ON THE SECURITY OF MESH-BASED MEDIA HASH-DEPENDENT WATERMARKING AGAINST PROTOCOL ATTACKS

Chun-Shien Lu and Chia-Mu Yu

Institute of Information Science, Academia Sinica, Taipei, Taiwan 115, ROC

ABSTRACT

A common way of resisting protocol attacks is to employ cryptographic techniques so that provable security can be retained. However, some desired requirements of watermarking such as blind detection and robustness are lost. This paper studies the issue of security against protocol attacks based on a mesh-based media hash-dependent image watermarking approach while maintaining the aforementioned requirements. Our main contributions include (1) media hashing instead of cryptographic hashing is used so that blind detection is still satisfied; (2) robustness against signal processing attacks is retained; (3) the difficulty of resisting ambiguity attack is derived to be equivalent to that of resisting challenging geometric attacks including cropping with larger parts discarded and rotation with larger degrees so that an acceptable trade-off between false positive and false negative can be achieved.

1. INTRODUCTION

In digital watermarking, robustness against attacks is the critical issue affecting the practicability of a watermarking system. Usually, attacks can be classified into four categories: (1) removal attacks; (2) geometric attacks; (3) cryptographic attacks; and (4) protocol attacks. Among them, resistance to signal processing attacks (i.e., removal and geometric attacks) has been widely studied [6, 10] by inserting multiple watermarks locally. However, studies of resistance to protocol attacks are rare.

Protocol attacks are devoted to foiling a watermarking system. At present, copy attack and ambiguity attacks are known to belong to protocol attacks. After applying the copy attack, a watermark can be detected from an image that was not watermarked before to create the false positive problem. In ambiguity attack, more than one rightful owners can be identified to create the dispute problem in claiming the ownership. Although the importance of security against the protocol attacks has been recognized, its studies are mostly found in the cryptographic field.

Corresponding author: Dr. C. S. Lu (lcs@iis.sinica.edu.tw)

In order to provide security against protocol attacks, an intuitive way is to incorporate the principle of cryptography with watermarking so that provable security can be retained. In this manner, the DES technique [7] and the one-way-hashing function [9] were proposed to achieve non-invertibility. However, the main weakness of directly employing cryptographic techniques is their inherent fragility, which implies that even a slight change of an input will result in a totally different output. This poor error-resilience cannot fit the need of robustness in digital watermarking. Furthermore, the operation of one-way-hashing on an original data inherently leads to non-blind detection, which violates the basic requirement of digital watermarking.

An alternative is to employ the cryptographic signature of trusted third party [1] by exploiting the property of provable security based on the assumption that the attacker cannot query the trusted third party. However, an undesirable side effect that yields longer cryptographic signatures and watermarks is produced. Such an obstacle discourages [1] from being a robust watermarking method in resisting geometric attacks [6, 10].

On the other hand, it can also be observed that the existing methods that were proposed to deal with the protocol attacks lost the resistance to signal processing attacks. This is partially due to that either the cryptography-based schemes are over emphasized or the trade-off between resistance to signal processing attacks and resistance to protocol attacks is not achieved. Thus, taking resistance to both the signal processing attacks and protocol attacks into consideration has not been found in the literature. In view of these, we present a strategy of overcoming protocol attacks based on a mesh-based content-dependent image watermarking scheme [6] that has been extensively verified to resist extensive signal processing attacks.

2. MESH-BASED MEDIA HASH-DEPENDENT WATERMARKING

In this paper, our discussion of security against protocol attacks will be built based on a mesh-based content-dependent image watermarking scheme [6] so that resistance to both

0-7803-9332-5/05/$20.00 ©2005 IEEE

signal processing and protocol attacks can be achieved simultaneously. Based on robust feature extraction and mesh generation processes, a cover image is first divided into a set of triangular meshes, $Mesh = \{M_i\}_{i=1,2,\cdots,M}$, where M denotes the number of meshes. For each mesh M_i, it is treated as an embedding unit and is embedded with a content-dependent watermark. In addition, with respect to the set of meshes $Mesh$ a set of media hashes $Hash = \{MH_i\}_{i=1,2\cdots M}$ is extracted and has been verified to be robust against extensive geometric attacks [4]. The proposed mesh-based media hash-dependent watermark $MMHW_i$ is composed of a watermark K generated using a secret key and a MH_i as

$$MMHW_i = S(W, MH_i), \tag{1}$$

where S is a secret key-based shuffling function, which is used to control the combination of W and MH_i. In addition to resisting collusion attack and copy attack [5], here we are interested in the problem of how the mesh-based media hash-dependent watermark (MMHW) can be used to deal with the ambiguity attack. We can also observe from Eq. (1) that error-resilient media hash instead of fragile cryptographic hash is adopted in this paper. Furthermore, our watermark W is not dependent on the cover image, instead W is made to (highly) correlated with the cover image by incorporating the media hash.

