Security Models of Digital Watermarking

Qiming Li and Nasir Memon

Polytechnic University

Abstract. Digital watermarking, traditionally modeled as communication with side information, is generally considered to have important potential applications in various scenarios such as digital rights managements. However, the current literature mainly focuses on robustness, capacity and imperceptibility. There lacks systematic formal approach in tackling secure issues of watermarking. One one hand, the threat models in many previous works are not sufficiently established, which result in somewhat superficial or even flawed security analysis. On the other hand, there lacks a rigorous model for watermarking in general that allows useful analysis in practice. There has been some efforts in clearing the threat models and formulate rigorous watermarking models. However, there are also many other cases where security issues are lightly or incorrectly treated. In this paper, we survey various security notions and models in previous work, and discuss possible future research directions.

1 Introduction

A digital watermarking scheme, in general, is a set of algorithms that allow us to embed some information (i.e., *watermarks*) into some host signal (or *cover objects*) in such a way that these watermarks can later be extracted or detected, even if the cover objects are corrupted by a small amount of permissible noise.

A watermarking scheme usually consists of three major components. A *watermark generator* generates desired watermarks for a particular application, which are optionally dependent on some keys. An *embedder* embeds the watermark into the cover object, sometimes based on an embedding key. A *detector* is responsible for detecting the existence of some predefined watermark in a cover object, and sometimes it is desirable to extract an message from the watermarked cover object.

Watermarking schemes are potentially for a number of applications. For example, authentication and tamper detection, ownership and copyright protection, and fingerprinting and traitor tracing. Clearly, it is very important to address the security issues in these applications. However, many of the previous works on digital watermarking mainly focus on the robustness, where the attacker model is often over simplified. In particular, it is often assumed that the attacker only adds some sort of random noise to the cover objects, hence robustness against random noise would be sufficient to withstand watermark removal attempts. However, in this case nothing can be concluded for noise types that are not considered, or for smart attackers with carefully designed modifications, especially

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 60–64, 2007.

when a watermark detector is publicly available. Zero-knowledge watermark detection is proposed to limit the information leakage due to public watermark detectors [3]. However, it has been increasingly evident that a capable attacker with the access to the detection oracle would be able to remove watermarks without much difficulty even when the watermark embedding algorithm is not known, as shown by the results of the *Break Our Watermarking System* (BOWS) contest [5].

Furthermore, even if a watermarking scheme is perfectly robust, it is still possible to launch protocol attacks in particular applications. For instance, when watermarks are used as evidence of ownership, it is important that an attacker cannot hinder the claim of ownership from legitimate owners, which can be achieved without estimating or removing the watermark. These attacks are often referred to as *invertibility attacks* [6] or, in a more general form, *ambiguity attacks* [2]. In these attacks, the attacker finds a watermark that is already detectable in a given work, and invert the watermark embedding to obtain a fake original, so as to make a forged claim of ownership that may be difficult to distinguish from that of the real owner. Later work regarding these attacks include [11, 12, 9, 10, 13].

Another type of protocol attacks is called *copy attacks*, where the attacker attempts to copy the watermark in a given cover object to a dissimilar object without introducing much distortion to it [8]. There is some work addressing this type of attacks, including [4, 1, 7].

To address watermarking security issues in a formal and rigorous manner, one can take one of the following two approaches. First, we can examine each particular application scenario, consider a relatively narrow range of attacks, and formulate the security problems for each type of the attacks. Alternatively, we can start by giving a formal definition and model for watermarking itself, and see if we can prove general results that would apply to a wide range of scenarios. Most previous formal approaches belong to the first category, such as [2, 9], and only recently some attempts has been made using the second approach [7].

In this paper, we are going to discuss in detail some of the security notions, and explore possible future research directions.

2 Security Against Protocol Attacks

Craver et al. propose a method to combat invertibility attacks by generating the watermark in a one-way manner [6]. In particular, given a cover object, a one-way hash function is applied to obtain a hash value, and then use a pseudo-random number generator to expand the hash value into a watermark, which is then embedded into the object. It is conjectured that an attacker, after finding a detectable watermark in a cover object first, would have to invert the one-way hash function to make a ownership claim.

Ramkumar et al. [12] point out that such a conjecture may not hold, since an attacker could always perturb the watermarked object and apply the hash function to obtain a random watermark, and then check if the watermark is detectable. Hence, if the false-positive of the underlying watermarking scheme is high, an

attacker could succeed with high probability. An improved scheme is proposed in [12], where the original is required during the ownership proof, and the attacker has to make sure that the true owner's watermark cannot be detected in his fake original. This technique would make a random attacker infeasible.

Observing that it is difficult to design non-invertible watermarking schemes, a scheme based on a trusted third party (TTP) is proposed in [2]. Although provably secure, it may be difficult to find such TTP in practice.

It is later shown that it is possible to build non-inverible schemes without a trusted third party [9]. Their scheme involves a cryptographically secure pseudo-random number generator in the watermarking generation process, which is similar to the use of hash function in [6]. An important difference is that the need to distinguish *valid* and *invalid* watermarks is highlighted, and by limiting the number of valid watermarks to be only a negligible fraction of the total number of possible watermarks, it is then possible to prove the security of the scheme, together with the assumption that the false-positive rate is negligible. A zero-knowledge version of the scheme appears in [10].

