Simple Countermeasure to Non-Linear Collusion Attacks Targeted for Spread-Spectrum Fingerprinting Scheme

Minoru KURIBAYASHI\(^{(a)}\), Senior Member

**Summary** Based upon the Kerckhoffs’ principle, illegal users can get to know the embedding and detection algorithms except for a secret key. Then, it is possible to access to a host signal which may be selected from frequency components of a digital content for embedding watermark signal. Especially for a fingerprinting scheme which embeds user’s information as a watermark, the selected components can be easily found by the observation of differently watermarked copies of a same content. In this scenario, it is reported that some non-linear collusion attacks will be able to remove/modify the embedded signal. In this paper, we study the security analysis of our previously proposed spread-spectrum (SS) fingerprinting scheme\([1,2]\) under the Kerckhoffs’ principle, and reveal its drawback when an SS sequence is embedded in a color image. If non-linear collusion attacks are performed only to the components selected for embedding, the traceability is greatly degraded while the pirated copy keeps high quality after the attacks. We also propose a simple countermeasure to enhance the robustness against non-linear collusion attacks as well as possible signal processing attacks for the underlying watermarking method.

**key words:** fingerprinting, spread spectrum, Kerckhoffs’ principle, host signal

1. **Introduction**

Although the robustness against attacks is the major requirement for watermarking schemes, the difficulty in the intentional modification and removal of embedded information is also important factor for the security assessment. According to the Kerckhoffs’ principle\([3]\), it is reasonable to assumed that all parameters and algorithms except for a secret key are known at watermarking had been studied. T. Kalker\([4]\) stated that an adversary. Under such a condition, the security notion in security referred to the inability by unauthorized users to have access to the raw watermarking channel. The Kerckhoffs’ principle comes from cryptographic community but is also widely used in watermarking community\([5–7]\). The security analysis on SS (spread spectrum) watermarking method had been discussed in\([8]\) considering the Kerckhoffs’ principle.

There are two major measurements for watermarking schemes; robustness and security. The robustness measurement is based on unintentional attacks, for instance, signal processing operations on watermarked copy such as lossy compression and filtering. These attacks are assumed to be done by content provider before the legal use and by adversaries who do not know the information about the watermarking system at all. On the other hand, security attacks could be performed by clever adversaries who intend to fool a watermark detector considering the watermarking system including the embedding and detection algorithms. When adversaries know the embedding algorithm except for a secret key, they will be able to focus on the components into which watermark signal may be embedded. Such components are called host signal for convenience. From the sufficient number of observations, they will identify the host signal and will perform attacks intensively.

In this paper\(*\), we study the effects of the non-linear attacks on our previously proposed SS fingerprinting scheme\([1,2]\). Generally, a color image is represented by RGB color components and the color space is sometimes translated into luminance and chrominance signals depending on applications. It is well-known that the chrominance bandwidth is usually reduced in lossy compression like JPEG algorithm, and hence, the luminance is usually used to embed watermark. Meanwhile, frequency components are generally suitable for embedding rather than spatial components. Therefore, there are at least three layers, RGB, luminance, and frequency components, to perform collusion attacks. We evaluate the impact on each layer by the measurement of traceability, and derive the following characteristics. If colluders directly modify the components into which an SS sequence is embedded, they can effectively degrade the SS sequences involved in a pirated copy. In such a case, regretfully, the traceability of conventional fingerprinting schemes is destructively dropped by the non-linear collusion attacks presented in\([10]\). Although the experiments are performed only for the specific SS fingerprinting scheme, the drawback under the Kerckhoffs’ principle covers the similar schemes such that an SS sequence is embedded into the frequency domain of luminance components of a color image.

From the other perspective, our interest is how to prevent colluders from performing the destructive collusion attacks. One solution is to make it difficult to identify the components into which an SS sequence is embedded. In this paper, we propose a simple but effective method for obfuscating the components to be selected for embedding. According to a secret key, we choose a pseudo-random number (PN) sequence like M-sequence\([11]\), and multiply lu-

---

\(^{a}\)The author is with the Graduate School of Natural Science and Technology, Okayama University, Okayama-shi, 700–8530 Japan.

E-mail: kminoru@okayama-u.ac.jp

DOI: 10.1587/transinf.2015MUP0005

---

\(*\)The preliminary version of this paper is appeared in the proceedings of ISITA2014\([9]\).
minence components by the PN sequence before transforming to the frequency domain. Then, selected components for embedding involve the high frequency components of an image, and hence, the robustness against lossy compression is not sufficiently high. To improve the drawback, we apply the DWT (Discrete Wavelet Transform) to separate the high and low frequency components, and embed only into the low frequency components. The performance of the proposed method is intensively evaluated using the specific SS fingerprinting scheme by executing the collusion attacks to some possible domains. Because of its simplicity, the proposed method can be easily applied for the similar SS fingerprinting schemes.

