A Provable Watermark-based Copyright Protection Scheme

Pei-Yih Ting, Shao-Da Huang, Tzong-Sun Wu, and Han-Yu Lin
Department of Computer Science and Engineering
National Taiwan Ocean University,
Keelung, Taiwan 20242, R.O.C.
Email:pyting@mail.ntou.edu.tw, ml80318@gmail.com, y456@ntou.edu.tw, lin.hanyu@msa.hinet.net

***Abstract*—Watermark-based copyright protection techniques have been investigated for more than two decades in the signal processing and the digital rights management communities. Most efforts have been devoted on hiding the watermark and increasing the robustness of the embedded watermark under common signal processing operations and geometric transformations. In this paper, we build our scheme based on these previous well developed signal processing techniques but focus on how to employ unpredictable signature-seeded pseudo random bit sequence to make the false negative watermark detection rate computationally negligible. The ultimate goal is to resolve the ownership dispute of an exhibited digital media under adversarial watermark removal attacks.**

Keywords: watermark, copyright protection, digital signature, pseudo random bit sequence

I. Introduction

Perceptually invisible watermark-based fingerprinting techniques have been widely investigated [5], [13], [4], [14], [7], [12], [15] for copyright protection or authentication of digital contents like images or audio/video streams. As a common steganography mechanism, watermarks are designed to blend into the cover image such that they are unobtrusive (perceptually invisible to everyone) and can serve as footmarks for identifying the ownership. Watermarks should also be robust (difficult to remove) such that the embedded image sticks to its cover image when the stego image undergoes common signal processing operations and geometric transformations in the course of transmission. Most watermarking schemes are based on signal processing techniques that focus on exploiting the spatial and frequency domain properties of the watermark images and the cover images, the perception models of the human visual system, and the source coding algorithms in order to achieve the above two goals.

As long as it is not easy to tell a cover image (original image) apart from a stego image (watermarked image), it seems to be an overkill from an engineering point of view to prove that the embedded watermark cannot be detected, removed, or extracted without degrading significantly the quality of the stego image, or to prove that no valid stego image can be forged when an adversary is given the knowledge of the watermark generation and embedding procedures. Now that these schemes are not designed to withstand adversarial attacks, there are always possibilities that a watermark is detected, removed, or even forged. Hence, the digital watermarking techniques are usually treated as a complementary tool in helping the owner or the law enforcing officers identify the ownership of a disputed copy. The false positive possibilities and the lack of proof for the intent of redistribution prevent the detection of watermark from being an effective direct evidence in an intellectual property litigation.

In this paper, we propose a provable, cryptographic watermarking scheme that operates in the 'exhibition' model to provide 'proof of ownership' mechanism such that when a piece of exhibited digital content is duplicated, modified, or reproduced without proper authorization, the copyright owner can provide sufficient direct evidences in proving his/her ownership when a lawsuit is filed for the perceived infringement of intellectual property ownership.

There are several major distinctions between the proposed scheme and previous watermarking schemes: First, the watermarked digital contents are used in the public by the owner while many previous schemes [4], [3], [11] are designed for the 'tracing' model, in which the owner sells his watermarked contents to several buyers and demands only private usages. A scheme secure designed for the 'exhibition' model can also be used in the 'tracing' model such that a buyer can embed his own watermark to the purchased digital contents and use the watermarked content in the public. Second, previous schemes take a pattern matching view in determining whether a piece of digital content contains plaintiff's watermark, namely, these schemes answer questions like "whether the detected watermark looks more like A's or B's?" or "whether it is more likely that there exists a specified watermark or not?" through the comparison of a customized similarity measure. This sort of relative measures lead to ambiguous decisions when there exist candidates with competitive scores. On the contrary, a scheme that gives decisions with confidence level negligibly close to 100% is very much preferred in resolving legal disputes. Third, we start with a common cryptographic setup, where the existence of watermark in the stego image is announced, the watermark generation and embedding algorithms are public, and a watermark embedding oracle is provided for the adversary, and try to establish a provable watermarking scheme that can be used in a courtroom to provide reliable evidence just like a digital signature scheme provides according to the "electronic signature acts".

In a conventional watermarking scheme, the watermark to be embedded might be an image showing the textual identity string, a logo image of the company, or a registered trademark. These watermarks do not bind existentially to its owner and can be obtained by anyone for embedding in arbitrary cover media,

e.g. images with defamation or illegal contents, to frame the owner. Also, such watermarks are independent of their cover images. If the embedded watermark in one stego image is extracted and put into a second cover image, it still provides the same functionalities as it did to the first image. These two reasons explain further why a watermarking scheme is only supplementary in the copyright protection scenario.

