Security Analysis of Transformed-Key Asymmetric Watermarking System

I-Te Chen and Yi-Shiung Yeh

***Abstract*—Asymmetric watermarking schemes have been widely discussed recently. Transformed-key asymmetric watermarking (TKAW) system is one of the schemes, which was proposed by Choi *et al.* This letter discusses the weaknesses of TKAW. In TKAW, the inner product of the received signal and public key almost equals to zero. As a result, it cannot resist projection attack. To prove this point, we will demonstrate how to find the most relevant unwatermarked signal $\widetilde{y}$ for a given watermark embedded signal y and prove that $\|y - \widetilde{y}\| \leq |\alpha|$, where α is a constant adjusted to a value that makes the watermark imperceptible. Moreover, we also find that the feature $y - \widetilde{y}$, a predicted watermark of a stego image, can be copied from a stego image to another image, which means that the TKAW system cannot resist copy attack either.**

***Index Terms*—Asymmetric watermarking, copy attack, projection attack, transformed key.**

I. Introduction

DIGITAL watermarking is widely used in, but not limited to, authentication, copyright protection, and fingerprint. Most watermarking schemes are symmetric, which means the keys used to embed and detect watermark are identical. Therefore, the symmetric watermarking schemes cannot be applied in some applications; several asymmetric watermarking schemes have been proposed [1]–[5]. In this letter, we discuss the weaknesses of a proposed asymmetric scheme—transformed-key asymmetric watermarking (TKAW) system [1]—by demonstrating the way of finding the unwatermarked signal $\widetilde{y}$ for a given watermark-embedded signal y by means of projection attack. Furthermore, we will also prove that $\|y - \widetilde{y}\| \leq |\alpha|$, where α is a constant adjusted to a value that makes the watermark. Moreover, we will show that in TKAW, the vector of $y - \widetilde{y}$, which contains enormous watermark components, can be copied from a stego image to another image, which is called a copy attack.

II. Related Work

A. TKAW System

Choi *et al.* proposed the TKAW system in 2004 [1], which provided asymmetry through a transform matrix. In the proposed TKAW, let $\{u_i : i = 1, \ldots, k\} \subseteq \mathbf{R}^n$ be an orthonormal set, $\mathbf{A}$ be an invertible matrix of size $n \times n$, and $\mathbf{A}^{-1}$ and $\mathbf{A}^{\mathrm{T}}$ denote the inverse and transpose of $\mathbf{A}$, respectively. The normalized secret key $\mathbf{w}_s$ and public key $\mathbf{w}_p$ of the TKAW system are

$$\begin{aligned} \mathbf{w}_{s,i} &= \frac{\mathbf{A}u_i}{\|\mathbf{A}u_i\|} = r_s\mathbf{A}u_i, \text{ where } r_s = \frac{1}{\|\mathbf{A}u_i\|} \\ \mathbf{w}_{p,i} &= \frac{(\mathbf{A}^{-1})^{\mathrm{T}}u_i}{\|(\mathbf{A}^{-1})^{\mathrm{T}}u_i\|} = r_p(\mathbf{A}^{-1})^{\mathrm{T}}u_i, \text{ where} \\ r_p &= \frac{1}{\|(\mathbf{A}^{-1})^{\mathrm{T}}u_i\|}. \end{aligned}$$

For a host signal x, the watermark embedding process is given by $y = x + \alpha\mathbf{w}_{s,i}$, where α is a constant adjusted to a value that makes the watermark imperceptible. Let $\hat{x} = x + n$ denote a host signal with additive noise n and $\langle a, b\rangle$ denote the inner product of a and b. The public detection on received signal $\boldsymbol{r}_i = x + n + \alpha\mathbf{w}_{s,i}$, which is the watermark-embedded signal y containing additive noise n. The inner product of the public key $\mathbf{w}_p$ and $\mathbf{r}_i$ can be expressed as