This paper is organized as follows. In Sect. 2, we review the recent security analysis of watermarking schemes, and classify the domains for attack. In Sect. 3, we study the characteristic of fingerprinting scheme and vulnerability against non-linear collusion attacks. The proposed method is described in Sect. 4, and the performance evaluation is shown in Sect. 5. Finally, we conclude this paper in Sect. 6.

2. Watermarking Security

2.1 Kerckhoffs’ Principle

Auguste Kerckhoffs [3] introduced six principles for cryptography. Five of them were fairly straight-forward, and may appear anachronistic today. The second principle in contrast is occasionally controversial, and is absolutely fundamental to modern practice. Now known simply as Kerckhoffs’ principle.

Kerckhoffs’ principle is well understood in the context of symmetric cryptosystems. A symmetric cryptosystem is given by a one-to-one parameterised function $Enc$ using a secret key $k$. Given a plaintext $M$, the ciphertext is given as

$$C = Enc(M, k).$$

The encryption function is illustrated in Fig. 1. The adversary’s objective is either to recover the plaintext $M$ from the ciphertext $C$, or to recover the key $k$ from a set of ciphertexts or ciphertext/plaintext pairs.

A watermarking scheme is different from a cryptosystem. We have an embedding function,

$$y = Emb(x, w, k),$$

where $x$ is a host signal, $y$ is the watermarked copy, and $w$ is watermark. The embedding function is illustrated in Fig. 2. The key $k$ has a similar interpretation as the cryptosystem. An adversary tries to create an unauthorized copy $\hat{y}$ from $y$ and the knowledge of the function $Emb$, or to recover the key $k$. The copy $\hat{y}$ as well as $y$ must be perceptually equivalent to $x$.

In cryptoanalysis, an adversary recognizes a valid plaintext if he succeeds. On the other hand, he may not know whether he fools the detector of watermark or not.

2.2 Security Level of Watermarking Scheme

There are two attack frameworks for security assessment of watermarking scheme [6]. Known message attack (KMA) framework assumes that an adversary have access to watermarked signals and their embedded watermarks. In the watermarked only attack (WOA) framework, only the watermarked signals are available to the adversary.

Under the WOA framework we should consider the statistical distribution of host signal. It is assumed that an adversary knows embedding and detection algorithms except for a secret key, and he can obtain more than one watermarked content of the same [7]. Then, he tries to get the secret key from gathering information. In case of SS watermarking, the secret key can be regarded as a secret carrier like SS sequences. When a watermark is embedded in a digital content according to a secret key, the statistical distribution of the watermarked content may be changed. With a sufficient number of observations, an adversary will be able to estimate the secret key [8], [12]. When an SS watermarking method is used, the watermark signal modulated by the SS sequence is embedded into some frequency components selected according to a secret key. In such a case, the identification of the selected components are easily accomplished by the observations of some differently watermarked copies [13].

It is also possible for an adversary to modify/delete the embedded watermarks by combining some copies of a same content without analyzing the secret key. In some cases, he may be able to reduce the candidates of components into which watermark signals are embedded if the algorithm is known. Thus, he can selectively perform an attack to fool the detector of watermark without degrading the perceptual quality.

2.3 Host Signal

Generally, it is assumed that the components into which the watermark is embedded are selected according to a secret key, and the embedding operation also uses the other key for the security reason. The flowchart of embedding procedure is depicted in Fig. 3, where the feature extraction means the selection of frequency components [14], [15].
The feature extraction essentially depends on the targeted digital content. In raw format of a color image, the color space is typically represented by RGB color space. At the compression, the color space is translated into luminance(Y) components and the others, e.g. YCrCb format in the JPEG algorithm. In conventional works, Y components are used for embedding because of the down-sampling of Cr and Cb components in the JPEG algorithm. Therefore, a host signal is preferred to be selected from the frequency domain of the Y components. Thus, the feature extraction of a color image may be composed of two operations as shown in Fig. 4: one is the separation of color components, and the other is the extraction of frequency components. If the system parameters except for a secret key are known, the first operation is known at the adversary’s side.

Under the WOA framework, however, it is easy for an adversary to identify the host signal even if its elements are selected according to a secret key. Therefore, the security of watermarking schemes in general only depends on the secret key input in the embedding operation. Namely, the secrecy of two secret keys $k_f$ and $k_e$ shown in Fig. 3 are supposed to be 1) $k_f$ is easy to estimate and 2) $k_e$ is difficult to estimate.