Because the above two problems are also the characteristic problems of a cryptographic digital signature scheme in authenticating valuable documents, it is quite intuitive to apply the framework by signing the image, treating the signature as the watermark, and embedding it into the cover image [5, Chap.10.2]. In this way, the unforgeability of the digital signature scheme assures that only the holder of the private key can generate the watermark. The exclusive dependency of the watermark on the cover image protects the watermark itself from being used on other images even if the watermark embedding algorithm is not strong enough to hold the watermark and the cover image together. However, there is an intrinsic disparity between the authentication problem and the intellectual property protection problem such that the above intuitive application of cryptographic signature can be easily subverted. When a digital signature σ is used to authenticate a piece of data x, both x and σ are left in the public and any single change of the bits of x or σ is considered violation of the data integrity and leads to the failure of signature verification. That is, a cryptographic signature is fragile. In the above application, when the stego image is transmitted and undergoes common lossy compressions, the inevitable changes of bits fail the signature verification. As a result, semi-fragile watermark schemes [10], [6], [7], [12] resorting to non-cryptographic, content-based, robust signatures or hashes were proposed instead.

It is the goal of this paper to solve this dilemma by transforming the unforgeable cryptographic signature to an unpredictable pseudo random bit sequence and using the sequence as both the embedding keys and the embedded watermark such that the watermark is extremely robust not only against common signal processing operations and geometric transformations but also against adversarial removal attempts. A cryptographic pseudo random sequence is proven computationally unpredictable under the assumption of the existence of one-way functions. If such a pseudo-random generator is seeded with a uniformly random λ-bit sequence, where λ is the security parameter, the resulting sequence is unpredictable in the sense that, for any polynomial time predicting adversary $\mathcal{A}$, the probability that the single bit output of $\mathcal{A}$ after reading i bits matches the $(i+1)$-th bit is negligibly close to $1/2$. With sufficiently long pseudo random bit sequence as the watermark and a fixed watermark embedding algorithm, we show that the probability for a poly-time adversary to output an image that contains partially (e.g. $\delta = 60$ percent) matched portion of the pseudo random bit sequence without seeing any data related to the embedded watermark is lower than a value in the order 2^{-100}. Hence, in the case that the watermark is hard to remove cleanly from a disputed image, it can be concluded with confidence level negligibly close to 100 percent that if the watermark extracted from a disputed image matches the signature-seeded pseudo random sequence over δ percent and the disputed image has a high peak signal to noise ratio (PSNR) with respect to the cover image, this disputed image must be an unauthorized derivation of the cover image.

Note that although the watermark bits are embedded in pseudo random positions, there is no way to guarantee that no more than, say, 100*(1-δ) percent of the watermark can be removed. Indeed, if the watermark is added to a totally black cover image, it can be completely removed. Thus, the percentage of watermark bits removed depends both on the contents of the cover image and the intelligence used in the removal algorithm. Our scheme nevertheless guarantees that any adversary cannot verify whether sufficient percentage of watermark is removed or not. When the adversary uses the processed watermarked image publicly, he puts himself in a dangerous situation that there might be sufficient amount of evidences left behind to prove this illegal usage.

There were also some fingerprinting researches with cryptographic setups: Boneh and Shaw's collusion resistant fingerprinting [3] is symmetric, in which both the merchant and the buyers know the stego images. Pfitzmann and Schunter's asymmetric fingerprinting [11] combines a generic secure two-party protocol and unforgeable signature scheme to hide the stego image from the merchant. They also implemented a relatively efficient scheme with homomorphic commitments and zero knowledge proofs. However, both the above schemes are designed to work in the 'tracing' model and achieve provable security at the cost of large amount of computation and communication resources.

Section 2 reviews necessary backgrounds in both cryptography and signal processing. Section 3 describes the watermark embedding / detection algorithms and the security of the scheme. Section 4 presents the experimental results and the discussions. The last section is the conclusion.

II. Backgrounds

Some signal processing and cryptography backgrounds are introduced in this section.

Wavelet transform

Wavelet transform is a spatial-frequency decomposition with the bases of various spatial and frequency localities. It has many applications in signal compression, detection, and communications. The following is a one dimensional m-level discrete Haar transform: A signal vector $\mathbf{x} = (x_1, x_2, \ldots, x_{2^n})$ is first split into two parts: the *running average* $\mathbf{c}^{m-1} = (c_1^{m-1}, c_2^{m-1}, \ldots, c_{2^{n-1}}^{m-1})$ is the low frequency part and the *running difference* $\mathbf{d}^{m-1} = (d_1^{m-1}, d_2^{m-1}, \ldots, d_{2^{n-1}}^{m-1})$ is the high frequency part, where $c_j^{m-1} = (x_{2j-1} + x_{2j})/\sqrt{2}$ and $d_j^{m-1} = (x_{2j-1} - x_{2j})/\sqrt{2}$. Recursively decompose $\mathbf{c}^{m-i}$ into low frequency part $\mathbf{c}^{m-i-1}$ and high frequency part $\mathbf{d}^{m-i-1}$, where the size of each vector is 2^{n-i-1}, until $\mathbf{c}^0$ and $\mathbf{d}^0$ are obtained. The size of $\mathbf{c}^0$ or $\mathbf{d}^0$ is 2^{n-m}. Thus, the signal vector $\mathbf{x}$ is decomposed as $(\mathbf{c}^0 \| \mathbf{d}^0 \| \mathbf{d}^1 \| \ldots \| \mathbf{d}^{m-1})$. Define the level-$i$ *scaling functions* $\{\mathbf{v}_j^{m-i}\}_{j=1,\ldots,2^{n-i}}$ and *wavelets* $\{\mathbf{w}_j^{m-i}\}_{j=1,\ldots,2^{n-i}}$, where the size of each vector is 2^n, as follows:

