Resistance of Anti-Disclosure Image Watermark to Collusion and Copy Attacks: An Approach of Combining Perceptual Hash and Watermark

Chun-Shien Lu and Chao-Yong Hsu

Institute of Information Science, Academia Sinica

Taipei, Taiwan 115, Republic of China

E-mail: {lcs,cyhsu}@iis.sinica.edu.tw

Abstract

Robustness is a critical requirement regarding the practicability of a watermarking scheme. Current watermarking methods usually claim a certain degree of robustness against those attacks that aim to destroy the hidden watermark at the expense of degrading the quality of media data. However, there exists a kind of watermark-estimation attack (WEA) such as the collusion attack that can remove watermarks meanwhile the attacked data can be made further transparent to its original. Another kind is the copy attack that can create the protocol ambiguity problem within a watermarking system. The motivation of this paper is dedicated to cope with the WEA that is clever at disclosing hidden information for unauthorized purposes. To this end, we begin by gaining an insight into the WEA that leads to formal definitions of optimal watermark estimation and perfect cover data recovery. Subject to these definitions, content-dependent watermark (CDW) is proposed to resist watermark-estimation attack. The key point is to introduce a media hash as a constituent component of the CDW. Mathematical analyses and experiment results have consistently verified the effectiveness of the content-dependent watermarking scheme. To our knowledge, the proposed anti-disclosure watermark is the first to resist both the collusion and the copy attacks.

Keywords: Collusion attack, Copy attack, Content-dependent watermark, Hash, Robustness, Watermark estimation

Technical Report **TR-IIS-03-012**, **Academia Sinica**

1 Introduction

Data hiding is a technology of embedding a piece of information into cover media data to carry out specific missions. In this emerging field, despite different requirements have been concerned, digital watermarking and steganography are two main branches that lead to a great amount of researches in the last decade [5, 6, 18, 20]. This paper will focus on digital watermarking, which has been popularly employed for copyright protection, content authentication, access control, and so on.

However, no matter what kinds of applications are considered, robustness is the popular and critical issue that will affect the practicability of a watermarking system. In data hiding, robustness refers to the capability of resistance to attacks, which are used to destroy, remove, or disable watermark detection. As previously discussed in [27], attacks can be classified into four categories: (1) removal attack; (2) geometrical attack; (3) cryptographic attack; and (4) protocol attack. The robustness of current watermarking methods was popularly examined on either removal attack or geometrical attack or both. In particular, removal attack contains those operations, including filtering, compression, and noise adding, that will more or less degrade the quality of media data. Even though the employed removal attack cannot guarantee to remove the hidden watermark successfully, media's quality has been inevitably destroyed. Despite their different removal capabilities, there indeed exists a kind of attacks that can fail a watermarking system and generate attacked data further perceptually close to its cover version simultaneously. Among currently known attacks [27], the collusion attack [8, 21, 22] that belongs to the removal attack and the copy attack [9] that belongs to the protocol attack are the typical examples, which can achieve the aforementioned goal. The common step of realizing the collusion and the copy attack is the watermark estimation, which is commonly accomplished by means of a denoising procedure. Consequently, we call both of the collusion attack and the copy attack, watermark-estimation attack (WEA).

The aim of collusion attack is to collect and analyze a set of watermarked media data[†] in order that an unwatermarked copy is constructed to raise the false negative problem. In digital watermarking, collusion attack naturally occurs in video watermarking because a video is composed of many frames and one way of watermarking a video is to embed the same watermark into all frames. This scenario has been firstly addressed in [22]. However, we argue that collusion attack is not exclusive to video watermarking. In the past few years, image watermarking with resistance to geometrical attack has

[†]This set of watermarked media data in fingerprinting [24] is generated from the same cover data but individually embedded with different watermarks, while in watermarking it is generated from visually similar/dissimilar image blocks or video frames embedded with the same watermark.

received much attention because even a slight geometrical distortion may have disordered the hidden watermark to disable watermark detection. In view of this, some researches [1, 23, 28] have been presented by inserting multiple redundant watermarks into an image with the hope that it suffices to maintain robustness as long as at least one watermark exists. The common framework is that some kinds of image units such as blocks [28], meshes [1], or disks [23] are extracted as carriers for embedding. With this unique characteristic, we propose to treat each image unit in an image like a frame in a video, and thereby, collusion attack can be equally applied to those image watermarking methods by embedding multiple redundant watermarks. Therefore, once the hidden watermarks are successfully removed by means of a collusion attack, the false negative problem occurs even no geometrical attack is imposed on stego images. Of particular interest are the possible fidelity improvements of attacked images by means of a collusion attack. Hence, the collusion attack shows a promising advantage over its counterparts of removal attack.

