On The Invertibility Of Invisible Watermarking Techniques

Conference Paper · November 1997

DOI: 10.1109/ICIP.1997.647969 · Source: IEEE Xplore

4 authors, including:

Nasir D. Memon
New York University
404 PUBLICATIONS **13,563** CITATIONS

SEE PROFILE

Boon-Lock Yeo
Princeton University
73 PUBLICATIONS **5,388** CITATIONS

SEE PROFILE

Minerva M. Yeung
Princeton University
91 PUBLICATIONS **4,534** CITATIONS

SEE PROFILE

ON THE INVERTIBILITY OF INVISIBLE WATERMARKING TECHNIQUES

Scott Craver, Nasir Memon
Northern Illinois University,
DeKalb, IL 60115.

Boon-Lock Yeo and Minerva M. Yeung
IBM T.J. Watson Research Center,
Yorktown Heights, NY 10598.

Abstract

In this paper we address the invertibility of invisible watermarking schemes for resolving rightful ownerships, and present attacks which can cause confusion to rightful claims. We shall show that non-invertibility is a necessary but not sufficient condition in resolving ownership disputes. We then define quasi-invertible *watermarking schemes, and, present analysis that links invertibility and quasi-invertibility to some classes of watermarking techniques with different properties (which may or may not require original versions in watermark decoding), as well as to the different classes of attacks we have developed.*

1. INTRODUCTION

The rapid growth of digital imagery coupled with the ease of duplication and distribution of digital information, have generated interest in the development of effective digital copyright protection mechanisms. Various invisible watermarking techniques have been proposed in recent literature. It may have appeared from some ensuing work that the most important property of the watermarking schemes for providing unique proof of ownership was their robustness — their ability to survive despite malicious attempts at removal, or common image processing operations. Watermarking schemes have been proposed and shown to be robust. However, in recent work [1] we demonstrated that the ability to embed robust watermarks does not necessarily imply the ability to establish ownership, unless certain requirements are imposed on the schemes used in inserting the watermarks. In the absence of such requirements, we show how confusion can be created by developing counterfeit watermarking schemes that allow multiple claims of rightful ownerships. This led us to formulate *invertibility* of invisible watermarking schemes and we presented *non-invertible* watermarking schemes as a remedy.

In this paper we will address the invertibility of invisible watermarking schemes in depth and present more powerful attacks on some existing watermarking schemes. In section 2 we review non-invertibility, and show it is a necessary but not sufficient condition for ownership resolution. We then formulate *quasi-invertibility*, and present a non-invertible watermarking technique that to the best of our knowledge is also non-quasi-invertible in Section 3. In section 4 we demonstrate an attack on *public* watermarking schemes (which do not require the use of an original version in decoding the watermark). A detailed study of invertibility and quasi-invertibility in several classes of watermarking techniques and different types of attacks that invalidate ownership claims is reported in [2].

2. NON-INVERTIBLE WATERMARKING

We denote an image by I, a watermark comprising of a sequence of *ownership labels* $S = \{s_1, s_2, \cdots\}$ and the watermarked image by $\hat{I}$. $\mathcal{E}$ is an encoder function if it takes an image I and a watermark S, and generates a new image which is called the *watermarked* image $\hat{I}$, i.e., $\mathcal{E}(I, S) = \hat{I}$.

A decoder function $\mathcal{D}$ takes an image J whose ownership is to be determined, and recovers a watermark T from the image, with reference to the original image I: $\mathcal{D}(J, I) = T$. The extracted watermark T is then compared with the owners watermark by a comparator function $\mathcal{C}_\delta$, and a binary output decision generated indicating a match or otherwise:

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

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

Given only $\hat{I}$ which is watermarked by some scheme ($\mathcal{E}$,$\mathcal{D}$,$\mathcal{C}_\delta$), we demonstrated in [1], how to reverse-engineer a counterfeit original image $\hat{I}'$, a watermark S' and a decoding function $\mathcal{D}'$ such that the following properties are satisfied:

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

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

where δ' and δ are sufficiently large thresholds. $\mathcal{D}'$ can be the same as, or different from, the decoding function $\mathcal{D}$. By virtue of the above properties, the attacker has acquired equal amount of ownership evidence through the use of invisible watermarking as the true owner. This is what we call a SWICO (**S**ingle-**W**atermarked-**I**mage-**C**ounterfeit-**O**riginal) attack, which makes possible multiple claims to ownerships of an image. Notice that it is not necessary to remove an existing watermark from an image in order to create this ownership deadlock, and thus the robustness of

Proceedings of the 1997 International Conference on Image Processing (ICIP '97)
0-8186-8183-7/97 $10.00 © 1997 IEEE

a watermark to survive image corruption is not enough to guarantee ownership. The counterfeiting is essentially accomplished by inverting a watermarking schemes encoding function $\mathcal{E}$, in order to "remove" a watermark from an image rather than insert one. We refer such as an *invertible* watermarking scheme.

