7-2012

A Pollution Attack to Public-key Watermarking Schemes

Yongdong WU
Institute of InfoComm Research, Singapore

Robert H. DENG
Singapore Management University, robertdeng@smu.edu.sg

DOI: https://doi.org/10.1109/ICME.2012.73

2012 IEEE International Conference on Multimedia and Expo
http://doi.ieeecomputersociety.org/10.1109/ICME.2012.73

A Pollution Attack to Public-key Watermarking Schemes

Yongdong Wu
Institute for Infocomm Research, Singapore
Email: wydong@i2r.a-star.edu.sg

Robert H. Deng
Singapore Management University, Singapore
Email: robertdeng@smu.edu.sg

***Abstract*—Public-key watermarking schemes are required to possess two desirable properties: allowing everyone to determine whether a watermark exists in an image or not and ensuring high detection probability in case of malicious modification. In this paper we propose an attack which pollutes the watermark embedded in an image with an optimal colored noise so as to fool the detector of the underlying public-key watermarking scheme. We further show how to apply the proposed pollution attack to public-key subspace watermarking schemes to generate pirated images of high quality but of low detection probability. Our experiment results demonstrate that the proposed pollution attack is very effective.**

I. Introduction

Subliminal channels were historically used for military and political purposes. For example, in the 12th century, to overthrow the Yuan dynasty which ruled China, Yuanzhang Zhu, a leader of peasant uprising army, distributed a secret message hidden in moon-cakes (sweet paste-filled cakes). The secret note told people to stage a uprise in united revolt once they saw a lantern burning in their village's watchtower. On a mid-autumn's night (Aug. 15, lunar year), the lanterns were hoisted. The Yuan dynasty was overthrown and Zhu became the first Emperor of the Ming dynasty. Nowadays, moon-cake is a favorite food and an essential part of mid-autumn festival. Its role of hiding secret messages is taken over by digital content since the latter can be stored, transmitted and processed much more efficiently. This technique of embedding messages in digital content is referred to as digital watermarking. Based on the detection methods, watermarking schemes are classified into three categories [1] [1]: symmetric watermarking, weak asymmetric watermarking, strong asymmetric watermarking or public-key watermarking.

A symmetric watermarking scheme (*e.g.*, [3]-[5]) requires the detector to share secrets with the embedder. Hence, only designated users can detect/extract the watermark from an image. Technically, the embedder inserts a secret watermark $\mathtt{w}$ into a host original image with another optional secret so as to produce a watermarked image. To detect or extract the watermark from an unknown image, the detector must have the watermark and the optional secret. For example, the well-known direct sequence spread spectrum watermarking [5] adds a secret watermark $\mathtt{w}$ into some DCT (Discrete Cosine Transform) coefficients of an original image. Without the secret $\mathtt{w}$, the detector can not confirm whether an unknown image is watermarked or not. Besides being vulnerable to the de-synchronization attack and the average attack, symmetric watermarking schemes are also susceptible to the sensitivity attack or the oracle attack [6].

A weak asymmetric watermarking scheme (*e.g.*, [7]) enables a detector to identify a watermark using a detecting key which is derived from the embedding key with a one-way function. In this case, a detector (e.g., a judge or authority) can answer whether an unknown image is watermarked or not, but he can not recover the original image. Thanks to the complicated non-linear functions in the detecting method, Furon's scheme [7] greatly increases the complexity of the oracle attack.

A public-key watermarking (*e.g.*, [8]-[15]) is a special case of asymmetric watermarking as its detecting key is known to all. Therefore, a public-key watermarking enables anyone to determine the existence of a watermark using public parameters. This merit is very useful in some applications such as copyright protection. Unfortunately, because almost all asymmetric watermarking methods are not constructed based on hard mathematics problems as in cryptographic systems, the public parameters leak some information on the embedding secret, and may result in serious security flaws, *e.g.*, scheme [8] is vulnerable to the removal attack [16], and schemes [9][10] are vulnerable to the confusion attack [17].

With a correlation detecting strategy, the public-key subspace watermarking [11] is robust against the projection attack by selecting a watermark which is parallel to the original image feature; however, it generates a watermarked image which is proportional to the original image [18]. Strictly speaking, the subspace scheme [11] is not in the classic watermarking field but in the image searching field. Boato *et al.* [13] improved the subspace scheme [11] so that the watermarks can be selected arbitrarily. They further provided a detailed construction algorithm in [14] on their subspace watermarking method and then applied it to multilevel fingerprinting. The improvement enables everyone to detect the embedded watermark from a pirated image even in the presence of a certain amount of image degradation due to image processing operations, such as filtering, JPEG compression and their combination. Moreover, subspace watermarking schemes [13][14] are provably secure against the projection attack which minimizes the difference

[1]Indeed, it is controversial how watermarking schemes are classified. We omit the publicly detectable watermarks (e.g.,[2]) based on zero-knowledge as it requires multiple rounds of interaction between the detector and the embedder.

between the pirated image and the watermarked one.