In contrast to the collusion attack, copy attack [9] has been developed to create the false positive problem; i.e., one can successfully detect a watermark from unwatermarked data. The copy attack is first applied to image watermarking and is operated as follows: (i) a watermark is first predicted from a stego image; (ii) the predicted watermark is added into a target image to create a counterfeit watermarked image; and (iii) from the counterfeit image, a watermark can be detected that wrongly claims rightful ownership. Compared with the collusion attack, copy attack is not restricted to be executed on more than one media data, and thus, is more flexible. We will also show that the achievable capability of copy attack is rather easier than denoising attack (a special case of collusion attack). It is in this respect that copy attack must be involved in referring to the issue of robustness.

In this paper, our motivation is to propose a new scheme to cope with the watermark-estimation attack (WEA). Throughout this paper, we shall exemplify image watermarking to discuss our method. After introducing a general watermarking framework in Sec. 2, watermark estimation attack will be thoroughly explored in Sec. 3. We begin by investigating the achievable performance of the denoising attack and the copy attack to reveal that copy attack is, in fact, easier to achieve. Then, due to the prerequisite of WEA is watermark estimation, we formally define the so-called "optimal watermark estimation" and "perfect cover data recovery", and analyze the confidence of collusion attack. From our analyses, we know that both accurate estimation of watermark's sign and complete subtraction of watermark's energy are two indispensable components to achieve effective watermark removal. On the other hand, they also serve as the clues to break WEA. In order to withstand WEA, we propose to embed the content-dependent watermark (CDW), which is composed of an informative watermark

that carries information about an owner and a media hash that represents the cover carrier. The design of media hash will be addressed in Sec. 4. Based on the presented media hash, in Sec. 5 CDW is constructed and its properties will be examined. Furthermore, the validity of resistance to collusion attack and copy attack using CDW will be analyzed. We would like to clarify that the concept of CDW was firstly proposed in [22] to solve the deadlock problem, however, we further explore it in withstanding watermark-estimation attack. Finally, experimental results and concluding remarks will be summarized in Sec. 6 and 7, respectively.

2 Basic Framework of Digital Watermarking

A general digital watermarking scheme is illustrated in Fig. 1. In the embedding process, a watermark is a message (author key) first converted into a binary sequence and then encoded as $\mathbf{S}$ using an error correction code (ECC) to enhance error correction. Before embedding, the values of the ECC encoded sequence $\mathbf{S}$ are mapped from $\{0\ \ 1\}$ to $\{-1\ \ 1\}$ so that $\mathbf{S}$ is a Gaussian distribution $\sim \mathcal{N}(0,1)$. Then, $\mathbf{S}$ is shuffled by means of a secret key K known by the owner only. Finally, the resultant sequence $\mathbf{S}$ will be magnified under the constraint of perceptual masking $\mathbf{M_I}$ and embedded into a cover image $\mathbf{I}$ to produce a corresponding watermarked (or stego) image $\mathbf{I^s}$ as

$$I^s(i) = I(i) + S(i)M_I(i)\ \forall i \in [1\ \ L],$$

where L denotes the length of $\mathbf{S}$ and $\mathbf{M_I}$ stands for the masking matrix derived from $\mathbf{I}$ (of course, $\mathbf{M_I}$ and $\mathbf{I}$ have the same dimension). We call the finally embedded sequence $\mathbf{S} \cdot \mathbf{M_I}$ as a watermark $\mathbf{W}$. The hidden watermark is assumed to be a Gaussian sequence and satisfies $\mathbf{W} \sim \mathcal{N}(0, \rho^2)$. Note that the variance ρ^2 coming from $\mathbf{M_I}$ determines watermark's energy and varies with different images. It is mainly used to increase watermark's robustness. Besides, $\mathbf{S}$ determines watermark's sign and is secured by means of the secret key K. The energy $||\mathbf{W}||_2$ of $\mathbf{W}$ can be derived to be $\sqrt{L}\rho$.