$$\mathbf{v}_j^{m-i} = (\ldots, v_{j,k}^{m-i}, \ldots) \text{ and}$$

$$\mathbf{w}_j^{m-i} = (\ldots, w_{j,k}^{m-i}, \ldots), k = 1, \ldots, 2^n$$

where $v_{j,k}^{m-i} = \begin{cases} 1/\sqrt{2}, & \text{if } 2^i j - 1 \le k < 2^i j + 2^i - 1 \\ 0, & \text{otherwise} \end{cases}$

and

$$w_{j,k}^{m-i} = \begin{cases} 1/\sqrt{2}, & \text{if } 2^i j - 1 \le k < 2^i j + 2^{i-1} - 1 \\ -1/\sqrt{2}, & \text{if } 2^i j + 2^{i-1} - 1 \le k < 2^i j + 2^i - 1 \\ 0, & \text{otherwise} \end{cases}.$$

The signal vector $\mathbf{x}$ can be reconstructed as

$$\sum_{j=1,\ldots,2^{n-m}} c_j^0 \mathbf{v}_j^0 + \sum_{i=1,\ldots,m} \sum_{j=1,\ldots,2^{n-i}} d_j^{m-i} \mathbf{w}_j^{m-i}.$$

Digital signature scheme and its unforgeability:

A digital signature scheme consists of three algorithms:

`KeyGen`(1^λ): The key generation algorithm inputs the security parameter 1^λ and outputs a key pair (PK,SK).

`Sign`(SK, m): The signing algorithm inputs the secret key SK, the message m, and outputs the corresponding signature σ.

`Verify`(PK, m, σ): The verification algorithm inputs the public key PK, the message m, the signature σ, and outputs 1 if the signature is valid; 0 otherwise.

The security of a signature scheme is captured by the existential unforgeability (EUF) under the adaptive chosen message attack (CMA) [8]. In the following game, a challenger $\mathcal{C}$ interacts with an adversary $\mathcal{A}$ and the message space is denoted as $\mathcal{M}$.

Setup phase: $\mathcal{C}$ runs `KeyGen`(1^λ) to obtain a pair of keys (PK,SK) and sends PK to $\mathcal{A}$.

Query phase: $\mathcal{A}$ queries adaptively the signing oracle q_s times of arbitrary message $m^{(j)} \in \mathcal{M}$. $\mathcal{C}$ simulates the signing oracle and returns the corresponding signature $\sigma^{(j)}$ to $\mathcal{A}$.

Output phase: $\mathcal{A}$ outputs a message-signature pair (m^*,σ^*). If $m^* \notin \{m^{(j)}\}_{j=1,\ldots,q_s}$ and `Verify`(PK, m^*,σ^*)=1, then $\mathcal{A}$ wins the game with the advantage

$$Adv_{\mathcal{A}}^{\texttt{EUF-CMA}}(1^\lambda) = \Pr\left[\texttt{Verify}\left(PK, m^*, \sigma^*\right) = 1 \text{ and } m^* \notin \{m^{(j)}\}_{j=1,\ldots,q_s}\right].$$

A signature scheme (`KeyGen`, `Sign`, `Verify`) is `EUF-CMA` secure if the advantage $Adv_{\mathcal{A}}^{\texttt{EUF-CMA}}(1^\lambda)$ of an arbitrary probabilistic polynomial time adversary $\mathcal{A}$ in the above game is negligible.