Furon *et al*. [1] proposed an attack principle to the public-key watermarking methods assume that the watermark message is very long (*e.g.*, tens thousand bits in [9]). Thus, their attack is not able to defeat any public-key watermarking scheme whose watermark is short, *e.g.*, subspace watermarking. In this paper we propose a new attack called pollution attack by adding optimal colored noise into the watermarked image. In principle, it is not new to add noise to a watermarked image so as to fool the watermark detector, and almost all the proposals for robust watermarking schemes including [13]-[15] analyzed the addition attack with white noise so as to investigate the security of watermarking schemes. As a generation of scheme [18], the present pollution attack generates the noise craftily rather than randomly. Specifically, the noise is selected to minimize the detection value given that the noise energy is bounded. To demonstrate the effectiveness of the proposed attack to the public-key watermarking schemes [13][14], an experimental prototype is used to create pirated images which are of high quality but of low watermark correlation. It shows that the attack is effective.

This paper is organized as follows. Section II presents the framework of the proposed pollution attack. To make the paper self contained, Section III introduces the subspace watermarking scheme [13][14]. Section IV illustrates the attack to the subspace scheme. Section V describes our experiments and results. Section V concludes the paper.

II. Framework of Pollution Attack

A. Model of public-key watermarking

To embed a watermark $\mathtt{w}$ into an original host $\mathbf{I}_o$ so as to generate a watermarked image $\mathbf{I}_w$ with a public-key watermarking scheme, an embedder performs an embedding operation $\otimes$ as $\mathbf{I}_w = \mathbf{I}_o \otimes \mathcal{F}_0(\mathtt{w}, \mathbf{I}_o)$, with some function $\mathcal{F}_0(\mathtt{w}, \mathbf{I}_o)$.

To determine the existence of a watermark $\mathtt{w}$ in an unknown image $\mathbf{I}_u$, the detector will calculate a similarity measure as

$$\lambda_u = sim(\mathcal{F}_1(\alpha, \mathbf{I}_u), \mathcal{F}_2(\alpha, \mathbf{I}_u)),$$

where α, $\mathcal{F}_1(\cdot)$ and $\mathcal{F}_2(\cdot)$ are publicly known. Please note that α may be related to the watermark $\mathtt{w}$ and the original image $\mathbf{I}_o$. If $\lambda_u > \varepsilon$, the watermark $\mathtt{w}$ is identified; otherwise, no watermark exists in the unknown image $\mathbf{I}_u$.

B. Attack method

The pollution attack inserts noise $\mathtt{p}$ into a watermarked content $\mathbf{I}_w$ so as to generate a pirated content $\mathbf{I}_p$ as $\mathbf{I}_p = \mathbf{I}_w \otimes \mathcal{F}_0(\mathtt{p}, \mathbf{I}_w)$.

To fool the detector, the attacker should select the pollution noise $\mathtt{p}$ which meets the following requirements:

R1: High fidelity. The pirated image $\mathbf{I}_p$ should be perceptually similar to the original image, or the watermarked image. To evaluate the image quality, we adopt the PSNR (Peak Signal to Noise Rate) as a fidelity measurement in the following although other quality measures are also applicable.

R2: Low similarity λ_p. A detector can not identify the watermark from a pirated image $\mathbf{I}_p$ generated with the pollution signal $\mathtt{p}$ if and only if λ_p is small.

R3: No spike modification. The modification value should be limited to a small interval so as to eliminate pepper and salt artifacts.

According to the above requirements, the attacker should solve

$$\begin{cases} \mathtt{p} = arg\ \min_{\mathtt{x}} \mid sim(\mathcal{F}_1(\alpha, \mathbf{I}_{\mathtt{x}}), \mathcal{F}_2(\alpha, \mathbf{I}_{\mathtt{x}})) \mid \\ PSNR(\mathbf{I}_p - \mathbf{I}_w) \geq \gamma_0 \\ \theta_l \leq \mathtt{p} \leq \theta_h \end{cases} \quad (1)$$

where $PSNR(\mathtt{y}) = 20(\log_{10} 255 - \log_{10} \parallel \mathtt{y} \parallel)$, and γ_0 is a pre-defined threshold value to evaluate image quality in dB (say, 40dB). The three formulas in Eq.(1) match the requirements R1-R3 respectively.