In the detection process, a watermark $\mathbf{W^e}$ is first extracted and decoded into a bipolar sequence $\mathbf{S^e}$ using

$$\mathbf{S}^e(i) = sgn(\mathbf{W}^e(i)), \tag{1}$$

where the sign function, $sgn(\cdot)$, is defined as

$$sgn(t) \ = \ \begin{cases} +1, & \text{if } t \geq 0, \\ -1, & \text{if } t < 0. \end{cases}$$

Due to the high-frequency property of a watermark signal, denoising is naturally an efficient way of achieving blind watermark extraction [7, 9, 26]. It is said that a hidden watermark exists provided that the normalized correlation δ_{nc} between $\mathbf{S}$ and $\mathbf{S^e}$ (both are with equal energy $\sqrt{L}$) is larger than a pre-determined threshold T, where

$$\delta_{nc}(\mathbf{S}, \mathbf{S^e}) = \frac{1}{L} \sum S(i) S^e(i) \tag{2}$$

and $\delta_{lc}(\cdot, \cdot) \in [-1 \;\; 1]$. In fact, Eq. (2) is also equal to $1 - 2P_e$, where P_e stands for the bit error rate (BER).

While numerous methods have been claimed to be robust, they are actually only robust from one perspective. From another perspective, they are extremely fragile. In the next section, we shall discuss watermark-estimation attack that will make a watermark system do a wrong decision about the existence of a hidden watermark. In particular, collusion attack and copy attack are investigated.

3 Watermark Estimation Attack

In [27], watermark attacks are roughly classified into four categories: removal attack, geometrical attack, cryptographic attack, and protocol attack. Among them, removal attack tries to vanish the hidden watermark by manipulating a stego image $\mathbf{I^s}$ so that the quality of an attacked image $\mathbf{I^a}$ is further destroyed. Specifically, $PSNR(\mathbf{I}, \mathbf{I^s}) \geq PSNR(\mathbf{I}, \mathbf{I^a})$ always holds for conventional removal attacks. However, a more clever removal attack exists to achieve $PSNR(\mathbf{I}, \mathbf{I^s}) \leq PSNR(\mathbf{I}, \mathbf{I^a})$, i.e., an attacked image $\mathbf{I^a}$ is even closer (in terms of mean square error (MSE)) to the cover image $\mathbf{I}$ than the stego image $\mathbf{I^s}$. The collusion attack is a typical example that will satisfy the above scenario. Usually, a collusion attack is applied to video watermarking, as depicted in Fig. 2, by averaging of a set of estimated watermarks to obtain the hidden watermark. In image watermarking, some recent work has been proposed to embed multiple redundant watermarks into local areas [1, 23, 28] in order that global/local geometrical distortions can be resisted. Provided we treat a local region in an image like a video frame in a video, then the same collusion attack can also be applied to region-based image watermarking. Once the hidden watermark is estimated, it is subtracted from all image blocks to yield an unwatermarked image. Under these circumstances, a false negative problem appears. It should be noted that the conventional denoising-based removal attack [26] only applied to one single image is a special case of the collusion attack.

On the other hand, the estimated watermark can be inserted into an unwatermarked media data to produce a counterfeit stego data. This is the so-called copy attack [9], which has been developed to

create the false positive problem; i.e., one can successfully detect a watermark from an unwatermarked data. The copy attack, as depicted in Fig. 3, belongs to the type of protocol attack. The use of copy attack is very simple because it can be performed on only one piece of media data (e.g., an image).

To our knowledge, both the collusion attack and the copy attack have not been, simultaneously, taken into consideration in investigating the robustness issue. Since watermark estimation is the first step towards the above two attacks, which are called watermark-estimation attack (WEA). In the next section, we shall analyze the respective performance of watermark-estimation attack. Here, blind watermark extraction is accomplished by means of a denoising process [7, 9, 15, 26].

3.1 Analysis of Achievable Performance of Denoising Attack and Copy Attack

Two typical examples of watermark-estimation attack, i.e., the denoising attack [26] (recall that it is a special case of the collusion attack) and the copy attack [9], will be discussed. Without loss of generality, suppose the decision on a watermark's existence will be based on the normalized correlation, as defined in Eq. (2). Let $\mathbf{X}$, $\mathbf{X^s}$, $\mathbf{Z}$, and $\mathbf{Z^s}$ represent the original image, watermarked image, faked original image, and faked watermarked image, respectively. Among them, $\mathbf{X^s}$ is generated from $\mathbf{X}$ through an embedding process, and $\mathbf{Z^s}$ is generated from the combination of $\mathbf{Z}$ and a watermark extracted from $\mathbf{X^s}$.

