Protocols for Watermark Verification

K. Gopalakrishnan*
Department of Mathematics
East Carolina University
Greenville, NC 27858
gopal@cs.ecu.edu

Nasir Memon
Computer Science Department
Polytechnic University
Brooklyn, NY 11201
memon@poly.edu

Poorvi Vora
Imaging Technology Department
Hewlett-Packard Labs
Palo-Alto, CA 94304 pvora@hpl.hp.com

Abstract

Digital watermarks have recently been proposed for the purpose of copy deterrence of multimedia content. Copy deterrence using digital watermarks is achieved by inserting a unique watermark into each copy of the image sold which could be used to trace unauthorized copies to the erring buyer. However, all the schemes proposed in the literature suffer from the following major limitation which is generally glossed over. The owner of the image reveals the embedded watermark to the buyer or to a "trusted" third party in the process of proving that a pirated copy originated from a specific buyer. Needless to say that the watermark could subsequently be removed from the watermarked image by the buyer or the "trusted" third party. This leads to the same fundamental problem viz., the buyer or the "trusted" third party could now sell illegal copies with complete impunity. While this limitation appears to be inherent, it can actually be eliminated by using appropriate cryptographic protocols. In this paper, we present a novel way of demonstrating the presence of a watermark in an image without revealing knowledge about the watermark to the other party, thereby eliminating the possibility of subsequent watermark removal.
keywords: Digital Watermarking, Copyright Protection and Fingerprinting.

1 Introduction

Consider an application where multimedia content is electronically distributed over a network. In order to discourage unauthorized duplication and distribution, the content owner can embed a unique watermark (or a fingerprint) in each copy of the data that is distributed. If, at a later point of time, unauthorized copies of the data are found, then origin of the copy and hence the identity of the erring buyer could be determined by retrieving the unique watermark corresponding to each buyer. This type of schemes are sometimes called *copy deterrence watermarking schemes* or *digital fingerprinting schemes*. Here, we will focus our attention on digital image watermarking, although the same problems exist for other multimedia data such as video or audio.

A *watermark* is a signal added to the digital image which can later be extracted or detected to make an assertion about the image [4]. There are two types of watermarks *visible* and *invisible*. A *visible* watermark typically contains a conspicuously visible message or a company logo indicating the ownership of the image. On the other hand, *invisible* watermarks are unobtrusive modification to the image and the invisibly watermarked content appears visually very similar to the original. The existence of an invisible watermark can only be determined by using an appropriate watermark extraction or detection algorithm. Invisible watermarks are generally preferred as their unobtrusiveness makes them more desirable.

Watermarking techniques can also be classified as *fragile* and *robust*. *Fragile* watermarks are easily corrupted by any form of image processing procedure. *Robust* watermarks can resist common image manipulation procedures such as rotation, reflection, scaling, cropping, smoothing, contrast or brightness adjustment or lossy compression. Clearly, a watermark used for the purpose of copy deterrence needs to be a robust one.

Yet another classification of watermarking techniques is into *oblivious* and *non-oblivious* schemes. A *non-oblivious* scheme requires an original or reference image in the watermark detection procedure. On the other hand, an *oblivious* scheme does not require the use of an original or reference. Obviously, oblivious schemes are attractive in many applications.

In copy deterrence watermarking schemes, the watermarks used are generally invisible, robust and oblivious. Recall that copy deterrence using digital watermarks is achieved by inserting a unique watermark into each copy of the image sold which could be used to trace unauthorized copies to the erring buyer. In such a scenario, in order to indict the erring buyer, the seller has to demonstrate the presence of the unique watermark of the erring buyer on an unauthorized copy of the image and also evidence binding the specific watermark to the buyer. The latter can be done by obtaining a digital certificate at the time of sale which, say, is in the form of the hash of the watermark, details of the terms of the sale, the identity of the buyer, all time stamped and signed by a trusted authority. To establish the former, it has been generally assumed in the literature that the seller reveals the embedded watermark to the buyer or to a "trusted" third party. Needless to say that once the watermark is revealed it could subsequently be removed from the watermarked image by the buyer (or by the "trusted" party) and now in fact the buyer (or the "trusted" party) can resell multiple copies of the image with complete impunity. While this limitation appears to be inherent, it can actually be eliminated by using appropriate tools from cryptography. In this paper, we present a novel way of demonstrating the presence of a watermark in an image without revealing knowledge about the watermark to the other party. This prevents the adversary from removing the watermark subsequently.