C. Attack Rationale

Clearly, a watermarking method should guarantee that a detector does not identify a watermark from the original image; otherwise, the watermarked image has no additional information in comparison with the original image. In other words, the detector outputs

$$\lambda_o = sim(\mathcal{F}_1(\alpha, \mathbf{I}_o), \mathcal{F}_2(\alpha, \mathbf{I}_o)) < \varepsilon.$$

Denote $\mathbf{I}_p^{min}$ as the pirated image constructed with $\mathtt{p}$, according to Eq.(1), the detector obtains

$$\begin{aligned} \lambda_p^{min} &= sim(\mathcal{F}_1(\alpha, \mathbf{I}_p^{min}), \mathcal{F}_2(\alpha, \mathbf{I}_p^{min})) \\ &= \min_{\mathtt{x}} sim(\mathcal{F}_1(\alpha, \mathbf{I}_{\mathtt{x}}), \mathcal{F}_2(\alpha, \mathbf{I}_{\mathtt{x}})) \\ &< sim(\mathcal{F}_1(\alpha, \mathbf{I}_o), \mathcal{F}_2(\alpha, \mathbf{I}_o)) = \lambda_o < \varepsilon, \end{aligned}$$

Therefore, the solution to Eq.(1) can be used to create the pirated image which is of high quality but of low watermark similarity. In the following, we will omit the superscript ^{min} for simplicity if there is no ambiguity.

D. Discussion on high-dimension puzzle

According to Subsection II-C, the solution to Eq.(1) can defeat the watermarking scheme. Therefore, the critical step in the pollution attack is to solve Eq.(1) so as to obtain the optimal colored noise $\mathtt{p}$. Let n be the dimension of a watermark. If n is very big, (*e.g.*, $n = 57600$ in the experiments of [15]), it may be too expensive to directly carry out the computation in terms of time and space. To handle the high-dimension problem, we can adopt any of the following methods to reduce the dimensions:

(a) Simply fix some elements of $\mathtt{p}$ as 0.
(b) Split the elements of $\mathtt{p}$ into groups, and all the group members have the same values.
(c) Divide the high-dimension problem into several low-dimension sub-problems and solve the sub-problems in parallel or in sequence.

Apparently the value of $\mathtt{p}$ obtained via the above simplifications may no longer be the optimal value. In this case, we have to check the applicability of $\mathtt{p}$against the requirements R1-R3 so as to defeat the detector.

III. Subspace Watermarking Scheme

This section will brief the subspace watermarking schemes. For sake of consisteny, the following sections adopt the notations of the original paper [13]. An integer $n > 1$ is the dimension of the watermark and $\mathcal{X}$ is the feature space of images (for instance, the space corresponding to the entries in the top left corner of the DCT coefficients).

A. Generating feature vector

In order to embed an arbitrary watermark $\mathtt{w}$ into an image $\mathbf{I}_o$, the embedder transforms (*e.g.*, DCT) an original image $\mathbf{I}_o$ to produce the transform-domain coefficients, then re-arranges the 2D coefficients into a 1D column vector ϕ_o with a public operator $\xi(\cdot)$. For example, $\xi(\cdot)$ means to scan left-right corner of DCT matrix row by row.

Next, the embedder decomposes $\mathcal{X}$ into two orthogonal subspaces $\mathcal{W}$ of dimension $2n$ and $\mathcal{V}$ of dimension m, and further splits $\mathcal{W}$ into two orthogonal subspaces $\mathcal{G}$ and $\mathcal{H}$ of dimension n with a secret random operator $\eta(\cdot)$. In other words, the feature space $\mathcal{X}$ is divided into three subspaces $\mathcal{G}$, $\mathcal{H}$ and $\mathcal{V}$, $\mathcal{X} = \mathcal{G} \oplus \mathcal{H} \oplus \mathcal{V}$. Technically, the embedder chooses a $(2n+m)\times(2n+m)$ orth-normal matrix M (*i.e.*, $M^T = M^{-1}$) randomly. Then, he/she produces a $(2n+m) \times m$ matrix V which is a sub-matrix of M to form the basis of $\mathcal{V}$. Similarly, he/she produces $(2n+m) \times n$ matrix[2] G (and H) which is a random sub-matrix of M. Please note matrices G, H, V are the ortho-normal bases of subspaces $\mathcal{G}, \mathcal{H}, \mathcal{V}$ respectively, and they are orthogonal to each other. Hence, the original feature vector $\phi_o \in \mathcal{X}$ can be represented as

$$\phi_o = \psi_o + \sigma_o,$$

where $\psi_o \in \mathcal{W}$ and $\sigma_o \in \mathcal{V}$, and

$$\psi_o = G\mathtt{s} + H\mathtt{t},$$

for some $n \times 1$ vectors $\mathtt{s}$ and $\mathtt{t}$.