2.4 Security Measurement

Our interest is to investigate how to select the host signal $x$ according to a secret key. The host signal $x$ is a vector of sampled components of digital content, which could be selected from spatial domain and frequency domain. Considering the robustness, the host signal $x$ are usually selected from low- and middle-frequency domain according to a secret key.

Basically, there are two approaches to calculate frequency coefficients of Y components as shown in Fig. 5; one calculates by full-domain transform, and the other calculates from partitioned blocks which are not overlapped each other. Adversaries can get to know the approach employed at a watermarking system because such an operation cannot be protected using a secret key. It is reasonable to assume that adversaries focus only to the suspicious frequency components for embedding in order to control the perceptual quality of a pirated copy. Because it is much more efficient to modify those components intentionally if they know the system. It means that the robustness against JPEG compression and other filtering operation is not sufficient measurement in such a case. One simple but efficient attack is to add a white Gaussian noise only to the suspicious components. If small blocks are randomly sampled from Y components according to a secret key as shown in Fig. 6, the positions of the blocks could be kept secret. However, if the number of blocks are small, the spreading effects are restricted, and it results in the decrease of the robustness. On the other hand, if the number of blocks are large, the secrecy of the positions may be dropped, because the number of candidates for such positions is decreased.

3. Fingerprinting

One of the important applications of watermarking technique is the fingerprinting such that a coalition of illegal users, called colluders, will be able to identified from a pirated copy by uniquely embedding fingerprinting information in a same content based on a watermarking technique. After the observations of the difference among their uniquely watermarked copies, colluders will attempt to produce a pirated copy by combining their copies. Such an
attack is called collision attack. The scenario for collision attack is friendly to the WOA framework for watermarking scheme.

Studies on collusion-resistant fingerprinting systems can be categorized into two approaches. One is based on the SS technique [16]–[19], and the other is based on collusion-resistant codes [20]–[22]. The idea of SS-based fingerprinting scheme, proposed by Cox et al. [16], is to assign mutually (quasi-)orthogonal sequences to users as their fingerprints.

By employing statistical analysis, modeling of a variety of attacks has been studied for the SS-based fingerprinting scheme. It is reported in [17] that a number of nonlinear collisions such as interleaving attacks can be well approximated by averaging collusion plus additive noise. Hence, hereafter consider the security of the SS-based scheme which embeds an SS sequence of length ℓ as a fingerprint.

### 3.1 Collusion Attack

Because it is easy for collaborators to find the difference among their copies, they can focus on the domain into which watermark signal is embedded under the WOA framework. As discussed in Sect. 2.4, the access to the host signal is possible if the domain is calculated by the transformation illustrated in Fig. 5. In case of random sampling as shown in Fig. 6, the positions of blocks can be detected by observing some copies, and then the host signal can be identified accordingly.

Suppose that \(c\) malicious users out of \(N\) users collude to produce a pirated copy. They first produce a pirated version of host signal \(\hat{y} = \{\hat{y}_0, \ldots, \hat{y}_{ℓ-1}\}\) by performing a certain strategy of collusion attack, and then make a pirated copy combining with the other components of their copies. It is reasonable to assume that the copies of colluders are distorted by signal processing operations such as lossy compression and filtering. For convenience, we denote by \(y_{ρ,t}\) for \(t\)-th elements of \(ρ\)-th user’s copy. In a linear collusion strategy, the pirated copy is calculated as follows:

\[
\hat{y}_t = \sum_{ρ=1}^{c} a_ρ y_{ρ,t}^*,
\]

(3)

where \(a_ρ\) stands for a weighting parameter satisfying \(\sum a_ρ = 1\). For instance, an averaging attack is represented by \(a_ρ = 1/c\), namely

Average: \(\hat{y}_t^{avg} = \frac{1}{c} \sum_{ρ=1}^{c} y_{ρ,t}^*\)

Some typical examples of nonlinear attacks are given as follows:

- **Minimum**: \(\hat{y}_t^{min} = \min_{ρ} y_{ρ,t}^*\)
- **Maximum**: \(\hat{y}_t^{max} = \max_{ρ} y_{ρ,t}^*\)
- **Median**: \(\hat{y}_t^{med} = \text{median } y_{ρ,t}^*\)

Midpoint (MinMax):

\[
\hat{y}_t^{mid} = (\hat{y}_t^{min} + \hat{y}_t^{max})/2
\]

Roughly speaking, the average attack makes the histogram of detected SS sequence to be normal distribution with zero mean. The Minimum and Maximum attacks give normal distributions with nonzero means (negative and positive means, respectively). In order to improve the performance, Zhao et al. [17] presented a preprocessor that subtracts the mean to be zero. When such a preprocessor is introduced in a detector, a number of nonlinear collusion strategies can be well approximated by an averaging collusion plus additive noise.