Pseudo-random generator and its unpredictability:

A cryptographic pseudo-random generator (PRG) $\mathbf{G}(\mathbf{s})$ is a poly-time deterministic algorithm, which inputs a uniformly distributed λ-bit seed $\mathbf{s}$, outputs an $\ell(\lambda)$-bit sequence that is computationally indistinguishable from a uniformly distributed $\ell(\lambda)$-bit random sequence, where λ is the security parameter and the polynomial $\ell(\lambda) > \lambda$. Formally, for every poly-time probabilistic binary-output distinguishing algorithm $\mathcal{D}$, for every positive polynomial $p(\cdot)$, and every sufficiently large integer λ,

$$|\ \Pr[\mathcal{D}(\mathbf{G}(U_\lambda)) = 1] - \Pr[\mathcal{D}(U_{\ell(\lambda)}) = 1]\ | < 1/p(\lambda),$$

where U_λ and $U_{\ell(\lambda)}$ denote uniform random ensembles of λ-bit and $\ell(\lambda)$-bit sequences, respectively. In addition, it can be shown that the output of a PRG is unpredictable, i.e., for every probabilistic poly-time predicting algorithm $\mathcal{A}$, for every positive polynomial $p(\cdot)$, and every sufficiently large integer λ,

$$\Pr[\mathcal{A}(\mathbf{G}(U_\lambda)) = \text{next}_{\mathcal{A}}(\mathbf{G}(U_\lambda))] < 0.5 + 1/p(\lambda),$$

where $\text{next}_{\mathcal{A}}(\cdot)$ is a special function for definitional purpose: when the Turing machine $\mathcal{A}$ processes the first k bits, of the input tape, $\text{next}_{\mathcal{A}}(\cdot)$ outputs the $(k+1)$-th bit of the input tape; when $\mathcal{A}$ reads in all λ bits, $\text{next}_{\mathcal{A}}(\cdot)$ outputs a uniform random bit [9, Chap.3.3.5].

III. Construction & Security Analysis

The proposed provable watermark-based copyright protection scheme consists of three algorithms: (`WSetup`, `WEmbed`, `WVerify`) based on one-way functions from the integer factoring problem. `WSetup`(1^λ) is a probabilistic algorithm that inputs a security parameter 1^λ and outputs a public parameter PK and an embedding key EK. `WEmbed`(PK, EK, I) is the watermark embedding algorithm that uses PK and EK to generate a unique watermark, embeds it into the cover image I, and outputs the watermarked image I_w together with the extraction key $XK_I = (I, \sigma_I)$, where σ_I is the digital signature produced by the embedding key EK. `WVerify`(PK, XK_I, I_a) is the watermark verification algorithm that uses the public key PK and the extraction key XK_I to determine if an arbitrary image I_a contains the watermark corresponding to I.

These three algorithms are described in detail as follows:

- `WSetup`(1^λ): This algorithm first runs `KeyGen`(1^λ) to generate the parameters for an RSA signature scheme: It chooses randomly two $\lambda/2$-bit primes p_1, q_1, calculates the modulus $N_1 = p_1 \cdot q_1$, calculates the Euler totient function $\phi(N_1) = (p_1 - 1)(q_1 - 1)$, chooses the verification exponent e coprime to $\phi(N_1)$, and calculates the signing exponent $d \equiv e^{-1} \pmod{\phi(N_1)}$. The public verification key is (N_1, e) and the private signing key is d. The setup algorithm then chooses parameters for a BBS [1] pseudo random generator (PRG): It chooses a Blum integer $N_2 = p_2 \cdot q_2$ with random primes p_2, q_2 of $\lambda/2$ each and satisfying $p_2 \equiv q_2 \equiv 3 \pmod 4$. The Rabin function f_{N_2}: $QR_{N_2} \to QR_{N_2}$ is defined as $f_{N_2}(x) = x^2 \pmod{N_2}$, where QR_{N_2} is the set of quadratic residues in $\mathbf{Z}_{N_2}^*$. The BBS PRG $G_{f_{N_2}} : \{0,1\}^\lambda \to \{0,1\}^{k\lambda}$ is defined as $G_{f_{N_2}}(s) = LSB(f_{N_2}(s)) \parallel LSB(f_{N_2}{}^2(s)) \parallel \ldots \parallel LSB(f_{N_2}{}^{k\lambda-1}(s))$, where s is the λ-bit seed for the PRG and $f_{N_2}{}^2(s)$ denotes the composition of f_{N_2}, i.e. $f_{N_2}(f_{N_2}(s))$. Finally, it chooses a collision-resistant hash function $H(\cdot)$.
  The public parameter PK consists of N_1, e, N_2, and $H(\cdot)$. The embedding key EK is d.
- `WEmbed`(PK, EK, I): This algorithm first computes the digital signature $\sigma_I = H(I)^d \pmod{N_1}$ of the cover image I. Then, it generates $k\lambda$-bit pseudo random sequence $\mathrm{w}_I[1, k\lambda] = G_f(\sigma_I)$, of which the first λ bits $\mathrm{w}_I[1, \lambda]$ are used later as the watermark