Definition 1 [Invertible Watermarking Schemes]

A watermarking scheme $(\mathcal{E}, \mathcal{D}, \mathcal{C}_\delta)$ *is* **invertible** *if, for any image* $\hat{I}$*, there exists a mapping* $\mathcal{E}^{-1}$ *such that* $\mathcal{E}^{-1}(\hat{I}) = (\hat{I}', S')$, *and* $\mathcal{E}(\hat{I}', S') = \hat{I}$*, 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* **non-invertible**.

$\mathcal{E}^{-1}$ is the inverse mapping, and $\hat{I}'$ the inverse image of $\hat{I}$. 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. The definition implies that $\mathcal{C}_\delta(\mathcal{D}(\hat{I}, \hat{I}'), S') = 1$. The scheme in [3] is an invertible scheme.

Removing this invertibility is a crucial step in foiling our attack. We note that to fabricate a counterfeit original $\hat{I}'$, the attacker may have to choose 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. The attacker would have to compute S' dependent upon the final value of $\hat{I}'$, but that final value cannot be known until S' is used in its computation. This could be achieved by computing a bit sequence $B = \{b_1, b_2, \ldots\}$ from I (for example, via a one-way hash) which is then used in watermarking I.

Example 1 *A modified non-invertible version of the scheme described by Cox et. al. [3].*

We first produce a 1000-bit one-way hash $\{b_1, b_2, \cdots b_{1000}\}$ of the original image before computing it's 2D 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. Specifically, we used two versions of the second update formula in [3] as follows:

$$v_i' = \begin{cases} v_i(1 + \alpha s_i), & b_i = 0; \\ v_i(1 - \alpha s_i), & b_i = 1, \end{cases} \qquad (3)$$

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 coefficients, one of the formula was used depending on the value of the hash bit b_i.

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.

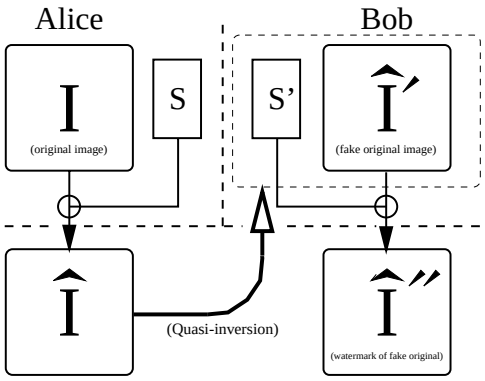

Figure 1: The TWICO-attack.

The above scheme is non-invertible and appears to circumvent the SWICO attack. It is, however, not effective against another attack which we shall introduce as a more general instance of the counterfeit attack. In other words, we shall demonstrate that non-invertibility is a necessary but not sufficient condition for resolving ownership.

3. QUASI-INVERTIBILITY

We have assumed previously 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. Our new and more general attack uses two watermarked images, one created by Alice, $\hat{I}$, from her original I with her watermark S embedded which is the true watermarked version, and the other created by Bob, $\hat{I}''$, from his counterfeit original $\hat{I}'$ with his watermark S' embedded, which is a fake watermarked image. We show that Bob can successfully counter Alice's claim of ownership by engineering an equal amount of evidence in terms of watermark presence as Alice, through the construction of a counterfeit original $\hat{I}'$ and the new watermarked image $\hat{I}''$, which are perceptually similar to $\hat{I}$, with the presence of his watermark S' in both Alice's original I and watermarked image $\hat{I}$, as well as in the new watermarked image $\hat{I}''$. In other words, given only $\hat{I}$, we construct $C_{\delta'}$,$\mathcal{D}'$, $\hat{I}'$, $\hat{I}''$, and S' such that the properties in (1) and (2) hold; but unlike the SWICO attack, we do not require that the insertion of the fake watermark S' onto the counterfeit original $\hat{I}'$ produce the same watermarked image $\hat{I}$. Instead, we produce another watermarked image, $\hat{I}''$. We call such a counterfeit-attack involving two watermarked versions of an image, a TWICO (**T**win-**W**atermarked-**Images** **C**ounterfeit-**O**riginal) attack. This is illustrated in Figure 1.