$$\begin{aligned} C_{ji} = \langle \mathbf{w}_{p,j}, \boldsymbol{r}_i\rangle &= w_{p,j}^{\mathrm{T}}\boldsymbol{r}_i \\ &= w_{p,j}^{\mathrm{T}}(x + n + \alpha\mathbf{w}_{s,i}) \\ &= w_{p,j}^{\mathrm{T}}(\hat{x} + \alpha\mathbf{w}_{s,i}) \\ &= r_p u_j^{\mathrm{T}}\mathbf{A}^{-1}\hat{x} + \alpha r_s r_p u_j^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{A}u_i \\ &= r_p u_j^{\mathrm{T}}\mathbf{A}^{-1}\hat{x} + \alpha r_s r_p u_j^{\mathrm{T}}u_i. \end{aligned}$$

Choi let $u_j^{\mathrm{T}}\mathbf{A}^{-1}\hat{x} \cong 0$, then $C_{ji} \ll C_{ii}$, for all $i, j = 1, \ldots, k$. Let $E[\cdot]$ denote the expected value. By comparing C_{ji} and threshold $\mathrm{T} = (E[C_{ii}] + 2E[C_{ji}])/3$, the watermark can be detected. Although the TKAW system is an efficient asymmetric watermarking system, it is, after all, a linear transform system, and Miller already found a way to attack a linear transform system in [7, Sec. 2.1]. Furthermore, the TKAW system also needs $u_j^{\mathrm{T}}\mathbf{A}^{-1}\hat{x} \cong 0$, which means it can be defeated by a standard closest point or projection attack [6], [7]. As discussed in [6, Sec. 5], a secure asymmetric watermarking must be able to resist the projection attack.

B. Projection and Copy Attack

In closest point attack [7], also called projection attack in [2], the pirate finds the unwatermarked signal, which is closest to a given watermarked signal, either by means of analysis or an efficient search. This unwatermarked signal can be as close as identical to the original watermarked signal.

The watermark is an imperceptible signal; hence, the watermark is one kind of perturbation to the original image. In the symmetric watermarking system, the detection method is unknown to a pirate; therefore, it is difficult to predict the

Manuscript received August 26, 2005; revised November 18, 2005. This work was supported in part by the National Science Council under Grant NSC-94-2213-E-009-118. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. James E. Fowler.

The authors are with the Department of Computer Science and Information Engineering, National Chiao-Tung University, Hsinchu 30050, Taiwan, R.O.C. (e-mail: itchen@csie.nctu.edu.tw; ysyeh@csie.nctu.edu.tw).

Digital Object Identifier 10.1109/LSP.2005.863677

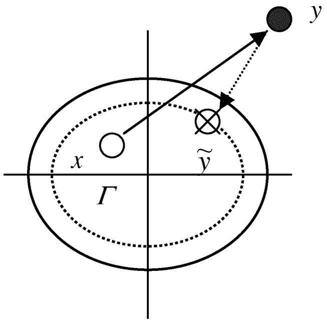

Fig. 1. The relationship among x, y, and $\widetilde{y}$.

direction between watermark-embedded image and original image. However, in the asymmetric watermarking system, the detection method is public; a pirate can try at will to trace the direction between watermark-embedded image and original image and therefore locate the original image. This searching method that locates the perceptible unwatermarked image by tracing the watermark-embedded image is called the closest point attack.

Let x, y denote original image and watermark-embedded image, respectively. The set Γ within the dotted line denotes image collection that cannot be watermark-detected. The closest point attack is equivalent to the procedure of finding the direction from y to $\widetilde{y}$. The relationship among x, y, and $\widetilde{y}$ is illustrated in Fig. 1.

If the detection method is linear, the closest point attack is also called projection attack [7]. In the recent linear asymmetric watermarking techniques, $\mathbf{A}$, which is a linear operator, applies and compares watermark image with watermark by sim value. Sim value $= w^{\mathrm{T}}\mathbf{A}y = (\mathbf{A}^{\mathrm{T}}w)^{\mathrm{T}}y$, where w is the watermark. When applying closest point attack to linear asymmetric watermarking, we find the closest point in the plane of $(\mathbf{A}^{\mathrm{T}}w)^{\mathrm{T}}y = 0$, where $\mathbf{A}^{\mathrm{T}}w$ is the normal vector of this plane.