B. Embedding watermark into feature subspace

The embedding algorithm is defined by

$$\phi_w = \phi_o + G\mathtt{w}, \tag{2}$$

where $\mathtt{w} \in \mathbb{R}^n$ is an arbitrary watermark vector or watermark for short. Next, the embedder re-arranges 1D vector ϕ_w to a 2D matrix with the public inverse operator $\xi^{-1}(\cdot)$, and performs inverse transform (*e.g.*, IDCT) to produce the watermarked image $\mathbf{I}_w$.

In order to provide the capacity of public detection, the embedder chooses a symmetric matrix A (*i.e.*, $A^T = A$) satisfying

$$A(\mathtt{s} + \mathtt{w}) = \mathtt{s} + \mathtt{w}$$

and an orthogonal matrix B (*i.e.*, $B^T = B^{-1}$) satisfying

$$B\mathtt{t} = \mu(\mathtt{s} + \mathtt{w})$$

with $\mu := \| \mathtt{t} \| / \| \mathtt{s} + \mathtt{w} \|$, and defines a matrix

$$D = AG^T + \mu BH^T.$$

Let $\mathtt{q} = \mathtt{s} + \mathtt{w}$. The public key is $\alpha = (D, \mathtt{q}, \xi)$ which is known to everyone, but the secret G is known to the embedder only. Those who are interested in the selection of A, B, please refer to the original paper [13].

C. Detecting watermark in feature subspace

With the public key α, a detector will extract the feature vector from an unknown image $\mathbf{I}_u$ as the embedder does in Subsection III-B. Let ϕ_u be the extracted feature of the image $\mathbf{I}_u$, then the detector calculates the similarity

$$\lambda_u = sim(\mathtt{q}, D\phi_u) = \frac{(\mathtt{s} + \mathtt{w})^T D\phi_u}{\| \mathtt{s} + \mathtt{w} \| \cdot \| D\phi_u \|}. \tag{3}$$

If $\lambda_u > \varepsilon$ for some predefined threshold value $0 < \varepsilon \ll 1$, the watermark $\mathtt{w}$ is confirmed.

IV. Pollution Attack to Subspace Watermarking

Due to the non-linearity of embedding formulas, the subspace scheme [14] is robust against the attack of white noise addition. For instance, in the experiments of the paper [14], despite a watermarked image $\mathbf{I}_w$ is degraded greatly by adding a white noise power up to 20dB, the detection probability is still very high. Nonetheless, it does not mean that the subspace scheme [14] is secure against all the addition attacks.

Unlike white noise generated randomly, the present colored noise is generated with the public key $\alpha = (D, \mathtt{q}, \xi)$. After the optimal colored noise $\mathtt{p}$ is inserted into the watermarked image $\mathbf{I}_w$ to produce a pirated image $\mathbf{I}_p$, the similarity $\lambda_p < \varepsilon$, *i.e.*, the detector can not confirm the existence of the watermark $\mathtt{w}$ from $\mathbf{I}_p$.

A. Attack method

Given an watermarked image $\mathbf{I}_w$, an attacker aims to generate a pirated image $\mathbf{I}_p$ such that both $\mathbf{I}_p$ and $\mathbf{I}_w$ are perceptually similar, but their watermark similarities are obviously different, *i.e.*, $\lambda_p = sim(\mathtt{q}, D\phi_p) < \varepsilon < sim(\mathtt{q}, D\phi_e) = \lambda_w$, where ϕ_p and ϕ_e are the extracted feature vectors from $\mathbf{I}_p$ and $\mathbf{I}_w$ respectively. To this end, the attacker generates a pirated feature vector $\phi_p = \phi_e + \mathtt{p}$ with the extracted vector ϕ_e and a pollution noise $\mathtt{p}$ based on the method described in Section II. Specifically, in order to guarantee the similarity

$$\begin{aligned}
\lambda_p &= sim(\mathtt{q}, D\phi_p) = \frac{(\mathtt{s} + \mathtt{w})^T D\phi_p}{\| \mathtt{s} + \mathtt{w} \| \cdot \| D\phi_p \|} \\
&= \frac{(\mathtt{s} + \mathtt{w})^T D(\phi_e + \mathtt{p})}{\| \mathtt{s} + \mathtt{w} \| \cdot \| D\phi_p \|} \\
&= \frac{(\mathtt{s} + \mathtt{w})^T D\phi_e + (\mathtt{s} + \mathtt{w})^T D\mathtt{p}}{\| \mathtt{s} + \mathtt{w} \| \cdot \| D\phi_e + D\mathtt{p} \|} < \varepsilon,
\end{aligned}$$

an attacker will utilize Eq.(1) to obtain the pollution signal $\mathtt{p}$ by solving

$$\begin{cases} \mathtt{p} = arg\ \min_{\mathtt{x}} \mid sim(\mathtt{q}, D(\phi_e + \mathtt{x})) \mid \\ PSNR(\mathbf{I}_p - \mathbf{I}_w) \geq \gamma_0 \\ \theta_l \leq \mathtt{p} \leq \theta_h \end{cases} \tag{4}$$