Notice that, a watermarking scheme need not be invertible to be susceptible to a counterfeit attack. Rather than having to compute an image $\hat{I}'$ and watermark vector S' such that marking $\hat{I}'$ with S' yields Alice's watermarked $\hat{I}$,

Proceedings of the 1997 International Conference on Image Processing (ICIP '97)
0-8186-8183-7/97 $10.00 © 1997 IEEE

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 perceptual quality, that S' can still be expected to lie in I. This leads us to define *quasi-invertible* watermarking schemes.

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* $\mathcal{E}^{-1}(\hat{I}) = (\hat{I}', S')$*, where* $\mathcal{E}^{-1}$ *is a* computationally feasible *mapping,* S' *belongs to the set of allowable watermarks, the images* $\hat{I}$ *and* $\hat{I}'$ *are perceptually similar, and* $C_\delta(\mathcal{D}(\hat{I}, \hat{I}'), S') = 1$*. Otherwise,* $(\mathcal{E}, \mathcal{D}, \mathcal{C}_\delta)$ *is* **non-quasi-invertible**.

We now show that the watermarking scheme in Example 1 is quasi-invertible and subjected to the TWICO attack: Bob extracts a 1000-element watermark S using the 1000-bit string "1111 ... 1" (that is, using the second insertion formula for every single watermark vector element), and a second watermark T using the string "0000 ... 0". The two resulting images are averaged, to yield an image $\hat{I}'$ which he claims to be his original. This image is then hashed, and the watermark S' is computed thus: if the ith hash bit of the average is 0, 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 S' to produce a watermarked image $\hat{I}''$, which he claims to be the watermarked version of his "original." S' is then a vector composed of elements from S and T so as to match the 1000-bit hash of I.

We have found that Bob's mix-and-match method of constructing any watermark works astonishingly well. For the experiment, a watermarked image was used to fabricate a fake original using this averaging technique, increasing the value of α from 0.1 to 0.27 so that the watermark exists more prominently. 1000 watermarks were created based on random bit strings, and each was tested for presence in Alice's original image. The mean correlation value was 9.97, well above random, and the highest computed correlation was 13.8. Alice's original watermark is present in the fake original with a significance of 14.2. The data do not present a good margin for distinguishing the true owner.

This is an example of a non-invertible watermarking scheme that, because of its quasi-invertibility, still falls prey to counterfeit attacks. Hence we have to require that a good scheme for resolving ownerships must be practically non-quasi-invertible. While a frightening reminder that an algorithm cannot be made secure merely by implementing a one-way function, this pitfall can be avoided by using another non-invertible watermarking scheme presented below. As in Example 1, the technique makes use of a bit string generated by a one-way hash of the image. However, the string is used in a different way that eliminates the potential for mix-and-match attack described above.

Example 2 *A more secure (and further modified) version of the scheme described by Cox et. al. [3].*

We first produce an n-bit $\{b_1, b_2, \cdots b_n\}$ one-way hash of the original image, with the possibility of incorporating a relatively small standard identifier in order to allow multiple distinct watermarks to be inserted into the same image. We then use this hash as the seed for the pseudo-random-number generator used to generate the normally distributed watermark vector elements. A vector, then, is not considered an allowable watermark unless it is 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 is very unlikely to produce an allowable watermark. Determining the seed that generates an arbitrary allowable watermark is hard. The true owner 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 non-quasi-invertible, and secure against the different types of attacks presented in this paper.

4. INVERTIBILITY OF PUBLIC SCHEMES

We have used as examples of counterfeit-attack on the "private" watermarking schemes like [3], in which an original image is necessary for watermark extraction and decoding. We can also engineer similar attacks on "public" watermarking schemes which do not involve an original image in detecting the watermark. We show here an implementation using the technique proposed in [4]. 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). We shall call the division strategy S which is also the watermark. To yield the watermarked image, k is added to each pixel in A. Both S and k are needed for proper decoding. To detect the watermark from an image $\hat{I}$, based on watermark S, we compute $\bar{w} = \bar{A} - \bar{B}$, where $\bar{A}$ and $\bar{B}$ are the mean pixel values in A and B respectively. If $\hat{I}$ is watermarked using S and k, then $\bar{w} \approx k$, Otherwise, $\bar{w} \approx 0$ if S is randomly chosen. The test statistics $q = \frac{\bar{w}}{s_{\bar{w}}}$, where $s_{\bar{w}}$ is the sample standard deviation of $\bar{w}$, will then indicate the confidence of the watermark absence or presence. It has a distribution with mean $\frac{k}{s_{\bar{w}}}$ or 0, depending on whether there is a watermark presence or not in $\hat{I}$. We shall write

