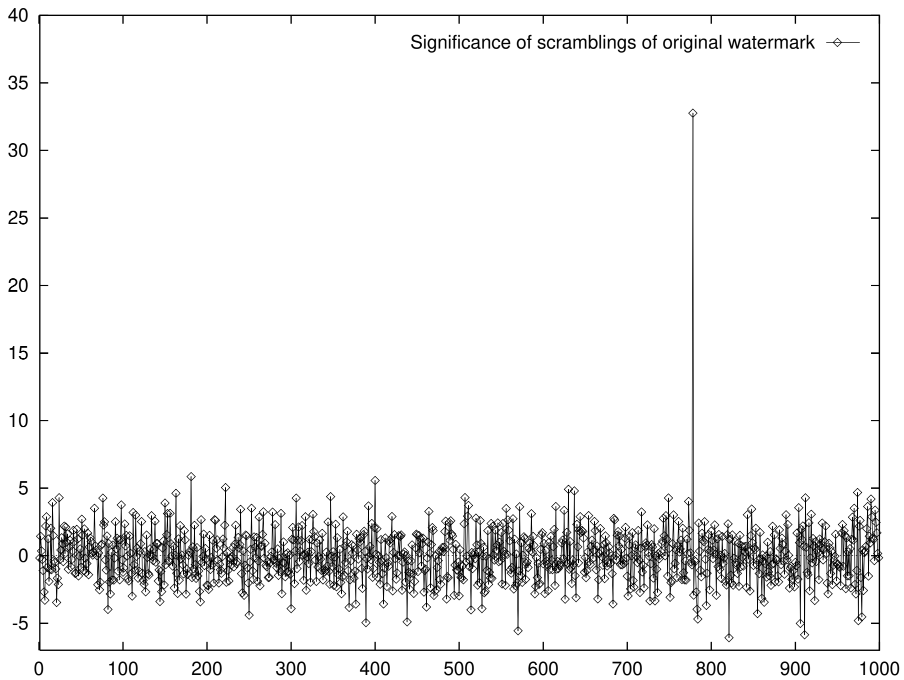

Figure 4: Results of scrambling hash bits in watermark vectors. The original (as seen by the spike) and 999 copies with scrambled hashes

In conclusion, requiring non-invertibility of a watermarking scheme may be necessary to prevent fabrication attacks such as those we just illustrated. Without such requirements, one could use a non-invertible scheme such as the one described in Section 4 to generate the necessary confusion. Thus, from a legal standpoint, to be able to establish rightful ownership through invisible watermarks, it would be necessary that watermarking schemes satisfy certain requirements. We have demonstrated in this section that non-invertibility could be an answer.

6 Conclusions and Discussions

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[4]. 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 which can be so easily copied and manipulated. This 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 could be variations on the requirement of the 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 very 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

[4] We quote several sentences in [7] 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.

requirements are imposed legally on the watermarking schemes. In the absence of such requirements, Alice cannot simply lock away an original image which 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.

But then, what would be the need of watermarking? Watermarking would still be very much needed for protecting her interests. For example, Alice could embed a different watermark in each copy of the image that she sells. This unique watermark would enable her to determine the identity of her specific customer who may be making unauthorized copies and selling them for a profit. A watermark could 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, which 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 would require different types of watermarking schemes with different requirements. For example, non-invertibility may not be an issue when there is a centralized authority with whom copyrighted images are registered. However, in such an application it would be useful for an user to query an image for copyright protection. In certain applications such as the use of invisible watermark to ensure the integrity of digital fingerprints, it is important only to verify that a fingerprint image has not been tampered with; in such a case, the robustness of the watermark against image modification is *not* an issue. On the other hand, it is also important to investigate into crytographic protocols that make it difficult for general attacks (such as that proposed in [5]) to *remove* or *diminish the presence of* the watermark in the image.

References

[1] I.J. Cox, J. Kilian, T. Leighton and T. Shamoon, “Secure Spread Spectrum Watermarking for Multimedia”, *NEC Research Institute Technical Report*, 95-10.

[2] J. Zhao and E. Koch, “Embedding Robust Labels into images for Copyright Protection”, *Intellectual Property Rights and New Technologies, Proceedings of the KnowRight’95 Conference 1995*, pp.242–51 .

[3] N. Nikolaidis and I. Pitas, “Copyright Protection of Images using Robust Digital Signatures”, *Proceedings, IEEE International Conference on Acoustics, Speech and Signal Processing 1996*.

[4] F.M. Boland, J.J.K. O’Ruandaidh and C. Dautzenberg, “Watermarking Digital Images for Copyright Protection”, *Proceedings, IEE Image Processing And its Applications Conference 1995*, pp.326–330.

[5] H.S. Stone, “Analysis of Attacks on Image Watermarks with Randomized Coefficients”, *NEC Research Institute Technical Report*, 1996.

[6] G. Braudaway, K. Magerlein and F. Mintzer, “Protecting publicly available images with a visible image watermark”, *Proc. SPIE: Optical Security and Counterfeit Deterrence Techniques*, Vol. 2659, pp. 126–133, 1996.

[7] U.S. Copyright Office, Circular 1, “Copyright Basics”, March 28, 1996. The circular is available on Internet via World Wide Web URL *http://lcweb.loc.gov/copyright* and via Gopher *marvel.loc.gov*.