[2]From the items $(2n+m) \times 1$ vector ψ_o, $\mathtt{w} \in \mathbb{R}^n$ and $G\mathtt{w}$ in Eq.(2), we know that G consists of $2n+m$ rows and n columns. *i.e.*, G is a $(2n+m)\times n$ matrix.

Based on the theory of linear algebra, the attacker can find at least one solution of Eq.(4) which can defeat the watermark detector (cf. Subsection IV-B).

In order to obtain the solution of Eq.(4) quickly, we prefer to simplifying the formulas. In order to guarantee invisibility, $\| \mathtt{s} \| \gg \| \mathtt{p} \|$, thus we have

$$(\mathtt{s}+\mathtt{w})^T D\phi_e = (1+\mu^2)(\mathtt{s}+\mathtt{w})^T(\mathtt{s}+\mathtt{w}) \gg (\mathtt{s}+\mathtt{w})^T D\mathtt{p}$$

Therefore,

$$\begin{aligned} \lambda_p &= sim(\mathtt{q}, D\phi_p) = \frac{(\mathtt{s}+\mathtt{w})^T D\phi_e + (\mathtt{s}+\mathtt{w})^T D\mathtt{p}}{\| \mathtt{s}+\mathtt{w} \| \cdot \| D\phi_e + D\mathtt{p} \|} \\ &\approx \frac{(\mathtt{s}+\mathtt{w})^T D\phi_e}{\| \mathtt{s}+\mathtt{w} \| \cdot \| D\phi_e + D\mathtt{p} \|} \end{aligned} \quad (5)$$

Note that Eq.(5) does not hold all the time, but it is enough to provide an approximate solution. As a result, Eq.(4) can be replaced with its simplified form which merely maximizes the denominator $\| D\phi_e + D\mathtt{p} \|$ That is, the attacker can obtain the pollution signal $\mathtt{p}$ by solving

$$\begin{cases} \mathtt{p} = arg \ \max_{\mathtt{x}}(\| D\phi_e + D\mathtt{x} \|) \\ PSNR(\mathbf{I}_p - \mathbf{I}_w) \geq \gamma_0 \\ \theta_l \leq \mathtt{p} \leq \theta_h \end{cases} \quad (6)$$

or a simpler one

$$\begin{cases} \mathtt{p} = arg \ \max_{\mathtt{x}}(\| D\phi_e + D\mathtt{x} \|) \\ PSNR(\mathtt{p}) \leq \gamma_1 \\ \theta_l \leq \mathtt{p} \leq \theta_h \end{cases} \quad (7)$$

where γ_1 is the predefined noise energy threshold in dB (say 20dB). Since Eq.(7) is a good approximation formula of Eq.(4), its solution $\mathtt{p}$ usually suffice to defeating the watermark detector. Therefore, we elaborate the attack with Eq.(7) only.

B. Attack rationale

Following the principle that a watermark detector should not identify the watermark from an original image, the public-key subspace watermarking scheme [14] guarantees

$$\begin{aligned} sim(\mathtt{q}, D(\phi_e - G\mathtt{w})) &\approx sim(\mathtt{s}+\mathtt{w}, D(\phi_w - G\mathtt{w})) \\ &= sim(\mathtt{s}+\mathtt{w}, D\phi_o) < \varepsilon. \end{aligned}$$

For an image $\mathbf{I}_p$ generated with the solution $\mathtt{p}$ to Eq.(4), the similarity is

$$\begin{aligned} \lambda_p &= sim(\mathtt{q}, D(\phi_e + \mathtt{p})) \\ &= \min_{\mathtt{x}} sim(\mathtt{s}+\mathtt{w}, D(\phi_e + \mathtt{x})) \\ &< sim(\mathtt{s}+\mathtt{w}, D(\phi_e - G\mathtt{w})) < \varepsilon. \end{aligned}$$

That is to say, the solution of Eq.(4) can fool the detector.