Let $\mathbf{W}$ be a watermark (described in Sec. 2) to be hidden in $\mathbf{X}$, and let $\mathbf{W^e}$ be an estimated watermark obtained by denoising $\mathbf{X^s}$. For the purpose of watermark removal, $\mathbf{W^e}$ will be subtracted from $\mathbf{X^s}$ to produce an attacked image $\mathbf{X^a}$, i.e.,

$$\mathbf{X^a} = \mathbf{X^s} - \mathbf{W^e}.$$

In the watermark detection process, a watermark, $\mathbf{W^a}$, is extracted from $\mathbf{X^a}$ and correlated with $\mathbf{W}$. If denoising-based watermark removal is expected to succeed, then $\delta_{nc}(sgn(\mathbf{W}), sgn(\mathbf{W^a})) \leq T$ must hold. This result indicates that the ratios of the correctly (C_w) and wrongly (NC_w) decoded watermark bits should, respectively, satisfy

$$C_w \leq \frac{1+T}{2}$$

and

$$NC_w \geq \frac{1-T}{2}, \tag{3}$$

where $C_w + NC_w = 1$ and NC_w corresponds to BER. Based on the false analyses of normalized correlation (pp. 186 of [2]), if we would like to have a false positive probability at the level of 10^{-8}

when $|\mathbf{W}| = 1024$, then the threshold T should be set to be 0.12. As a consequence, it is evident from the above analyses that an efficient watermark removal attack should be able to vanish *most* watermark bits since T is usually small. In fact, the actual number of bits has been specified in Eq. (3).

As for the copy attack, the estimated watermark $\mathbf{W^e}$ is added to the target image $\mathbf{Z}$ to form a counterfeit image $\mathbf{Z^s}$, i.e.,

$$\mathbf{Z^s} = \mathbf{Z} + \mathbf{W^e}. \tag{4}$$

In the watermark detection process, a watermark, $\mathbf{W^z}$, is extracted from $\mathbf{Z^s}$ and correlated with $\mathbf{W}$. Copy attack is claimed to succeed if $\delta_{nc}(sgn(\mathbf{W}), sgn(\mathbf{W^z})) > T$ holds. This implies that the ratio of the correctly decoded watermark bits C_w only needs to be at least increased from $\frac{1}{2}$ (due to the randomness of an arbitrary image, $\mathbf{Z}$) to $\frac{1+T}{2}$. Actually, the amount of increase, ξ^{copy}, only needs to satisfy

$$\xi^{copy} \geq \frac{1+T}{2} - \frac{1}{2} = \frac{T}{2}. \tag{5}$$

Comparing Eqs. (3) and (5), we can conclude that a copy attack is easier to perform successfully than a denoising attack because $\frac{1-T}{2}$ is quite larger than $\frac{T}{2}$ based on the fact that T is usually a smaller number. Under the situation of $T = 0.12$, we can obtain $\frac{1-T}{2} = 0.44 > \frac{T}{2} = 0.06$. However, if the denoised results (i.e., more than one estimated watermarks) are collected and colluded to generate an estimation more close to its origin, then the collusion attack will show more powerful performance than the denoising attack, as evidenced in [21, 22].

3.2 Optimal Watermark Estimation and Perfect Cover Data Recovery

In this section, we shall explore the definitions of "optimal watermark estimation" and "perfect cover data recovery" from which a strategy that can combat with watermark-estimation attack is addressed.

3.2.1 Formal Definition

From an attacker's perspective, the energy of each watermark bit must be accurately predicted so that the previously added watermark energy can be completely subtracted to accomplish effective watermark removal. Especially, an estimated watermark's energy is closely relevant to the accuracy of watermark removal attack. These scenarios are shown in Fig. 4, which illustrates the energy variations of (a) an original watermark; (b)/(d) an estimated watermark (illustrated in gray-scale); and (c)/(e) a residual watermark generated by subtracting the estimated watermark from the original

watermark. From Fig. 4(a)∼(c), we can realize that even watermark's sign bits are fully obtained but the corresponding energies cannot be completely discarded, the residual watermark still suffices to reveal the encoded message according to Eq. (1). Furthermore, if the sign of an estimated watermark bit is different from its original one (i.e., $sgn(W(i)) \neq sgn(W^e(i))$), any more energy subtraction will not helpful in improving removal efficiency. On the contrary, watermark removal in terms of energy subtraction operated in the opposite (wrong) polarity will undesirably destroy the media data's fidelity severely. Actually, this corresponds to even add a watermark with higher energy into a cover data without obeying the masking constraint, as shown in Fig. 4(d). By subtracting Fig. 4(d) from Fig. 4(a), the resultant residual watermark is illustrated in Fig. 4(e). By correlating Figs. 4(a) and (e), it is highly possible to reveal the existence of a watermark.

With this understanding, we shall give formal definitions of "optimal watermark estimation" and "perfect cover data recovery", respectively, as follows from an attacker's viewpoint for further analyses.