The rest of the paper is organized as follows. In Section 2, we present the protocol that demonstrates the presence of an watermark without revealing it to the verifier. In section 3, we discuss zero knowledge proofs and show that an earlier zero knowledge protocol based on the graph isomorphism problem for digital watermarking is seriously flawed. Finally, in Section 4, we provide some concluding remarks and suggest some open problems.

Figure 1: The Watermarking Decision Problem

Problem Instance: The digital image Y in dispute, seller's public-key and an encrypted spread-spectrum watermark $E(W)$. **Question:** Is the watermark W present in the digital image Y?

2 Proposed Protocol for Watermark Verification

The watermark verification protocol we propose works with linear and additive watermarking techniques where watermark detection is done by means of correlation. However, for ease of exposition, we present it in terms of the spread-spectrum watermarking technique proposed by Cox et al [1] that was demonstrated to be remarkably robust against malicious attacks aimed at its removal. Before we present our protocol, we first briefly review this technique.

Cox et al [1] embed a set of independent real numbers $W = \{w_1, w_2, \ldots w_n\}$ drawn from a zero mean, variance 1, Gaussian distribution into the n largest DCT AC coefficients of an image. Results reported using the largest 1000 AC coefficients show the technique to be remarkably robust against various image processing operations, and also after printing and re-scanning.

Specifically, they take the 2-dimensional DCT of an image X and the watermark W is inserted into the largest n AC coefficients $\{x_1, x_2, \ldots, x_n\}$ by a suitable insertion formula to yield modified coefficients $\{x'_1, x'_2, \ldots, x'_n\}$. For example, the insertion formula used could be

$$x'_i = x_i(1 + \alpha w_i)$$

where α is a small constant. An inverse 2D DCT is then taken, yielding the watermarked image X'. To determine if a given image Y contains the watermark W, the decoder first takes the 2-dimensional DCT of the image and extracts the largest N DCT coefficients $Y = \{y_1, y_2, \cdots y_n\}$. The confidence measure on the presence of the watermark W in Y is taken to be the correlation between W and Y. Note that this version of their technique is invisible, robust and oblivious.

Under our scenario of copy deterrence watermarking schemes using the spread-spectrum technique, the seller or the distributor inserts a unique watermark, that is distinct for each buyer, into the image before distributing it to the buyer. The seller also encrypts this watermark W using his public key of the well-known RSA public-key cryptosystem and obtains a time stamped digital certificate binding $E(W)$ to the specific buyer. At a later point of time, the seller encounters an image Y and contends that it is a pirated copy originating from a specific buyer. In order to establish it, the seller has to prove that the answer to the *watermarking decision problem* presented in Figure 1 is a resounding "yes".

Note that the seller can of course do so, by simply disclosing the watermark W and the digital certificate that binds $E(W)$ to the buyer. The verifier can check the certificate and also check that $E(W)$ is indeed the encryption of W and then finally check that W is present in Y by using the watermark detection procedure of a spread-spectrum technique in the standard manner. But then, the verifier now knows the watermark W and hence can

Figure 2: A Protocol for the Watermarking Decision Problem

input: The digital image Y in dispute, seller's public key and an encrypted spread-spectrum watermark $E(W)$.

1. Repeat the following steps k times:
2. The seller chooses a random number r and uses it to generate a sequence ϵ in a "one-way manner". The seller then adds ϵ to Y to get an "image" $Y' = Y + \epsilon$. The seller encrypts Y' and sends $E(Y')$ to the verifier.
3. The verifier chooses a random integer $j = 1$ or 2 and sends it to the seller.
4. If $j = 1$, seller reveals Y' and r and as a consequence also reveals ϵ. The verifier checks $E(Y')$ is consistent i.e. $E(Y') = E(Y + \epsilon)$ and ϵ is indeed random.