and the remaining bits $\mathrm{w}_I[\lambda+1, k\lambda]$ are used as the embedding keys, $\mathrm{w}_I[\lambda+1, k\lambda]$ are partitioned as λ groups, each specifying one out of $\ell = 2^{k-1}$ pixels for embedding a single bit of the watermark. Next, it obtains the one-level discrete Haar wavelet transform of the cover image I as $DWT(I) = (LL, LH, HL, HH)$ and embeds the watermark $\mathrm{w}_I[1, \lambda]$ into LL as follows: For $i = 1$ to λ, it uses the $(k-1)$-bit subsequence $\mathrm{w}_I[\lambda+i]\ \mathrm{w}_I[2\lambda+i] \ldots \mathrm{w}_I[(k-1)\lambda+i]$ to specify one out of ℓ possible pixels. Then, it replaces the d-th bit of that pixel with the i-th watermark bit $\mathrm{w}_I[i]$. If I is a 256-level grey-scale image or a 256-level luminance color image, putting the watermark at the d-bit is equivalent to a noise of amplitude $\{2^{d-1}, 0, -2^{d-1}\}$, where $1 \leq d \leq 8$. The parameter d is a tradeoff to be determined from our experiments such that the watermark is on one hand embedded in perceptually significant part of the cover image and can survive under common image processing operations or geometric transformations; on the other hand is still perceptually invisible. Because the watermark is hidden with a one out of ℓ strategy, in order to remove every watermark bit in a brute-force way, an adversary would need to clear/set all other $\ell - 1$ genuine pixels at the d-th bit and would cause significant quality degradation of the stego image. Finally, the embedding algorithm obtains the inverse one-level discrete Haar wavelet transform $IDWT(LL', LH, HL, HH)$ to get the spatial domain stego image I_w. I_w is the ultimate product that can be publicly used. The original cover image I and the signature σ_I are kept secret as the extraction key XK_I for resolving future copyright disputes.

- $\texttt{WVerify}(PK, I, \sigma_I, I_a)$: This algorithm generates the pseudo random sequence $\mathrm{w}_I[1, k\lambda] = G_f(\sigma_I)$. Next, it runs $DWT(I_a)$ to decompose I_a as the tuple (LL, LH, HL, HH), and extracts the bits $\mathrm{w}_{I_a}[1, \lambda]$ from LL according to original hiding locations specified by $\mathrm{w}_I[\lambda+1, k\lambda]$. Then the algorithm compares $\mathrm{w}_I[1, \lambda]$ with $\mathrm{w}_{I_a}[1, \lambda]$ bit by bit, and calculates the normalized Hamming distance as $1 - \frac{1}{\lambda}\sum_{i=1}^{\lambda}(\mathrm{w}_I[i] \bigoplus \mathrm{w}_{I_a}[i])$, which is also the common normalized cross correlation $NC(\mathrm{w}_I[1, \lambda], \mathrm{w}_{I_a}[1, \lambda])$ when each bit is represented by values $\{1, -1\}$. This is the indicator of the ratio of bits which are identical in both sequences. If $NC(\mathrm{w}_I[1, \lambda], \mathrm{w}_{I_a}[1, \lambda]) \geq 0.5 + 2/\ell$, the algorithm outputs 1; otherwise it outputs 0.

Ownership resolution protocol:

Consider the following scenario of an ownership dispute: An owner $\mathcal{O}$ creates an original image I and holds I secretly. $\mathcal{O}$ calculates the watermarked image $I_w = \texttt{WEmbed}(PK^{(\mathcal{O})}, EK^{(\mathcal{O})}, I)$ and uses I_w in the public. If the owner $\mathcal{O}$ finds a disputed image I_d, which is both subjectively and objectively close to the original image I, in the merchandise of a plagiarist $\mathcal{P}$. Because $\mathcal{O}$ has never authorized anyone to use the image I_w, he files an ownership infringement lawsuit against $\mathcal{P}$ on the unauthorized reproduction, public usage, and redistribution of I_w. In the following, $(PK_{\mathcal{O}}, EK_{\mathcal{O}})$ and $(PK_{\mathcal{P}}, EK_{\mathcal{P}})$ denote the public and private key pairs of $\mathcal{O}$ and $\mathcal{P}$ respectively. The proposed procedure to resolve the ownership dispute on I_d in the court is as follows:

1) The plaintiff $\mathcal{O}$ sends the image I with his verifiable signature ${\sigma_I}^{(\mathcal{O})}$ to the court-designated trusted third party $\mathcal{T}$. $\mathcal{T}$ first verifies $\sigma_I^{(O)}$ with $\mathcal{O}$'s public key $PK^{(\mathcal{O})}$ by checking if $\texttt{Verify}(PK^{(\mathcal{O})}, I, \sigma_I^{(\mathcal{O})}) = 1$. $\mathcal{T}$ presents the verification results in the court.
2) $\mathcal{T}$ checks if $\texttt{WVerify}(PK_{\mathcal{O}}, I, {\sigma_I}^{(\mathcal{O})}, I_d) = 1$ to see if the disputed image I_d contains the watermark which only $\mathcal{O}$ can make exclusively with image I and secret key $EK_{\mathcal{O}}$. $\mathcal{T}$ then checks if the objective measure $PSNR(I, I_d)$ is greater than 30. The case is dismissed if any one of these two conditions fails.
3) The defendant $\mathcal{P}$ may enter a plea of not guilty by presenting another image I' with its signature ${\sigma_{I'}}^{(\mathcal{P})}$ to $\mathcal{T}$ such that $\texttt{Verify}(PK^{(\mathcal{P})}, I', \sigma_{I'}^{(\mathcal{P})}) = 1$ and $\texttt{WVerify}(PK^{(\mathcal{P})}, I', {\sigma_{I'}}^{(\mathcal{P})}, I_d) = 1$. If any one of the above two tests fails, $\mathcal{P}$ would be declared guilty.
4) $\mathcal{T}$ makes sure that $\texttt{WVerify}(PK^{(\mathcal{O})}, I, \sigma_I^{(\mathcal{O})}, I') = 1$ and $\texttt{WVerify}(PK^{(\mathcal{P})}, I', \sigma_{I'}^{(\mathcal{P})}, I) = 0$ to see if the image I' contains the watermark which only $\mathcal{O}$ can make exclusively with image I and secret key $EK_{\mathcal{O}}$ and the image I does not contain the watermark of $\mathcal{P}$. If this is the case, $\mathcal{P}$ would be declared guilty. Otherwise, the case is dismissed. The following theorem guarantees that the probability of the event that first check being 0 and the second check being 1 is computationally negligible.

If the proposed watermark embedding algorithm is sufficiently robust such that the embedded watermark is hard to remove cleanly, we prove the following theorem:

Theorem: If $\texttt{WVerify}(PK_{\mathcal{O}}, I, \sigma_I^{(\mathcal{O})}, I') = 1$, the probability that I' is not derived from I_w is computationally negligible.

The proof of this theorem follows the unpredictability of the pseudo random sequence. Basically we assume that the probability I' not derived from I_w is non-negligible and try to construct a probabilistic polynomial time adversary that can predict the bits of the pseudo random sequence. The details will be presented in the full version of this paper.

In the above ownership infringement scenario, if an adversary $\mathcal{P}$ wants to use $\mathcal{O}$'s image without authorization, he might try every possible operation to remove the watermark embedded in I_w to obtain a clean image I'. He then embeds his own watermark to I' and obtains I_d. If he removes successfully the watermarks in I_w, i.e. I' is free of $\mathcal{O}$'s watermark, I_d in step 2 would not contain the watermark of $\mathcal{O}$. However, if $\mathcal{P}$ does not remove completely $\mathcal{O}$'s watermark in I_w, $\mathcal{T}$ should find that $\mathcal{O}$'s watermark still in I'. By this result, the judge should declare $\mathcal{P}$ guilty since I' and I_d must be reproduced from some image that contains $\mathcal{O}$'s watermark, namely I_w. Put it in another way, even if the non-removability of the watermark cannot be proved, we force the plagiarist into a very risky situation that he has to try removing the watermark blindly without any verification measure. Because the unpre-

dictability of the watermark bit sequence, we establish a strong removal barrier that the remaining correct watermark bits must be less than $0.5 + 2/\ell$ to avoid the detection by algorithm `WVerify`. If $\mathcal{P}$ cannot cleanly remove the watermark in I_w, he would face the guilty verdict for sure. Although it is not the focus of this paper to improve the non-removability of the watermark, it has been the goal for almost all previous papers on digital watermarking. We suggest that a combination of spread spectrum watermarking technique [4] with our scheme by modulating each watermark bit to ℓ image pixels with again signature-seeded cryptographical pseudo random sequence as the PN-sequence would further improve the non-removability.

IV. Experimental Results

The common test image, 256-level 512×512 greyscale "Lenna", is used to evaluate the proposed scheme. The primary goals of these experiments are to decide the watermark intensity parameter d in the last section and to demonstrate the robustness of the watermarking scheme against common signal processing operations and geometrical transformations.

Figure 1 shows the $PSNR$ image quality of the published image under the "brute-force watermark removal attack" over the choices of the watermark strength parameter d, where the "brute-force watermark removal attack" means that an attacker transforms the stego image to the transform domain and clear the corresponding bit of every pixel in LL according to the public parameter d of the embedding algorithm; then transforms back to spatial domain to get an image without the owner's watermark.

The goal is to embed the watermark into perceptually significant part of the cover image while keeping it invisible. Therefore, $d = 6$ is used for the following experiments in demonstrating the robustness of the embedded watermark.