On the other hand, it is claimed in [10], [23] that the traceability of conventional fingerprinting schemes can be destructively dropped by well-designed two powerful nonlinear attacks called MMX attack and Uniform attack.

**Definition 1** (moderated minority extreme (MMX) attack). Let \(D_t = \hat{y}_t^{avg} - \hat{y}_t^{min}\). The MMX attack for a given threshold \(θ\) outputs the hybrid signal \(\hat{y}_t^{MMX(θ)}\), where

\[
\hat{y}_t^{MMX(θ)} = \begin{cases} 
\hat{y}_t^{min} & \text{if } D_t ≥ θ \\
\hat{y}_t^{avg} & \text{if } |D_t| < θ \\
\hat{y}_t^{max} & \text{if } D_t ≤ -θ
\end{cases}
\]

(4)

**Definition 2** (Uniform attack). The uniform attack takes \(c\) watermarked signals, and produces a hybrid copy \(\hat{y}_t^{uni}\) where each element \(\hat{y}_t^{uni}\) is drawn independently and uniformly at random on the interval \([\hat{y}_t^{min}, \hat{y}_t^{max}]\).

In other words, the element \(\hat{y}_t^{uni}\) is calculated by

\[
\hat{y}_t^{uni} = \hat{y}_t^{mid} + α_t r_t,
\]

where \(α_t = \hat{y}_t^{max} - \hat{y}_t^{min}\) and \(r_t\) is a random number independently and uniformly selected on the interval \([-γ, γ]\) for a threshold \(γ\).

### 3.2 Numerical Results

Suppose that an SS sequence of length \(ℓ = 8192\) is embedded into \(ℓ\) DCT (Discrete Cosine Transform) components randomly selected according to a secret key, where the DCT components are calculated from the Y components of an image after color translation from RGB components. To evaluate the effectiveness of these attacks at each domain, we embed an SS sequence into a 24-bit color image “lena” with 512×512 pixels by using the method in [1] under the constraint of PSNR 45 [dB], and colluders are attempted to detect by using the detector proposed in [2].

Figure 7 shows the sensitivity against nonlinear collusion attacks performed at three different domains. We use parameters of the MMX and Uniform attacks by setting \(θ = 0.5\) and \(γ = 1.0\) in this experiment. It is observed that the DCT components are seriously damaged both by the MMX and Uniform attacks. It is noted that the similar results are obtained for other images when these attacks performed at the DCT components. In case of Uniform attack, the RGB domain is more sensitive than the Y domain in this
It is because of the effects of additive noise represented by $\alpha_{trt}$ in Eq. (5). The number of total elements targeted for attack is just equal to the number of pixels at the Y domain, while it is three times bigger at the RGB domain. Thus, the noise added to three color components degrades the embedded watermark signal as the results.

The main reason why the DCT domain is extremely sensitive to these non-linear collusion attacks is that a fingerprint signal is embedded into some selected DCT components according to a secret key. If each copy of colluders is distorted by other attacks such as JPEG compression and filtering, the other DCT components that are not selected for embedding are also different with each other. Even in such a case, the MMX and Uniform attacks still work well because the output values $\hat{y}_{MMX}(\theta)$ and $\hat{y}_{uni}(\gamma)$ are adjusted to the amount of difference observed from $y^*\rho_t$. As a consequence, we can say that the secrecy of the domain used for embedding is very important to the countermeasure against these attacks.

Next, we examine the effects of attacks at the DCT domain by changing the parameters $\theta$ and $\gamma$. Figure 8 shows the results. It is observed from the figure that the traceability is dropped with the decrease of $\theta$ and with the increase of $\gamma$. It is possible to further drop the traceability by decreasing $\theta$, but it seems sufficient for the MMX attack to use $\theta = 0.5$. For the Uniform attack, the quality of pirated copy is also dropped with the increase of $\gamma$. Considering the trade-off between the quality of pirated copy and the traceability, we hereafter fix $\theta = 0.5$ and $\gamma = 1.0$.

### 4. Proposed Countermeasure

In order to prevent colluders from analyzing embedding components, we introduce an obfuscation method at the embedding procedure. Our goal is to make it difficult for colluders to derive the embedding components without a secret key even if they collect many pieces of copies.