   If $j = 2$, seller demonstrates that Y' and W do correlate.
5. The verifier accepts the seller's proof if the computation of step 2 and step 4 are verified in each of the k rounds.

remove it from the image Y and then resell multiple copies of it with complete impunity. The basic problem is that the seller has lost his power of demonstrating that a disputed copy is a pirated copy the moment he has disclosed the unique watermark. However, there is no reason why the seller should prove that the answer to the watermarking decision problem is "yes" in the above manner. It is possible to prove that the answer is "yes" without revealing the watermark by using tools from cryptography.

Specifically, the seller can use the protocol presented in Figure 2 to prove that the answer to the watermarking decision problems is "yes" without revealing the watermark itself. Although most of the protocol is self explanatory, some additional clarification on step 4 of the protocol is in order. When $j = 1$ the seller reveals I' and r. The verifier checks that ϵ was indeed generated by r in a manner that is "one-way" or difficult to invert. The verifier also checks that $E(Y') = E(Y + \epsilon)$. This step ensures that the sequence ϵ added to Y is indeed random does not correlate with W by design.

We now take a slight digression to explain the step $j = 2$. Let $a = (a_1, a_2, \ldots a_n)$ be a sequence and let $b = (b_1, b_2, \ldots b_n)$ be another sequence. Whether these two sequences correlate or not is essentially determined by the value of the inner product $a_1b_1 + a_2b_2 + \ldots + a_nb_n$. If $E(a)$ and $E(b)$ are available to the verifier, then the seller could disclose the sequence $(a_1b_1, a_2b_2, \ldots, a_nb_n)$ to the verifier. The verifier can simply add the elements of this sequence and thus determine whether or not the sequences a and b correlate. The verifier can be confident that the sequence given by the seller is not arbitrary by checking that

$$E(a_ib_i) = E(a_i)E(b_i)$$

for $i = 1, 2 \ldots n$. This checking is possible as the verifier is in possession of both $E(a)$ and $E(b)$, has been given the plaintext values of $a_i b_i$ by the seller and the RSA cryptosystem has multiplicative homomorphic property.

Now, if $j = 2$, the seller discloses the sequence $(y_1^{'} w_1, y_2^{'} w_2, \ldots y_n^{'} w_n)$ to the verifier. The verifier can then check the legitimacy of the sequence given to him, using the seller's public key, as he is already in possession of both $E(Y^{'})$ and $E(W)$. He can then add up the elements of this sequence and use the result to check that $Y^{'}$ and W indeed do correlate.

As the verifier knows that $Y^{'}$ is derived from Y by insertion of ϵ, $Y^{'}$ correlates with W and the random sequence ϵ does not correlate with W, he can conclude that Y must correlate with W and therefore the pirated copy must have originated from the specific buyer.

Note that in each round of the protocol, the seller only proves one of two statements viz., $Y^{'}$ correlates with W or Y' was constructed as agreed upon. However, as he does not know which one of these two statements he will be asked to prove before he commits to $E(Y^{'})$, he cannot choose ϵ by malicious design.

Several technicalities need to be addressed now. These include a formal proof that the protocol absolutely does not reveal any information about the watermark other than what could be gained from the input without participating in the protocol(that is, the protocol constitutes a zero knowledge proof as explained in the next section), the "random" nature of ϵ, a precise estimate of the confidence with which the verifier will be convinced that the image in dispute actually originated from the alleged buyer and finer details of implementation using finite field arithmetic. We defer addressing these technicalities to the full version of this paper.

3 Zero Knowledge Proofs and Digital Watermarks

The protocol presented in the previous section closely follows a well known tool in cryptography called *zero knowledge proofs*. A refreshingly non-mathematical introduction to zero knowledge proofs is provided in [8]. Zero knowledge proof systems is an active area in cryptography and a formal and detailed introduction to it can be found in the texts [7, Ch. 13] and [6, Ch. 10]. Informally, a zero knowledge proof system allows one person, Peggy, to convince another person, Vic, of some fact without revealing any information about the proof. At the end of the protocol, Vic is "completely convinced" of the same fact, but does not gain any additional knowledge whatsoever.