The solution of this closest point is to justify the project point from the watermark image to this plane. Hence, in linear asymmetric watermarking, the closest point attack is also called projection attack [2]. Tzeng remarks that if we cannot detect some component of watermark using the detection method on the original image in one linear asymmetric watermarking system, that linear asymmetric watermarking system cannot resist projection attack [2].

Choi's method is not only weak to projection attack but also copy attack. Because the watermark-embedded signal is the noise that has little influence upon the original image, when we apply the denoise operator on the watermarked image, the extraction will still contain enormous watermark components. The copy attack is used to predict the watermark from the watermark embedded signal y and to add this predictive watermark to another image [8]. According to copy attack processes, we can derive those predictive watermark signals and copy it to another image. On the other hand, we can, improving the method of copy attack, apply projection attack to calculate the different value between y and $\widetilde{y}$, to prove the existence of the main watermark component and add this feature $y - \widetilde{y}$ to another image. Hence, we prove that the TKAW system can resist neither the projection attack nor the copy attack.

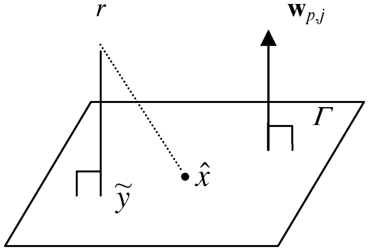

Fig. 2. Relationship among $\hat{x}$, $\boldsymbol{r}$, and $\widetilde{y}$.

III. Projection Attack on TKAW

Choi *et al.* [1] designed an asymmetric watermarking system TKAW with $u_j^{\mathrm{T}}\mathbf{A}^{-1}\hat{x} \cong 0$, which forces $\langle(\mathbf{A}^{-1})^{\mathrm{T}}u_j, \hat{x}\rangle = \langle\mathbf{w}_{p,j}, \hat{x}\rangle = 0$. Because the inner product of $\mathbf{w}_{p,j}$ and $\hat{x}$ is zero, we can find a hyperplane Γ, on which $\hat{x}$ falls. We denote the normal vector of the hyperplane as $\mathbf{w}_{p,j}$, shown as in Fig. 2.

In Fig. 2, $\boldsymbol{r} = y + n$ is the received signal (which is the watermark-embedded signal y combined with an additive noise n), and $\hat{x} = x + n$ denotes the host signal with an additive noise n. Therefore, we can derive the unwatermarked signal $\widetilde{y} = \boldsymbol{r} + t\mathbf{w}_{p,j}$ from the following equation:

$$\langle\mathbf{w}_{p,j}, \widetilde{y}\rangle = (\mathbf{w}_{p,j})^{\mathrm{T}}(\boldsymbol{r} + t\mathbf{w}_{p,j}) = (\mathbf{w}_{p,j})^{\mathrm{T}}\boldsymbol{r} + t\|\mathbf{w}_{p,j}\|^2 = 0$$

and we can derive $t = -(\mathrm{w}_{p,j})^{\mathrm{T}}\boldsymbol{r}/\|\mathrm{w}_{p,j}\|^2)$; therefore

$$\widetilde{y} = \boldsymbol{r} - \frac{(\mathrm{w}_{p,j})^{\mathrm{T}}\boldsymbol{r}}{\|\mathrm{w}_{p,j}\|^2}\mathbf{w}_{p,j}.$$

$\widetilde{y}$ without the watermark is the closest solution against watermark-embedded signal y. Furthermore, $\widetilde{y}$ is the projection of y on the hyperplane Γ. By triangle inequality, we know that

$$\|y - \widetilde{y}\| \leq \|y - x\| = \|\alpha\mathbf{w}_{s,i}\| = |\alpha|$$

which means that the difference between y and $\widetilde{y}$ is less than the watermark energy α. Hence, y and $\widetilde{y}$ are imperceptible in common situations. This means the TKAW system cannot resist projection attack.