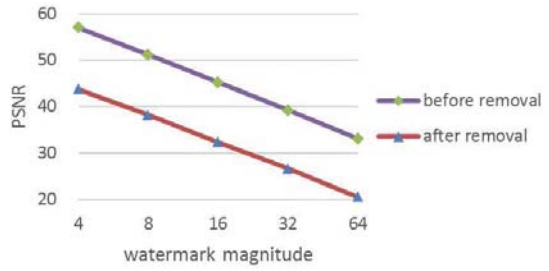

Fig. 1. Quality of the watermarked image under brute-force watermark removal attack versus the intensity of the watermarks

In order to clarify one's ownership of a particular image, it requires that the disputed image looks very close to the one provided by the owner, namely, $PSNR > 30$. If some spatial transformations have been performed on the disputed image, some preprocessing calibrations with respect to the original image have to be carried out before the watermark extraction, the calculation of the $PSNR$ and the residual correct percentage of the extracted watermark. For example, the calibrations include shifting, rotating, or scaling linearly back to the owner's original image.

Figure 2 shows the $PSNR$ image quality of the published image and correct percentage of the residual watermark versus JPEG encoding quality factors from 80% down to 40%. Note that the NC values stay far above 60%.

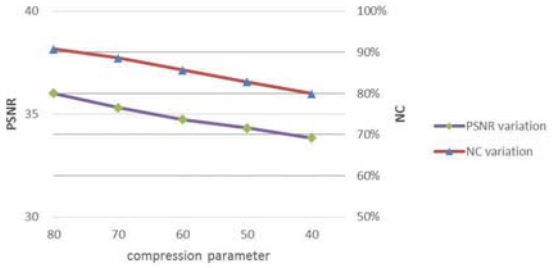

Fig. 2. Image quality and correct percentage of the residual watermark versus JPEG encoding quality factors

Figure 3 shows the quality of the watermarked image and correct percentage of the residual watermark versus scaling factors from 150% down to 25%. We should focus on those results with $PSNR$ above 30.

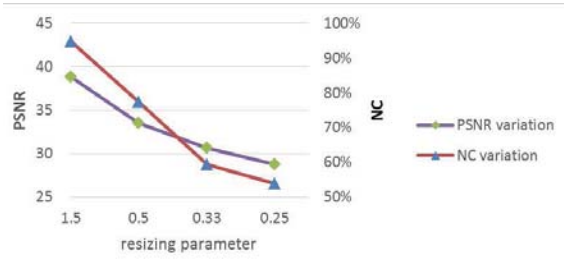

Fig. 3. Image quality and correct percentage of the residual watermark versus scaling factors

Figure 4 shows the quality of the watermarked image and correct percentage of the residual watermark versus rotation angles from 0.1 degrees to 0.5 degrees.

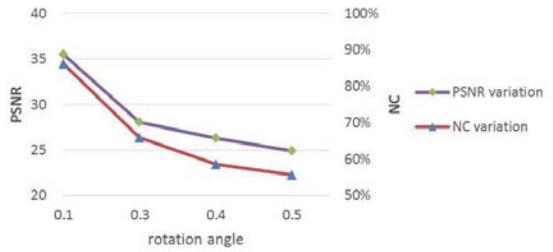

Fig. 4. Image quality and correct percentage of the residual watermark versus rotation angles

Figure 5 shows the quality of the watermarked image and correct percentage of the residual watermark versus horizontal shifting magnitude from 0.2 pixels to 0.8 pixels, which simulate the slight errors after the calibration.

Figure 6 shows the quality of the watermarked image and correct percentage of the residual watermark under the addition of the "salt and pepper" type noise with the density varying from 0.005 to 0.025, where the density indicates the percentage of polluted pixels.

Figure 7 shows the quality of the watermarked image and correct percentage of the residual watermark versus the cutoff

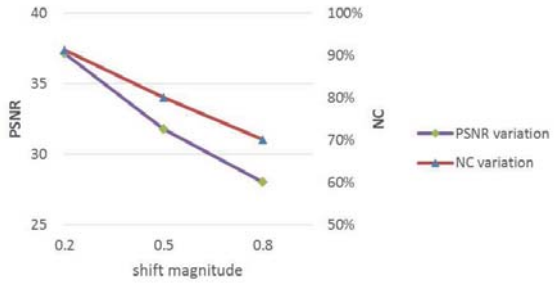

Fig. 5. Image quality and correct percentage of the residual watermark versus shifting offsets

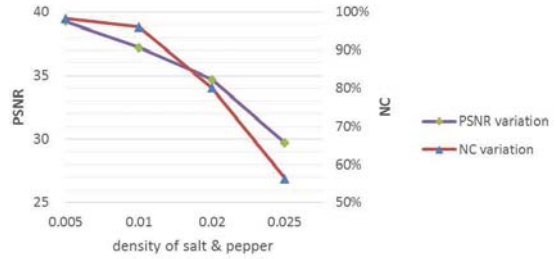

Fig. 6. Image quality and correct percentage of the residual watermark versus intensity of noise

frequencies of low pass filtering performed using DCT in the frequency domain.