In [3] Kinoshita Hirotsugu attempted to use the zero knowledge interactive proofs to assert ownership rights on an image. He used the well known zero knowledge interactive proof for the graph isomorphism problem presented in [2]. In this section, we briefly review his scheme and show that it is fundamentally flawed.

Kinoshita's scheme works in the spatial domain of the image. Essentially, he generates a graph with say n nodes, called the region graph G_r, from the most significant bits of the pixels of the image in a fixed manner. He then applies a permutation σ on n points to the region graph G_r to obtain an isomorphic graph called the concealed graph G_c. He then conceals the graph G_c in the least significant bits of the pixels. To assert ownership rights over the image, it is suggested that one could extract the region graph from the most significant bits and the concealed graph from the least significant bits and then the owner could demonstrate that these two graphs are isomorphic to each other without revealing the permutation σ using the zero knowledge interactive proof.

While the zero knowledge interactive proof for the graph isomorphism problem presented in [2] is perfect, the way it is used here is fundamentally flawed. The first problem is that this watermarking scheme is not robust. The concealed graph is encoded into the least significant bits of the pixels. The adversary can always modify the least significant bits thus preventing the real owner from proving his ownership of the image. More importantly, the adversary can construct the region graph G_r from the most significant bits of the pixels in exactly the same manner as the owner. The adversary can then apply a permutation ρ, known only to himself, to the region graph G_r and obtain an isomorphic graph $G_{c'}$. He can then embed $G_{c'}$ into the least significant bits of the image and can claim that the image actually belongs to him. Moreover he can prove so by using the very same zero knowledge interactive proof for graph isomorphism.

The above example shows that one has to be very careful in applying the subtle concept of zero knowledge interactive proofs to practical problems. The question that arises is whether the protocol we have presented is zero knowledge. We believe this is true and hope to present a formal proof to this effect in the full version of this paper.

4 Concluding Remarks

In copy deterrence watermarking schemes, it is important to be able to demonstrate the presence of a watermark in an image without revealing the watermark. In this paper, we developed a novel way of demonstrating the presence of a watermark in an image without revealing information about the watermark that could lead to the possibility of the adversary removing the watermark and re-selling multiple copies of the image with impunity.

For the sake of brevity, we exclusively focused on the problem of demonstrating the presence of a watermark in an image without revealing it. Some other aspects of copy deterrence watermarking schemes such as preventing the ability of a malicious seller to frame the buyer are discussed in [5]. Indeed, the protocol presented here could be coupled with the buyer-seller protocol presented in [5] to form a more comprehensive solution to the problem of copy deterrence.

In addition to copy deterrence applications, the fundamental problem that we have pointed to in this paper also applies to watermarking for ownership assertion. Here, current techniques assume that the watermark needs to be revealed in order to assert ownership. This problem could also be addressed by a similar protocol. We hope to report on such a protocol in future work.

References

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

[2] O. Goldreich, S. Micali, and A. Wigderson, "Proofs that yield nothing but their validity or all languages in NP have zero-knowledge proof systems," *Journal of the Association for Computing Machinery*, **38**, 691–729, 1991.

[3] Hirotsugu Kinoshita, "An Image Digital Signature System with ZKIP for the Graph Isomorphism Problem," *Proceedings of the IEEE International Conference on Image Processing — ICIP'96*, **VIII**, 247–250.

[4] Nasir Memon and Ping Wah Wong, "Protecting Digital Media Content: Watermarks for Copyrighting and Authentication," *Communications of the Association for Computing Machinery*, 35–42, July 1998.

[5] Nasir Memon and Ping Wah Wong, "A Buyer-Seller Watermarking Protocol," *Proceedings of the Second Workshop on Multimedia Signal Processing — MMSP'98.*

[6] A. J. Menezes, P. C. van Oorschot and S. A. Vanstone, Handbook of Applied Cryptography, *CRC Press* (1997).

[7] D. R. Stinson, Cryptography - Theory and Practice, *CRC Press* (1995).

[8] J. J. Quisquater, L. Guillo and T. Berson, "How to explain zero knowledge protocols to your children," *Advances in Cryptology — CRYPTO'89, Lecture Notes in Computer Science*, **435**, 628–631.