#### 4.1 Obfuscation Method

According to a secret key, we generate a PN-sequence $p_t \in \{\pm 1\}$ such as M-sequence, Gold-sequence [11], and so on. The length of the sequence is equal to or more than the number of pixels in an image. For simplicity, we suppose to use a pseudo-random number generator (PRNG) to generate the PN sequence. Let $Y_t$ be a $t$-th elements of luminance (Y) components of an image. In the embedding procedure, we first multiple the PN sequence to the Y components.

$$Y'_t = Y_t \cdot p_t$$

Then, we perform the full-domain DCT to the $Y'$, and embed an SS sequence. Such a procedure is illustrated in Fig. 9. For convenience, the detection procedure is also shown in Fig. 10.

Ideally, the obtained frequency components are randomized from the original DCT components by the multiplication of PN sequence, and the energy of the image signal is uniformly distributed in the whole components because
of the characteristic of SS system. Without the exploited PN sequence, it is difficult to exactly obtain the frequency components from the image or its fingerprinted one.

4.2 DWT Domain

The above obfuscation method is simple and effective to prevent colluders from directly obtaining the components for embedding, and hence, it is expected to improve the robustness against the MMX and Uniform attacks. However, the scheme is vulnerable against compression attacks because of the following reason. Due to the multiplication of PN sequence before transforming into frequency domain, the signal embedded as a fingerprint is spread all over the frequency domain after the inverse operation. As lossy compression like JPEG algorithm mainly reduces the energy of high frequency components, the signal spread over such frequency components will be affected seriously by compression attacks. As the results, the performance of the detector is decreased from the conventional method.

In order to improve the robustness against compression attacks, we excludes high frequency components for embedding. The idea is to divide an image into four sub-bands \( LL_1, LH_1, HL_1, HH_1 \) using the DWT. Since the \( LL_1 \) represents a half scaled image, the high frequency components are separated from the original image. By performing the proposed obfuscation method, we embed an SS sequence into the \( LL_1 \) components. Due to the characteristic of DWT, the \( LL_1 \) components can be further divided by four sub-bands and its low sub-bands can be divided iteratively. Therefore, the above embedding procedure can be performed to the lower sub-band components. However, the number of components in the lower sub-band becomes smaller by a factor of 4, which may give an advantage for attackers to reduce the components selecting for non-linear attacks.

4.3 Security Assessment

For the security assessment of the proposed method, the linear complexity of PN sequences is important measurement. The linear complexity of a given sequence is considered as one of the measurements for evaluating the complexity of the function generating a sequence. In a stream cipher, a binary key sequence is generated by a certain PRNG from a secret key, and is XOR-ed with an input message. A large linear complexity of the key sequence is necessary (but far from sufficient) condition for its practical security.

The linear complexity of M-sequence is so small that we can not use it for the key sequence in the stream cipher even if it has good properties of randomness. It is sufficient to employ a secure stream cipher for the generation of good PN sequence according to a secret key. It is possible to prevent colluders from estimating the secret key even if they observe their uniquely watermarked copies. It is still left for an open problem whether the PN sequence generated by a PRNG in a stream cipher retains a good SS property like the characteristic of M-sequence, Gold-sequence and so on. One possible candidate for PRNG may be a Mersenne Twister [24]. Since it is not clear whether the sequence generated by the Mersenne Twister can spread the signal energy uniformly over the whole Y components, we use the M-sequence in the following experiments for the evaluation of proposed obfuscation method.

5. Experiments

For the evaluation of the proposed detection method, we implemented the algorithm, and evaluated the number of detected colluders from a pirated copy.

5.1 Condition

As a host image, we use 8 images “baboon”, “f16”, “fruits”, “lena”, “peppers”, “sailboat”, “splash”, “ TIFFany”, which are shown in Fig. 11. The host images have 24-bit RGB color components with a size of \( 512 \times 512 \) pixels. Under a distortion constraint with a PSNR of 45 [dB], we embed randomly selected fingerprint signal into the host image. The number of users is \( N = 10^6 \), and the total false-positive probability that is the probability of accusing any innocent users by mistake is fixed to \( 10^{-4} \) in this experiment. At the detection, we use the iterative detector with removal operation proposed in [2]. We use a M-sequence as the PN sequence \( p_t \) in this experiment. Although the linear complexity is small, it has
Fig. 11 Test color images with $512 \times 512$ pixels.

(a) baboon  (b) f16  (c) fruits  (d) lena

(e) peppers  (f) sailboat  (g) splash  (h) tiffany

Fig. 12 Three attack types depending on the targeted domain.

(a) attack 1  (b) attack 2  (c) attack 3

The proposed scheme (prop.) introduces the obfuscation method and the DWT to obtain the host signal. In order to check the effects of obfuscation method in the $LL_1$ domain, we also implement three restricted methods; one excludes the DWT (noDWT), one excludes the obfuscation method (noPN), and the other excludes both of them from the proposed method (noPN-DWT). Notice that the “noPN-DWT” method is just equal to the conventional method.