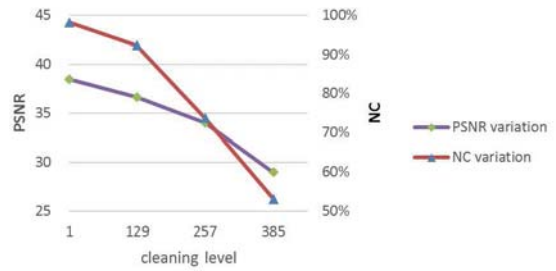

Fig. 7. Image quality and correct percentage of the residual watermark versus cutoff frequencies of low pass filtering

Through the above experiments, we can observe that as long as the $PSNR$ of the image is larger than 30, the amount of watermark bit sequence is far above the threshold we suggested such that the false negative probability is essentially negligible.

V. Conclusion

In this paper, a provable, cryptographic watermarking scheme that operates in the 'exhibition' model to provide 'proof of ownership' mechanism is proposed such that when a piece of exhibited digital content is duplicated, modified, or reproduced without proper authorization, the copyright owner can provide sufficient direct evidences in proving his/her ownership when a lawsuit is filed for the perceived infringement of intellectual property ownership. We build our scheme based on previous successful signal processing techniques but focus on how to employ unpredictable signature-seeded pseudo random bit sequence to make the false negative watermark detection rate computationally negligible. Supporting experiments are presented for the choice of scheme parameter and demonstrating the robustness of the scheme to unintentional watermark removal attacks. Most important is that we set up a theorem that proves affirmatively that as long as a small percentage of watermark survives from the adversarial watermark removal attack, it is a sufficient legal evidence that proves the infringement of copyright.

References

[1] L. Blum, M. Blum, and M. Shub, "A Simple Unpredictable Pseudo-Random Number Generator," SIAM Journal on Computing 15 (2): 364 - 383, 1986.

[2] M. Blum and S. Micali, "How to Generate Cryptographically Strong Sequences of Pseudo-random Bits," SIAM Journal on Computing 13 (4): 850-864, 1984, (extended abstract in FOCS 1982).

[3] D. Boneh and J. Shaw, "Collusion-Secure Fingerprinting for Digital Data," Advances in Cryptology - Crypto'95, LNCS 963, Springer-Verlag, 452 - 465, 1995.

[4] I. J. Cox, J. Kilian, T. Leighton, and T. Shamoon, "Secure Spread Spectrum Watermarking for Multimedia," IEEE Trans. on Image Processing, 6, 12, 1673 - 1687, 1997.

[5] I. Cox, M. Miller, J. Bloom, J. Fridrich, and T. Kalker, Digital Watermarking and Steganography, 2nd Ed., Morgan Kaufmann Publishers Inc., 2008.

[6] J. Fridrich, "Robust Bit Extraction from Images," Proc. IEEE Intern. Conf. Multimedia Computing and Systems (ICMCS'99), 1999.

[7] C. Fei, D. Kundur, and R. H. Kwong, "Analysis and Design of Secure Watermark-Based Authentication Systems," IEEE Trans. on Information Forensics and Security 1(1), 2006.

[8] S. Goldwasser, S. Micali, and R. Rivest, "A Digital Signature Scheme Secure against Adaptive Chosen-Message Attacks," SIAM J. Computing, Vol. 17, No. 2, pp.281–308, 1988.

[9] O. Goldreich, Foundations of Cryptography: Volume 1, Basic Tools, Cambridge University Press, 2000.

[10] C.-Y. Lin and S.-F. Chang, "Generating Robust Digital Signature for Image/Video Authentication," Multimedia and Security Workshop at ACM Multimedia'98, 1998.

[11] B. Pfitzmann and M. Schunter, "Asymmetric Fingerprinting," Advances in Cryptology - Eurocrypt'96, LNCS 1070, Springer-Verlag, 84 - 95, 1996.

[12] K. SriSwathi and S. G. Krishna, "Secure Digital Signature Scheme for Image Authentication over Wireless Channels," Int. J. Comp. Tech Aool. 2 (5): 1472 - 1479, 2011.

[13] L. F. Turner, "Digital Data Security System," Patent IPN WO 89/08915, 1989.

[14] X.-G. Xia, C. G. Boncelet, and G. R. Arce, "Wavelet Transform Based Watermark for Digital Images," Comm. ACM 57(3): 86 - 95, 2014.

[15] E. Zielinska, W. Mazurczyk, and K. Szczypiorski, "Trends in Steganography," Comm. ACM 57(3): 86 - 95, 2014.