Furthermore, the vector $y - \widetilde{y}$ projects r to $\widetilde{y}$; as a result, $y - \widetilde{y}$ contains the main component of the watermark signal. Given the other images G_i with $(\mathbf{w}_{p,j})^{\mathrm{T}}G_i = 0$, we can copy $y - \widetilde{y}$ to other images G_i with the following equation:

$$\boldsymbol{r}_G = G_i + \beta(y - \widetilde{y})$$

where β is a constant adjusted to a value that makes $(y - \widetilde{y})$ imperceptible. From

$$\begin{aligned} C_{ji} = \langle\mathbf{w}_{p,j}, \boldsymbol{r}_G\rangle &= w_{p,j}^{\mathrm{T}}\boldsymbol{r}_G \\ &= w_{p,j}^{\mathrm{T}}\left(G_i + \beta(y - \widetilde{y})\right) \\ &= r_p u_j^{\mathrm{T}}\mathbf{A}^{-1}G_i + \beta r_p u_j^{\mathrm{T}}\mathbf{A}^{-1}(y - \widetilde{y}) \\ &\cong \beta r_p u_j^{\mathrm{T}}\mathbf{A}^{-1}(y - \widetilde{y}) \end{aligned}$$

we know that $\boldsymbol{r}_G$ is a detected watermark, which means the TKAW system cannot resist copy attack either.

IV. Conclusion

To reduce the weaknesses of the current symmetric watermarking system, Choi *et al.* [1] proposed an asymmetric watermarking system called the TKAW system. Unfortunately, the TKAW system, which designed $u_j^{\mathrm{T}}\mathbf{A}^{-1}\hat{x} \cong 0$, is still under the threat of the projection attack [2] and copy attack [8]. We demonstrate in this letter the projection attack on the TKAW system and show how to find the closest unwatermarked $\widetilde{y}$ to the watermark-embedded signal y. In addition, we also show that the difference between y and $\widetilde{y}$ is less than the watermark energy α. Furthermore, the vector $y-\widetilde{y}$, which contains the main component of the watermark signal, can be copied to other images. Therefore, the TKAW system can resist neither projection nor copy attack.

References

[1] H. Choi, K. Lee, and T. Kim, "Transformed-key asymmetric water-marking system," *IEEE Signal Process. Lett.*, vol. 11, no. 2, pp. 251–254, Feb. 2004.

[2] J. Tzeng, W.-L. Hwang, and I.-L. Chern, "An asymmetric subspace watermarking method for copyright protection," *IEEE Trans. Signal Process.*, vol. 53, no. 2, pp. 784–792, Feb. 2005.

[3] J. J. Eggers, J. K. Su, and B. Girod, "Asymmetric watermarking schemes," in *Proc. Tagungsband des GI Workshops Sicherheit in Mediendaten*, Berlin, Germany, Sep. 2000.

[4] R. G. van Schyndel, A. Z. Tirkel, and I. D. Svalbe, "Key independent watermark detection," in *Proc. IEEE Int. Conf. Multimedia Computing Systems*, Florence, Italy, Jun. 1999, pp. 580–585.

[5] T. Furon and P. Duhamel, "An asymmetric public detection watermarking technique," in *Proc. 3rd Int. Information Hiding Workshop*, Dresden, Germany, Oct. 1999, pp. 88–100.

[6] M. Barni, F. Bartolini, and T. Furon, "A general framework for robust watermarking security," *Signal Process.*, vol. 83, pp. 2069–2084, 2003.

[7] M. L. Miller, "Is asymmetric watermarking necessary or sufficient?," in *Proc. European Signal Processing Conf.*, Toulouse, France, 2002.

[8] M. Kutter, S. Voloshynovskiy, and A. Herrigel, "The watermark copy attack," in *Proc. SPIE*, vol. 3971, 2000, pp. 371–380.