As confirmed at Sect. 3.2, the performance is drastically degraded if colluders can directly modify the domain into which watermark is embedded. Among some candidates for the domains targeted for collusion attack, we perform three attack domains as shown in Fig. 12 considering the observable domains by colluders under the WOA framework.

During a legal purchase, each fingerprinted copy is compressed in order to reduce the distribution cost. Then, colluders decompress their copies and produce a pirated copy by performing the MMX or Uniform attacks with/without the double JPEG compression. Figures 13-16 show the average number of detected colluders. The performance of “noPN-DWT” and “noPN” is drastically dropped with the increase of the number of colluders. It is because the obfuscation method is not exploited in these methods. As the watermark is embedded into some selected DCT coefficients calculated from Y components of a color image in the “noPN-DWT” method, it is obvious that these methods are sensitive to the attack 1. It is noticed that the “noPN-DWT” method is also sensitive to attack 2 though it is slightly less sensitive to attack 1. The reason comes from the following characteristic of frequency transformation. When watermark is embedded into the low- and middle-frequency DCT coefficients of full Y components, the embedded signal is also concentrated on the low- and middle-frequency domain of $LL_1$ components, and hence, the attack 2 can effectively modify/delete the embedded signal. The similar effect is appeared to the “noPN” method. On the other hand, the “noDWT” and “prop.” methods retain higher robustness. Because the “noDWT” method can spread over wider frequency domain than the “prop.” method, the performance is better. However, when the JPEG compression is performed, the “noDWT” method is more sensitive than the “prop.” method. Considering the total performance against attacks, the “prop.” method is the best one among these methods.

We also derive the similar results for other images. Tables 1 and 2 enumerate the number of detected colluders for

| Table 1 | Number of detected colluders for the “prop.” method against the MMX attack plus double JPEG compression when $c = 10$. |
|---------|-------------------------------------------------------------|
| image   | attack 1         | attack 2         | attack 3         |
| baboon  | 3.877            | 2.434            | 2.452            |
| f16     | **8.414**        | 9.062            | 8.818            |
| fruits  | 7.475            | 8.319            | 8.088            |
| lena    | **8.580**        | 9.252            | 9.195            |
| peppers | 7.735            | 8.721            | 8.590            |
| sailboat| 7.684            | 8.100            | 7.864            |
| splash  | **8.777**        | 9.578            | 9.556            |
| tiffany | **3.299**        | 3.682            | 3.598            |

| Table 2 | Number of detected colluders for the “prop.” method against the Uniform attack plus double JPEG compression when $c = 10$. |
|---------|-------------------------------------------------------------|
| image   | attack 1         | attack 2         | attack 3         |
| baboon  | 2.933            | 0.969            | 0.928            |
| f16     | **7.034**        | 6.672            | **6.417**        |
| fruits  | 6.066            | 5.356            | 5.537            |
| lena    | 7.358            | **6.870**        | 7.144            |
| peppers | 6.234            | 5.947            | 6.024            |
| sailboat| 6.221            | 4.984            | 4.709            |
| splash  | **7.192**        | 7.359            | 7.799            |
| tiffany | **2.219**        | 2.140            | 2.025            |

simplicity, the quality factor (QF) of the JPEG algorithm is 75% for the first compression and 50% for the second compression in this experiment.

5.2 Traceability

Using $10^3$ patterns of colluders, we produce pirated copies by performing the MMX and Uniform attacks with/without the double JPEG compression. Figures 13-16 show the average number of detected colluders. The performance of “noPN-DWT” and “noPN” is drastically dropped with the increase of the number of colluders. It is because the obfuscation method is not exploited in these methods. As the watermark is embedded into some selected DCT coefficients calculated from Y components of an color image in the “noPN-DWT” method, it is obvious that these methods are sensitive to the attack 1. It is noticed that the “noPN-DWT” method is also sensitive to attack 2 though it is slightly less sensitive to attack 1. The reason comes from the following characteristic of frequency transformation. When watermark is embedded into the low- and middle-frequency DCT coefficients of full Y components, the embedded signal is also concentrated on the low- and middle-frequency domain of $LL_1$ components, and hence, the attack 2 can effectively modify/delete the embedded signal. The similar effect is appeared to the “noPN” method. On the other hand, the “noDWT” and “prop.” methods retain higher robustness. Because the “noDWT” method can spread over wider frequency domain than the “prop.” method, the performance is better. However, when the JPEG compression is performed, the “noDWT” method is more sensitive than the “prop.” method. Considering the total performance against attacks, the “prop.” method is the best one among these methods.