Proceedings of the 1997 International Conference on Image Processing (ICIP '97)
0-8186-8183-7/97 $10.00 © 1997 IEEE

$q = \mathcal{D}(\hat{I}, S, k)$ 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. A crucial requirement for the watermarking scheme to be susceptible to attack is the presence of attacker's (Bob's) watermark on the watermarked image $\hat{I}$, as well as on the true original I. Bob has to reconstruct his watermark directly from, and without changing, the pixel values of Alice's watermarked image. 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' which forms two sets, A^* and B^*, such that the difference $\bar{w}$, 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 $\bar{w}'$ too unrealistic in typical images (for example, $\bar{w}' \approx 87$ for the image Lena — k embedded commonly have values range from 2 to 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 $\bar{w}$ — this is possibly because increasing l decreases $\bar{w}$. 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 $\hat{I}$.

Image	k	k'	q_w	q'_w	q_0	q'_0	Quality
Lena	3	3	14.5	15.3	14.5	15.3	44.6
Pepper	3	4	10.2	14.3	9.9	14.3	40.3

Table 1: A summary of confidence measurements q.

Table 1 summarizes the confidence measurements q in two test images by constructing a partition using the above method to form Bob's fake watermark. q_w, q'_w, q_0 and q'_0 denote $\mathcal{D}(\hat{I}, S, k)$, $\mathcal{D}(\hat{I}, S', k')$, $\mathcal{D}(\hat{I}', S, k)$ and $\mathcal{D}(I, S', k')$ respectively. The quality measures the PSNR of the counterfeit originals in dB. The attack uses under 1 second computing time on a Pentium 100MHZ PC — there are many possible set partitions in most images that can give the desired test statistic values for invalidating rightful claims. In general, the presence of Bob's watermark on Alice's original and watermarked image are the same or better than, Alice's measurements of her watermark presence in Bob's counterfeit original and her own watermarked image. In other words, the watermarking schemes like those presented in the above are invertible. There are remedies that can be deployed to make the attack more difficult by imposing stringent requirements on how the sets can be partitioned. One method is to use similar solution suggested in Example 2 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.

5. CONCLUSION

In conclusion, we have shown that although non-invertibility of a watermarking scheme is necessary to prevent counterfeit fabrication attack, it is not sufficient unless "invertibility" is taken in a very general sense. Designers of non-invertible schemes must therefore take extra care to ensure that theirs is non-quasi-invertible as well; a one-way function may not guarantee a secure scheme. We have shown in this paper how we can circumvent the step of inverting a one-way function, thereby defeating a non-invertible scheme which has a one-way function embedded in the watermarking process. In addition, we have shown that attacks to generate ownership confusion are not confined to the class of private watermarking schemes. We have illustrated that even the public watermarking schemes, if not designed carefully and used in the proper context, may be susceptible to counterfeiting attacks.

6. REFERENCES

[1] S. Craver, N. Memon, B. L. Yeo, and M. M. Yeung, "Can invisible watermarks resolve rightful ownerships?", IBM Technical Report RC 20509, July 25, 1996. IBM CyberJournal: http://www.research.ibm.com:8080.

[2] S. Craver, N. Memon, B. L. Yeo, and M. M. Yeung, "Resolving rightful ownerships with invisible watermarking techniques: Limitations, attacks, and implications", accepted for publication, *IEEE Journal on Selected Areas of Communications*. (IBM RC20755, Mar. 1997).

[3] I.J. Cox, J. Kilian, T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for multimedia", Technical Report 95-10, NEC Research Institute, 1995.

[4] I. Pitas, "A method for signature casting on digital images", *in Proceedings the International Conference on Image Processing 1996*, vol. 3, pp. 215–218.

Proceedings of the 1997 International Conference on Image Processing (ICIP '97)
0-8186-8183-7/97 $10.00 © 1997 IEEE