C. Attack process

Given an watermarked image $\mathbf{I}_w$ and the public detection key $\alpha = (D, \mathtt{q}, \xi)$, the attacker will

- Perform DCT on $\mathbf{I}_w$, and select the coefficients (*i.e.*, top-left corner of DCT) which are used for embedding.
- Re-organize the selected coefficients with algorithm ξ to form feature vector ϕ_e.
- Define a random $\mathtt{x}_0$ according to subsection IV-D.
- Obtain the pollution signal $\mathtt{p}$ by solving Eq.(7) in Subsection IV-A,
- Calculate $\phi_p = \phi_e + \mathtt{p}$, and perform the inverse operator ξ^{-1} to construct a DCT coefficient matrix.
- Perform IDCT on the DCT coefficient matrix so as to obtain a matrix in image domain.
- Round each element of said image matrix to the closest integer value in the designated interval (*e.g.* [0,255] for 8-bit gray image) so as to construct the pirated image $\mathbf{I}_p$.
- If either $\lambda_p > \varepsilon$ or the quality of image $\mathbf{I}_p$ is too low, repeat the above steps by re-selecting the initialization value $\mathtt{x}_0$ so as to find other local extreme vector $\mathtt{p}$.

As a result, the attacker produces a pirated image $\mathbf{I}_p$ which is of high quality, but a detector fails to detect the watermark $\mathtt{w}$ from $\mathbf{I}_p$.

D. Selection of initialization value

Although a solution $\mathtt{p}$ can be derived from Eq.(7) by randomly selecting an initialization value $\mathtt{x}_0$, it may not be satisfactory in terms of the quality of the pirated image $\mathbf{I}_p$ and watermark similarity λ_p because the solution is usually a local extreme rather than global one. Therefore, an attacker tries to select the initialization value which is close to the global extreme point of Eq.(7).

An ideal initialization value is $\mathtt{x}_0 = -G\mathtt{w} \in \mathcal{W}$ because this initialization value immediately produces a pirated image $\mathbf{I}_p = \mathbf{I}_o$ which is of the highest quality and of low watermark similarity. Of course, the attacker may not be so lucky to guess the ideal value. But, he/she may attempt to select an initialization value $\mathtt{x}_0 \in \mathcal{W}$ such that the local extreme is in the subspace $\mathcal{W}$. However, we assume that the attacker does not know the subspace of $\mathcal{W}$ because he/she does not know the watermark length n (Note that n is fixed in [14] but n is a variable in our attack. Hence, our assumption is weaker than that of [14]). In order to select an initialization $\mathtt{x}_0 \in \mathcal{W}$, the attacker calculates

$$\begin{aligned} D^T\mathtt{q} &= (GA^T + \mu HB^T)(\mathtt{s}+\mathtt{w}) \\ &= G(\mathtt{s}+\mathtt{w}) + Ht = \psi_o \in \mathcal{W}. \end{aligned}$$

Although $\psi_o \in \mathcal{W}$ is known to all, we do not initialize $\mathtt{x}_0 = c\psi_o$ for some scalar c due to two reasons: (1) $c \in \mathbb{R}$ is only an one-dimension variable, thus the number of available extreme points of Eq.(7) is much smaller than the total extreme points. It means that the success chance is much small; (2) the elements of ψ_o vary greatly, and hence the solution $\mathtt{p}$ may result in spike noise.

On the other hand, $\phi_e = \phi_o + \mathtt{r}$ where $\mathtt{r}$ is a residue error due to operations such as DCT/IDCT. Generally, $\| r \|$ is very small as shown in Fig.1. Next, the attacker computes

$$\sigma_e = \phi_e - D^T\mathtt{q} = \phi_o + \mathtt{r} - \psi_o = \sigma_o + \mathtt{r} \approx \sigma_o \in \mathcal{V}.$$

Therefore, to make sure that $\mathtt{x}_0$ is close to the subspace $\mathcal{W}$ which is orthogonal to $\mathcal{V}$, it is preferable that $\mathtt{x}_0$ is orthogonal

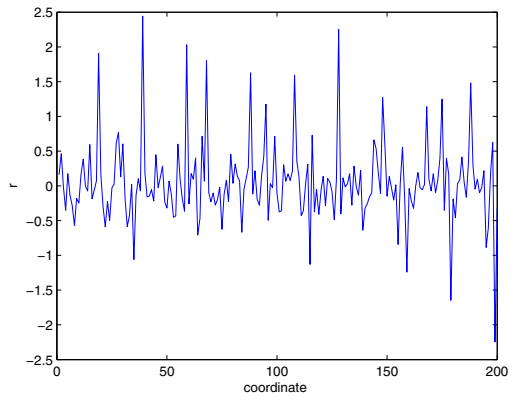

Fig. 1. The empirical residue error r with mean 0.0468 and standard variance 0.5719 due to image processing operations such as DCT/IDCT.

to σ_e. Thus, one simple way is to set

$$\mathtt{x}_0(i) = \begin{cases} 0 & \psi_o(i) \neq 0 \text{ and } \sigma_e(i) \approx 0 \\ a \in_R (\theta_l, \theta_h) & \text{otherwise} \end{cases}$$

where $\mathtt{z}(i)$ denotes the ith element of vector $\mathtt{z}$, and a is a random number.