We also derive the similar results for other images. Tables 1 and 2 enumerate the number of detected colluders for
Fig. 13. Traceability against MMX attack using an image “lena”.

Fig. 14. Traceability against MMX attack plus double JPEG compression using an image “lena”, where QF=75\% for the 1st compression and QF=50\% for the 2nd compression.

Fig. 15. Traceability against Uniform attack using an image “lena”.

Fig. 16. Traceability against Uniform attack plus double JPEG compression using an image “lena”, where QF=75\% for the 1st compression and QF=50\% for the 2nd compression.
Table 3 Comparison of worst case under the MMX attack plus double JPEG compression when \( c = 10 \).

| image  | noPN-DWT | noDWT | noPN | prop. |
|--------|----------|-------|------|-------|
| baboon | 1.178    | 0.000 | 2.434|
| f16    | 2.606    | 0.000 | 8.414|
| fruits | 1.762    | 0.000 | 7.475|
| lena   | 1.949    | 0.000 | 8.580|
| peppers| 0.461    | 0.000 | 7.735|
| sailboat| 0.705   | 0.000 | 7.684|
| splash | 2.784    | 0.000 | 8.777|
| tiffany| 0.236    | 0.000 | 3.299|

Table 4 Comparison of worst case under the Uniform attack plus double JPEG compression when \( c = 10 \).

| image  | noPN-DWT | noDWT | noPN | prop. |
|--------|----------|-------|------|-------|
| baboon | 0.578    | 0.002 | 6.417|       |
| f16    | 0.991    | 0.006 | 5.356|       |
| fruits | 0.637    | 0.002 | 6.870|       |
| lena   | 0.739    | 0.002 | 6.556|       |
| peppers| 0.149    | 0.002 | 5.947|       |
| sailboat| 0.351   | 0.003 | 4.709|       |
| splash | 0.947    | 0.001 | 7.192|       |
| tiffany| 0.109    | 0.008 | 2.025|       |

the “prop.” method when the number of colluder is \( c = 10 \), where the number represented by bold font stands for the worst case among three attacks. It is observed that the traceability depends on the characteristic of a host image. It is because of the sensitivity to the double JPEG compression. For the comparison with other methods, the results of worst case for each image are shown in Table 3 and 4. The results indicate that the obfuscation method suppresses the effects caused by non-linear collusion attacks and that the embedding into \( LL_1 \) domain avoids the effects caused by JPEG compression except for an image “baboon”. For an image “baboon”, the sensitivity to JPEG compression greatly degrades the traceability. Nevertheless, it is confirmed from experiments that the proposed method is still better than the others.

It is worth-mentioning that the robustness against JPEG compression could be improved by selecting lower DWT components though the side-effects should be considered. For instance, the number of components selecting for attacks becomes smaller. It is left for our future work. Furthermore, it is noted that the robustness against averaging attack plus additive noise is similar to the conventional method.

6. Conclusion

In this paper, we reviewed the study of security of watermarking schemes based on the Kerckhoffs’ principle. The robustness against attacks is not sufficient to assess the actual performance of a watermarking system. Under the WOA framework that all system parameters except for a secret key are public, we discussed the security of fingerprinting scheme against possible attack scenarios. It is revealed that illegal users can focus on the vulnerable points of the fingerprinting system based on an SS watermarking scheme and can fool the detector. We also proposed a simple countermeasure to the conventional SS-based fingerprinting scheme, and clarify the improved performance through experiments. As the results, we can say that the most important issue what we should mind is that users can access to the host signal under the WOA framework. In order to prevent them from directly attacking the embedded watermark signal, we should keep a secrecy of embedding function in essence using a secret key.

References

[1] M. Kuribayashi, “Hierarchical spread spectrum fingerprinting scheme based on the CDMA technique,” EURASIP J. Information Security, no.502782, p.16, 2011.

[2] M. Kuribayashi, “Interference removal operation for spread spectrum fingerprinting scheme,” IEEE Trans. Information Forensics and Security, vol.7, no.2, pp.403–417, 2012.

[3] A. Kerckhoffs, “La cryptographie militaire,” Journal des Sciences Militaires, vol.9, pp.5–83, 1883.

[4] T. Kalker, “Considerations on watermarking security,” Proc. MMSP’01, pp.201–206, 2001.

[5] M. Barni, F. Bartolini, and T. Furon, “A general framework for robust watermarking security,” Signal Processing, vol.83, no.10, pp.2069–2084, 2003.