3. SECURITY AGAINST PROTOCOL ATTACKS

Fig. 1 depicts the ambiguity attack, where Alice is the rightful owner and Bob is the attacker who exploits Alice's stego image as the available information to create his fake original image and fake watermark. The ambiguity attack is called a Single Watermarked Image Counterfeit Original (SWICO) attack if the stego images generated by Alice and Bob are the same and is otherwise called a Twin Watermarked Images Counterfeit Original (TWICO) attack [3]. When our watermarking method is used by Alice, the embedding component will embed a mesh-based media hash-dependent watermark for each mesh. According to the Kerckhoff's principle, the attacker Bob is assumed to fully know the watermarking algorithm adopted by Alice except for the secret key. Accordingly, how much available information the attacker Bob can have and what kind of information Bob cannot know in advance is specifically described as follows.

Available information: 1. the stego image I^w; 2. the watermarking algorithm (so, the MMHWs extracted by means of Wiener filtering I^w [6], $\left\{MMHW_i^f\right\}_{1\le i\le M}$, can be known); 3. the media hashing algorithm.

Un-available information: 1. the secret key used to generate the watermark W; 2. the secret key used for shuffling.

If a SWICO attack is considered as the ambiguity attack, Bob can try to forge his original image and watermark under the constraint that the stego image is kept unchanged. As a result, the attacker Bob can be successfully to construct an ambiguity attack only if MMHW can be correctly forged. His success relies on how to inverse our shuffling, as described in Sec. 3.1. On the other hand, if a TWICO attack, as proposed by Ramkumar and Akansu [8], is considered, its success depends on the potential higher false positive of robust watermarking instead of trying to break the non-invertibility, as described in Sec. 3.2.

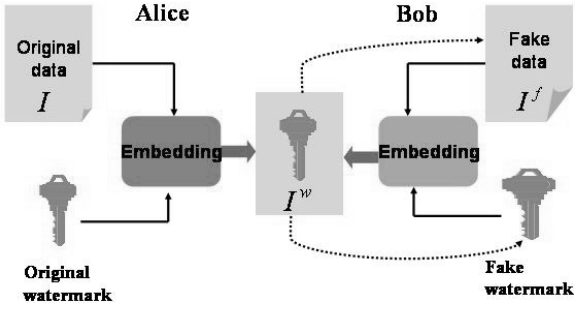

Fig. 1. **Ambiguity attack [3].**

3.1. Security vs. Non-invertibility

Since the proposed media hashing scheme is robust in terms of lower bit error rates (BERs), Bob can (slightly) process the stego image I^w to get a so-called fake original image I^f. Here, I, I^w, and I^f are perceptually similar according to the principle of media hashing [4]. Therefore, it can be assumed that the generated sets of media hashes of I, I^w, and I^f are also similar in terms of lower BERs. However, it cannot be guaranteed that the BERs can be ideally reduced to zero. Let $\left\{MH_i^f\right\}_{1\le i\le \bar{M}}$ be the set of extracted hashes, where $\bar{M}$ denotes the number of meshes in a fake image I^f.

In order to break the proposed watermarking scheme by means of SWICO, the attacker Bob needs to show he can produce $MMHW_i^f$ from his watermark and MH_i^f [3]. Therefore, his success relies on the probability of inversing the shuffling function. In this study, we consider the ciphertext-only attack and the known-plaintext attack.

Let the length of a mesh-based media hash-dependent watermark be N. It is not hard to calculate that the time-complexity of inversing one shuffling for the ciphertext-only attack is $O(N!)$ and, thus, the total time-complexity is $O(M \cdot N!)$. In practice, $N = 128$ is adopted in our mesh-based media hash-dependent watermarking scheme [6]. Thus, it is computationally impossible to inverse the employed shuffling with no bit errors.

As for the known plaintext attack, we refer to the assumption that the content-dependent watermark and its corresponding media hash can be perfectly extracted from the

stego image. When an attacker compares a pair of content-dependent watermark and its corresponding pair of media hashes, partial permutation is revealed. More specifically, a desirable permutation is partitioned into two subsets, where one set stands for the part that two hashes (and the CDWs) are the same and the other set represents those that are different. Based on this step, a total number of possible permutations is greatly reduced. By executing the above steps several times, it is possible to achieve the recovery of shuffling. In the worst case for attackers, at least $N-1$ such pairs are needed to fully achieve inverse *one* shuffle under the condition that the extracted content-dependent watermarks and media hashes don't incur errors. This imples that for each comparison a one-to-one position mapping of shuffling is found. Thus, the time complexity of inversing all shuffles in the worse case for an image is $O(M \cdot N)$. In the best case for attackers, $\lceil log_2^N \rceil$ pair comparisons are sufficient to break *one* shuffle under the condition that such pairs can be found. Thus, the time complexity of inversing all shuffles in the best case for an image is $O(M \cdot \lceil log_2^N \rceil)$.