V. Experiments

In this section, we tested the performance of the proposed pollution attack in terms of detection probability and the quality of pirated images.

A. Watermarking parameters

As the configuration in the original paper [14], the random watermark $\mathtt{w}$ is of length $n = 200$ with energy up to 20dB. We further chose the feature space $\mathcal{X}$ corresponding to a 32×32 sub-matrix in the upper-left corner of the DCT of the original image. Next, we split $\mathcal{X}$ into $\mathcal{W}$, corresponding to the upper-left 20×20 sub-matrix which are the perceptually robust components, and its complementary subspace $\mathcal{V}$ which are more susceptible to standard modifications of the image. Finally, we constructed the ortho-normal matrices G and H, and further constructed matrices A and B with the methods in [14]. Additionally, let $\varepsilon = 0.1$ since $\varepsilon \leq 0.1$ is required for low false-positive probability in [14]. For simplicity, we tested the formula Eq.(7) only.

B. Implementation of pollution attack

According to the attack process outlined in Subsection IV-C, the implementation code includes modules for (1) extracting feature; (2) generating pollution vector; (3) tampering feature vector and (4) re-constructing pirated image. We omit to describe the modules (1) and (4) because they are the same as those in the embedding process.

In the module for generating the pollution vector, let $\gamma_1 = 20$ (dB), and randomly select an initialization value $x_0 \in \mathcal{W}$, then solve Eq.(7) with function $\mathrm{fmincon}(\cdots)$ in the optimization toolbox of Matlab 7.0 as

θ_l=-20;
θ_h=20;
$\mathtt{p}$ = fmincon($\min(- \parallel D\phi_e + D\mathtt{x} \parallel)$,$\mathtt{x}_0$,$\cdots$,
θ_l,θ_h in Eq.(7), *options*);

where *options* can be used to decide the number of iterations for solving Eq.(7). Alternatively, we may merge the constraints of Eq.(6) and Eq.(7) for obtaining a good noise $\mathtt{p}$.

If the optimal solution is not found due to the limitation of *options*, we repeat the process. Technically, if

$$PSNR(\mathbf{I}_p - \mathbf{I}_w) < \gamma_0 = 40\text{dB},$$

we solve Eq.(7) with a new initialization value $\mathtt{x}_0$ such that the pirated image is of good quality.

C. Experiment result

In the experiment, we selected 123 images randomly from the image gallery. It took 817.95 minutes to pollute 123 watermarked images on a desktop Dell T3500 (CPU frequency 2.67GHz). Fig. 2 shows the PSNR of the polluted images, and Fig. 3 is the detected similarity value from the polluted images, the results of the experiment show that our attack can fool the detector with a pirated image of good quality.

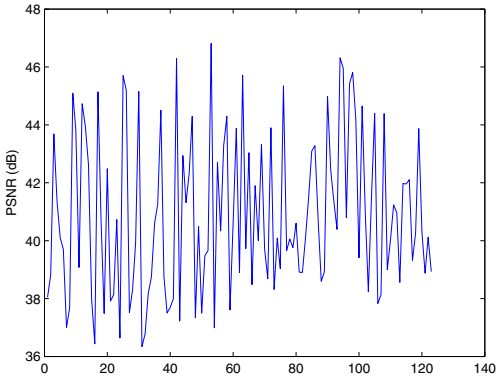

Fig. 2. PSNR values of polluted images. Its mean is 40.9dB, and standard deviation is 2.73dB.

Fig.4 and Fig.5 are examples for visually evaluating the pirated image quality, where the PSNR of the watermarked image $\mathbf{I}_w$ is 41.19dB while the PSNR of the pirated image $\mathbf{I}_p$ is 37.44dB, but the similarity $\lambda_p = 0.040 < 0.1 = \varepsilon$. That is to say, the pollution attack is more powerful than the projection attack investigated in [14].

VI. Conclusion

We presented a pollution attack which optimizes the interference noise so as to decrease detection probability dramatically but only sacrificing the image quality slightly. In comparison with the sensitivity attack, it is much faster and efficient.