[6] F. Cayre, C. Fontaine, and T. Furon, “Watermarking security: theory and practice,” IEEE Trans. Signal Processing, vol.53, no.10, pp.3976–3987, 2005.

[7] F. Cayre and P. Bas, “Kerckhoffs-based embedding security classes for WOA data hiding,” IEEE Trans. Information Forensics and Security, vol.3, no.1, pp.1–15, 2008.

[8] L. Perez-Freire and F. Perez-Gonzalez, “Spread-spectrum watermarking security,” IEEE Trans. Information Forensics and Security, vol.4, no.1, pp.2–24, 2009.

[9] M. Kuribayashi, “Countermeasure to non-linear collusion attacks on spread spectrum fingerprinting,” Proc. ISTA2014, pp.50–54, 2014.

[10] H.G. Schaathun, “Novel attacks on spread-spectrum fingerprinting,” EURASIP J. Information Security, vol.2008, no.803217, p.15, 2008.

[11] R. Gold, “Maximal recursive sequences with 3-valued recursive correlation functions,” IEEE Trans. Information Theory, vol.14, no.1, pp.154–156, 1968.

[12] L. Perez-Freire and F. Perez-Gonzalez, “Security of lattice-based data hiding against the watermark only attack,” IEEE Trans. Information Forensics and Security, vol.3, no.4, pp.593–610, 2008.

[13] B. Mathon, P. Bas, and F. Cayre, “Practical performance analysis of secure modulations for WOA spread-spectrum based image watermarking,” pp.237–243, 2007.

[14] P. Bas and T. Furon, “A new measure of watermarking security: the effective key length,” IEEE Trans. Information Forensics and Security, vol.8, no.8, pp.1306–1317, 2013.

[15] B. Mathon, F. Cayre, P. Bas, and B. Macq, “Optimal transport for secure spread-spectrum watermarking for still images,” IEEE Trans. Image Processing, vol.23, no.4, pp.1694–1705, 2014.

[16] I.J. Cox, J. Kilian, F.T. Leighton, and T. Shamson, “Secure spread spectrum watermarking for multimedia,” IEEE Trans. Image Processing, vol.6, no.12, pp.1673–1687, 1997.

[17] H.V. Zhao, M. Wu, Z.J. Wang, and K.J.R. Liu, “Forensic analysis of nonlinear collusion attacks for multimedia fingerprinting,” IEEE Trans. Image Processing, vol.14, no.5, pp.646–661, 2005.

[18] Z.J. Wang, M. Wu, W. Trappe, and K.J.R. Liu, “Group-oriented fingerprinting for multimedia forensics,” EURASIP J. Adv. Signal Processing, no.14, pp.2142–2162, 2004.

[19] Z.J. Wang, M. Wu, H.V. Zhao, W. Trappe, and K.J.R. Liu, “Anti-collusion forensics of multimedia fingerprinting using orthogonal modulation,” IEEE Trans. Image Processing, vol.14, no.6, pp.804–821, 2005.
[20] D. Boneh and J. Shaw, “Collusion-secure fingerprinting for digital data,” IEEE Trans. Information Theory, vol.44, no.5, pp.1897–1905, 1998.

[21] W. Trappe, M. Wu, Z.J. Wang, and K.J.R. Liu, “Anti-collusion fingerprinting for multimedia,” IEEE Trans. Signal Processing, vol.51, no.4, pp.1069–1087, 2003.

[22] G. Tardos, “Optimal probabilistic fingerprint codes,” J. ACM, vol.55, no.2, pp.1–24, 2008.

[23] H.G. Schaathun, “Attacks on Kuribayashi’s fingerprinting scheme,” IEEE Trans. Information Forensics and Security, vol.9, no.4, pp.607–609, 2014.

[24] M. Matsumoto and T. Nishimura, “Mersenne twister: A 623-dimensionally equidistributed uniform pseudorandom number generator,” ACM Trans. Modeling and Computer Simulations, vol.8, no.1, pp.3–30, 1998.

Minoru Kuribayashi received the B.E., M.E., and D.E degrees from Kobe University, Kobe, Japan, in 1999, 2001, and 2004 respectively. From 2002 to 2007, he was a Research Associate in the Department of Electrical and Electronics Engineering, Kobe University. In 2007, he was appointed as an Assistant Professor at Division of Electrical and Electronics Engineering, Kobe University. Since 2015, he has been an Associate Professor in the Graduate School of Natural Science and Technology, Okayama University. His research interests are in digital watermark, fingerprinting, information security, cryptography, and coding theory. He is a senior member of IEEE, and a member of EURASIP. He received the Young Professionals Award from IEEE Kansai Section in 2014.