Overall, the main differences distinguishing our approach from others in dealing with SWICO are that the length of a hidden watermark is not increased and blind detection is achieved.

3.2. Security vs. False Positive

The requirements of successfully achieving ambiguity attacks can be relaxed by exploiting the possibly high false positive probability of a robust watermarking method, i.e., the BER obtained by comparing a fake watermark and an extracted watermark can be larger than 0 but smaller than a threshold. The spirit is to avoid to run the risk of having difficulty in breaking cryptographically secure one-way hash function or signature scheme. This paradigm of such a protocol attack was first proposed by Ramkumar and Akansu [8] and was later extended by Adelsbach *et al.* [2].

In our watermarking scheme [6], bit detection is treated as an independent random Bernoulli trail with the probability p_b, which is defined to be the probability that the bit b (-1 or 1) occurs. Theoretically, the probability of truly detecting a watermark in a mesh when BER$\leq Th_{mesh}$ holds can be represented as:

$$p_{fp_{mesh}} = \sum_{j=(N-N\times Th_{mesh})}^{N} (\tbinom{N}{j} {p_b}^j (1-p_b)^{N-j}. \quad (2)$$

Eq. (2) also specifies the probability that a watermark can be found in a mesh that was, in fact, not watermarked before. In order to reasonably determine Th_{mesh}, $p_{fp_{mesh}}$ of Eq. (2) is better set to be consistent with practical results. The Corel image database containing 20000 images is used as inputs to our watermarking system and the obtained detection results show that $Th_{mesh} = 0.375$ is a reasonable choice that the probability of deciding a mesh to have been watermarked is 0.003. However, it cannot be simply concluded that a watermark is falsely detected with probability 0.003. This is because in our scheme multiple mesh-based watermarks are embedded to resist geometric attacks, the final decision of determining the existence of watermark is quite different from those methods that only embed a single watermark. In the following, the condition of determining the existence of a watermark for multiple watermarking methods will be derived.

Recall that M is the number of meshes in an image. Let D_M be the number of meshes detected to have been watermarked. Similarly, the probability of determining a suspect image to have been watermarked is derived as:

$$p_{fp_{image}} = \sum_{i=D_M}^{M} (\tbinom{M}{i} {p_{fp_{mesh}}}^i (1-p_{fp_{mesh}})^{M-i}. \quad (3)$$

In fact, Eq. (3) also reveals the probability that a random image will be "wrongly" determined to have been watermarked and/or attacked. *Furthermore, this also implies that different attacks lead to different $p_{fp_{image}}$'s, i.e., a more challenging attack will generate a higher false positive probability.* In order to claim the presence of a watermark with strong confidence (without causing non-negligible false positive), $p_{fp_{image}}$ should be low enough. On the other hand, $p_{fp_{image}}$ should be large enough to achieve robustness. Here, a reasonable threshold Th_{image} is required to satisfy the trade-off between robustness and false positive. Again, the Corel image database is adopted as a training database to derive Th_{image}.

In the following, we shall discuss the impact of our each watermarking step on robustness and security.

3.2.1. Impact of Our Watermarking on Security

As described in Sec. 2, in addition to media hashing feature point extraction and denoising-based blind detection are recognized as two main factors that may affect the performance of our method. Since the robustness of our media hashing has been verified in [4], it is not discussed here again. According to the experimental results shown in [6], it is important to know how many meshes of a stego image can be detected to contain watermarks without involving the effect of attacks. Two experiments are performed based on the conditions that (i) the feature points and media hashes extracted from the original image are directly applied to the stego image, which means that feature point extraction is perfect and we are only interested in the effect of Wiener filtering; (ii) all the processes are the same as those described in Sec. 2, which means that by comparing the results obtained from conditions (i) and (ii) we can know the effect of feature point extraction (and media hashing). The results of these two experiments are depicted in Table 1.