In our experiments, we observed that the non-linear embedding/detecting functions may increase the robustness of

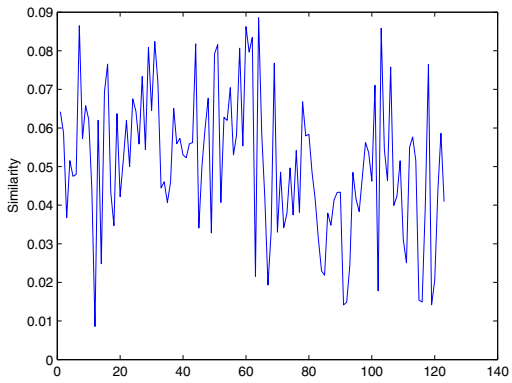

Fig. 3. Similarity values λ_p detected from polluted images. Its mean is 0.051, and standard deviation is 0.019.

Fig. 4. Watermarked image $\mathbf{I}_w$ with PSNR 41.19dB.

a public-key watermarking scheme because the optimization tools are not suitable for high-dimensional problems. Hence, it may be a countermeasure to the pollution attack to design public-key watermarking scheme with non-linear functions, especially discontinuity functions. However, it is not easy to design the non-linear functions because they increase the complexity of detection function too.

ACKNOWLEDGEMENT

This research is supported in part by A*STAR SERC Grant No. 102 101 0027 in Singapore.

REFERENCES

[1] T. Furon, I.Venturini, and P. Duhamel, "Unified approach of asymmetric watermarking schemes, SPIE Security and Watermarking of Multimedia Contents III, vol.4314, pp.269-279, 2001.

[2] S. Craver. "Zero Knowledge Watermark Detection, The 3rd International Information Hiding Workshop, pp. 102-115, 1999.

Fig. 5. Pirated image $\mathbf{I}_p$ with PSNR 37.44dB.

[3] C. H. Tzeng, Z. F. Yang and W. H. Tsai, "Adaptive Data Hiding in Palette Images by Color Ordering and Mapping with Security Protection," *IEEE Transactions on Communications*, 52(5):791-800, 2004.

[4] J. Tian, "Reversible data embedding using difference expansion," *IEEE Trans. on Circuits and Systems for Video Tech.*, 13(8):890-896, 2003.

[5] I. Cox, J. Kilian, F. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for multimedia," *IEEE Trans. Image Processing*, 6(12):1673-1687, 1997.

[6] Jean-paul M. G. Linnartz, Marten Van Dijk, "Analysis of the sensitivity attack against electronic watermarks in images," International Workshop on Information Hiding, Lecture Notes in Computer Science 1525, pp.258-272, 1998.

[7] T. Furon and P. Duhamel, "An asymmetric watermarking method, *IEEE Trans. Signal Processing*, 51(4):981-995, Apr. 2003.

[8] F. Hartung and B. Girod, Fast public-key watermarking of compressed video, IEEE Int. Conf. on Image Processing, pp.528-531, 1997.

[9] J. J. Eggers, J. K. Su, and B. Girod, "Public keywatermarking by eigenvectors of linear transforms," Euro. Signal Processing Conf., 2000.

[10] Hyuk Choi, Kiryung Lee, and Taejeong Kim, Transformed-key Asymmetric Watermarking System, Proc. SPIE Vol. 4314, Security and Watermarking of Multimedia Contents III, pp.280-289, 2001.

[11] J. Tzeng, W.-L. Hwang, and I.-L. Chern, "An asymmetric subspace watermarking method for copyright protection, *IEEE Transactions on Signal Processing*, 53(2):784-792, Feb. 2005.

[12] Yongdong Wu, "On an Asymmetric Subspace Watermarking Method," submission to *IEEE Transaction on Signal Processing*, Feb., 2005.

[13] G. Boato, F. G. B. De Natale, C. Fontanari, "An Improved Asymmetric Watermarking Scheme Suitable for Copy Protection, *IEEE Transactions on Signal Processing*, 54(6):2833-2834, Jul. 2006.

[14] G. Boato, F. G. B. De Natale and C. Fontanari, "A multilevel asymmetric watermarking scheme for digital fingerprinting" *IEEE Transaction on Multimedia*, 10(5):758-766, Aug. 2008.

[15] J. Picard and A. Robert, "Neural networks functions for public key watermarking, 4th Int. Workshop on Information Hiding, Lecture Notes in Computer Science 2137, pp. 142-156, 2001.

[16] G. Qu, "Keyless Public Watermarking for Intellectual Property Authentication," Lecture Notes in Computer Science 2137, pp.96-111, 2001.

[17] Yongdong Wu, Feng Bao, and Changsheng Xu, "On the Security of Two Public Key Watermarking Schemes," 4th IEEE Pacific-Rim Conference on Multimedia, pp.975- 979 vol.2, 2003

[18] Yongdong Wu, Robert H. Deng, "Fooling Public-key Watermarking Detectors with Optimal Color Noise," International Conf. on Multimedia Information Networking and Security, 2009.