Noting that low false-positive rate is essential in the security of non-invertible watermarking schemes, Sencar and Memon [13] propose the use of multiple watermarks instead of a single watermark in spread-spectrum based watermarking scheme to bring down the false-positive rate.

3 Formal Watermarking Model

As we mentioned earlier, there are two approaches dealing with watermarking security. Basically, we can either establish rigorous security for a particular application, or try to give a general formal model for watermarking and try to prove general results.

While the first approach may give sound and provably schemes in particular application scenarios, the implications of such techniques are often limited in the sense that new analysis and new proofs are often required to deal with different application scenarios and/or different attackers.

For example, although the security prove in [9] is sound, the scheme is based on a spread-spectrum watermarking scheme, and it is not straightforward to see how to adapt the proof for other types of watermarking schemes (e.g., QIM). Also, the use of a cryptographically secure pseudo-random number generator may not be suitable in some applications. However, from the proof itself it may be hard to see how the security is affected if the generator is replaced by other one-way functions.

The second approach ([7]) basically defines an *ideal* watermarking scheme, where the cover objects are in a well-defined metric space, and a point in the space is declared as watermarked if and only if it is near a watermarked object. Hence, a secure watermarking scheme can be defined as one that behaves very closely as the ideal counterpart. If such watermarking scheme can be built, it is clear that it is robust and resistant to copy attacks (which requires a distant object to be watermarked).

This approach may seem comprehensive and more formal at first, it however remains somewhat too theoretical and gives little insight or guideline regarding how to design such schemes in practice. For example, in real applications the most difficult part is actually the definition of the similarity measure (or distance function), which has to capture the perceptual characteristics of the data, and also take into consideration all possible permissible noise. Unfortunately, such formal model would not be able to tell a developer how to define such a metric space properly, let alone a secure scheme on top of it.

Furthermore, the model in [7] does give some positive results on how to construct secure watermarking schemes in some arbitrary metric space, but the main idea seems to make the scheme secure by making it difficult to sample in the neighbors of a watermarked object. Although this leads to provable security, it is hard to see how such difficulty in sampling can be imposed in practical scenarios.

4 Conclusions

To formulate the security problems and to design techniques to tackle them is a tricky business in digital watermarking. Many previous approaches are more or less heuristic, which often lead to schemes that are later proved insecure. Hence, we are going to need more rigorous approaches to better assess the security.

There are currently two main categories of approaches. One is application and attack specific, but easier to be applied. The other is more high level and theoretical, but allows more general results to be proved.

To achieve security in watermarking in general, we need a blend of these two approaches, where the results are rigorous and general, yet they allow detailed analysis of security in practice.

References

[1] Adelsbach, A., Katzenbeiser, S., Veith, H.: Watermarking schemes provably secure against copy and ambiguity attacks. In: Proceedings of the 2003 ACM Workshop on Digital Rights Management, pp. 111–119 (2003)

[2] Adelsbach, A., Katzenbeisser, S., Sadeghi, A.: On the insecurity of non-invertible watermarking schemes for dispute resolving. In: Kalker, T., Cox, I., Ro, Y.M. (eds.) IWDW 2003. LNCS, vol. 2939, Springer, Heidelberg (2004)

[3] Adelsbach, A., Sadeghi, A.: Zero-knowledge watermark detection and proof of ownership. In: Moskowitz, I.S. (ed.) Information Hiding. LNCS, vol. 2137, pp. 273–288. Springer, Heidelberg (2001)

[4] Barr, J., Bradley, B., Hanningan, B.T.: Using digital watermarks with image signatures to mitigate the threat of the copy attack. In: ICASSP, pp. 69–72 (2003)

[5] Break our watermarking system contest. `http://lci.det.unifi.it/BOWS/`

[6] Craver, S., Memon, N., Yeo, B.L., Yeung, M.M.: Resolving rightful ownerships with invisible watermarking techniques: Limitations, attacks, and implications. IEEE Journal on Selected Areas in Communications 16(4), 573–586 (1998)

[7] Hopper, N., Molnar, D., Wagner, D.: From weak to strong watermarking. In: Theory of Cryptography Conference (2007) http://eprint.iacr.org/2006/430.
[8] Kutter, M., Voloshynovskiy, S., Herrigel, A.: The watermark copy attack. In: Electronic Imaging 2000, Security and Watermarking of Multimedia Content II, vol. 3971 (2000)
[9] Li, Q., Chang, E.-C.: On the possibility of non-invertible watermarking schemes. In: Fridrich, J. (ed.) IH 2004. LNCS, vol. 3200, pp. 13–24. Springer, Heidelberg (2004)
[10] Li, Q., Chang, E.-C.: Zero-knowledge watermark detection resistant to ambiguity attacks. In: ACM Multimedia Security Workshop (2006)
[11] Qiao, L., Nahrstedt, K.: Watermarking schemes and protocols for protecting rightful ownership and customer's rights. Journal of Visual Communication and Image Representation 9(3), 194–210 (1998)
[12] Ramkumar, M., Akansu, A.: Image watermarks and counterfeit attacks: Some problems and solutions. In: Symposium on Content Security and Data Hiding in Digital Media, pp. 102–112 (1999)
[13] Sencar, H.T., Memon, N.: Combatting ambiguity attacks via selected detection of embedded watermarks. IEEE Transactions on Information Forensics and Security