Table 1. Impact of feature point extraction and denoising-based blind detection on the performance our watermarking method. "D_M/M" denotes "number of detected watermarked meshes/number of total meshes."

image	Condition (i)		Condition (ii)		Average shift of feature points
	D_M/M	$p_{fp_{image}}$	D_M/M	$p_{fp_{image}}$	
Baboon	67/103	6.2e-142	7/113	6.4e-008	4.1 pixels
Lena	88/100	9.8e-208	32/106	2.0e-054	2.6 pixels
Pepper	95/107	5.1e-225	55/109	7.3e-108	1.6 pixels

As we can see from Table 1 that when condition (i) is considered, denoising-based blind detection slightly affect the detection results. However, when condition (ii) is considered, the number of meshes detected to contain watermarks, when compared with the results obtained in condition (i), is dramatically reduced. This obviously implies that the stability of feature points play a major role in affecting the performance of our watermarking method. More specifically, it can be observed from Table 1 that the average displacement (in pixels) of feature points sufficiently explains the obtained detection results. According to condition (ii) of Table 1, we know that the total number of meshes that can be effectively exploited by attackers is reduced from M to D_M. In addition, the probability of an attacker's success as defined in Eq. (3) must be smaller than the threshold Th_{image}. *Based on these, we further find that the probability of an attacker's success is equivalent to that of our watermarking method in successfully detecting the existence of watermarks from the challenging attacks such as cropping with larger parts discarded and rotation with larger degrees.* In other words, the trade-off between false positive and false negative can be achieved reasonably.

On the other hand, due to the uncertainty (instability) of feature point extraction, the fake stego image generated by means of [8] is not the same as the stego image generated by Alice. Therefore, Ramkumar and Akansu's attack is treated as a Twin Watermarked Images Counterfeit Original (TWICO) attack [3] based on the prerequisite that our watermarking method is the target for attacking.

4. CONCLUSION

Protocol attacks were previously discussed without taking a practical watermarking method into consideration so that the achievable success is restricted to the common weakness of existing watermarking methods. In this paper, we investigate the issue of security against protocol attacks based on a mesh-based media hash-dependent image watermarking scheme that can achieve blind detection and has been verified to be robust against signal processing attacks. According to this study, our observations are summarized as follows: (1) the common use of cryptographic hashing found in the literature is not suitable for both owners (due ot its full fragility) and attackers (due to its recognized computational hard to break); (2) from an attacker's viewpoint, it is hard to succeed with SWICO because the permitted slight fragility of media hashing is not beneficial to attackers, as described in Sec. 3.1; and (3) although it is easier to construct TWICO than SWICO, the fake watermarks generated using TWICO cannot be guaranteed to be mostly detected if a robust watermarking method having lower false positive probability is considered, as described in Sec. 3.2.

5. REFERENCES

[1] A. Adelsbach, S. Katzenbeisser, H. Veith, "Watermarking Schemes Provably Secure Against Copy and Ambiguity Attacks," *ACM Workshop on Digital Rights Management*, pp. 111-119, Washington DC, 2003.

[2] A. Adelsbach, S. Katzenbeisser, A.-R. Sadeghi, "On the Insecurity of Non-Invertible Watermarking Schemes for Dispute Resolving," *Int. Workshop on Digital Watermarking*, LNCS 2393, pp. 355-369, Seoul, Korea, 2003.

[3] S. Craver, N. Memon, B. L. Yeo, and M. M. Yeng, "Resolving Rightful Ownership with Invisible Watermarking Techniques: Limitations, Attacks, and Implications," *IEEE Journal on Selected Areas in Comm.*, Vol. 16, No. 4, pp. 573-586, 1998.

[4] C. Y. Hsu and C. S. Lu, "A Geometric-Resilient Image Hashing System and Its Application Scalability," *Proc. ACM Multimedia and Security Workshop*, pp. 81-92, Germany, 2004.

[5] C. S. Lu and C.Y. Hsu, "Content-Dependent Anti-Disclosure Image Watermark," *Proc. 2nd Int. Workshop on Digital Watermarking,* LNCS 2939, pp. 61-76, Seoul, Korea, 2003.

[6] C. S. Lu, S. W. Sun, C. Y. Hsu, and P. C. Chang, "Robust Mesh-based Media Hashing-dependent Image Watermarking with Resistance to Both Geometric Attack and Watermark-Estimation Attack," journal version in preparation (the partial version was published in *Proc. SPIE: Security, Steganography, and Watermarking of Multimedia Contents VII*, 2005).

[7] L. Qiao and K. Nahrstedt, "Watermarking Schemes and Protocols for Protecting Rightful Ownership and Customer's Rights," *Journal of Visual Communication and Image Representation*, Vol. 9, No. 3, pp. 194-210, 1998.

[8] M. Ramkumar and A. Akansu, "Image Watermarks and Counterfeit Attacks: Some Problems and Solutions," *Symposium on Content Security and Data Hiding in Digital Media*, pp. 102-112, 1999.

[9] M. D. Swanson, M. Kobayashi, and A. H. Tewfik, "Multimedia Data Embedding and Watermarking Techniques," *Proceedings of the IEEE*, Vol. 86, No. 6, 1998.

[10] S. Voloshynovskiy, F. Deguillaume, and T. Pun, "Multibit digital watermarking robust against local nonlinear geometrical distortions," *Proc. IEEE Int. Conf. Image Processing,* Thessaloniki, pp. 999-1002, Oct. 2001.