Definition 1 (Optimal Watermark Estimation): A watermark is said to be optimally estimated and then removed if watermark detector fails to show the presence of a watermark in terms of correlation. Given an original embedded watermark $\mathbf{W}$ and its approximate version $\mathbf{W^e}$ estimated from $\mathbf{I^s}$ using either a watermark removal attack or a collusion attack, the necessary condition of optimal estimation of $\mathbf{W}$ as $\mathbf{W}^e$ is defined as[‡]

$$\delta_{nc}(sgn(\mathbf{W}), sgn(\mathbf{W^e})) > T, \tag{6}$$

where $sgn(\mathbf{v}) = \{sgn(v_1), sgn(v_2), ..., sgn(v_L)\}$ represents the signs of elements in a vector $\mathbf{v} = \{v_1, v_2, ..., v_L\}$ and Θ denotes the set of indices satisfying $sgn(W^e(i)) = sgn(W(i))$. This states the first step that a watermark is possible to be undetected by an owner if more than $\frac{L(1+T)}{2}$ sign bits[§] of a watermark can be obtained by an attacker. Beyond this, however, in order to avoid the residual watermark (as illustrated in Fig. 4(c)) left to reveal the trace about the hidden watermark, the accurate estimation of the energy of $\mathbf{W^e}$ is absolutely indispensable. In addition to Eq. (6), watermark removal can be really achieved if the watermark energy to be subtracted is also larger than or equal to the added energy, i.e., $mag(W^e(i)) \geq mag(W(i))$, where $mag(t)$ denotes the magnitude

[‡]In Eq. (6), the capability of collusion attack towards optimal estimation of watermark's sign can be further justified in Appendix.

[§]The importance of polarities of watermark bits has been previously emphasized in [12] by embedding two complementary watermarks that are modulated in different ways to resist different sets of attacks, and in [13] by exploiting the prior knowledge of watermark detector to design a watermarking scheme wherein both the denoising and the copy attacks can be resisted.

$|t|$ of t. Therefore, the sufficient and necessary condition for optimal watermark estimation can be defined to be

$$mag(W^e(i)) \geq mag(W(i)) \text{ and } sgn(W^e(i)) = sgn(W(i)) \ \forall i \in \Theta. \tag{7}$$

After employing the optimal watermark estimation derived in Definition 1, the extracted watermark will behave like a random signal so that no trace about the watermark cannot observed. In order to make collusion attack invalid, it is necessary to force attackers to get $\delta_{nc}(sgn(\mathbf{W}), sgn(\mathbf{W^e})) \leq T$. Intuitively, the basic requirement to deter collusion is to satisfy $\delta_{nc}(sgn(\mathbf{W}), sgn(\mathbf{W^e})) = 0$, which is equivalent to random guess, i.e, half of the watermark's sign bits are correctly guessed and the other half are wrongly guessed. Therefore, we propose in Sec. 4 a media hash extraction method to construct the so-called content-dependent watermark that attempts to lower the confidence of collusion (as derived in Eq. (25)).

Definition 2 (Perfect Cover Data Recovery): Under the prerequisite that Definition 1 (Eq. 7) is satisfied, it is said that $\mathbf{I^r}$ is an perfect recovery of $\mathbf{I}$ if

$$PSNR(\mathbf{I}, \mathbf{I^r}) \approx \infty, \tag{8}$$

where $\mathbf{I^r} = \mathbf{I} - sgn(\mathbf{W^e})mag(\mathbf{W^e})$ and $mag(\mathbf{v}) = \{mag(v_1), mag(v_2), ..., mag(v_L)\}$ represents the magnitudes of elements in a vector $\mathbf{v} = \{v_1, v_2, ..., v_L\}$. Of course, it is best to get $mag(W^e(i))$ as the upper bound of $mag(W(i))$; otherwise, even watermarks have been completely removed the qualities of attacked images would be poor. Typically, evaluation of $mag(\mathbf{W^e})$ can be either achieved by averaging [22] or remodulation [26].

In sum, under the condition of sufficiently large $\delta_{nc}(sgn(\mathbf{W}), sgn(\mathbf{W^e}))$, $PSNR(\mathbf{I}, \mathbf{I^s}) \leq PSNR(\mathbf{I}, \mathbf{I^r})$ will hold undoubtedly. Unlike other watermark removal attacks that will destroy the quality of the media data, collusion attack is possible to improve the quality of colluded data. To facilitate our later analyses, the following lemma will be introduced in advance.

Lemma 1 Given any two independent variables, $\mathbf{w_1}$ and $\mathbf{w_2}$ of the same length L, that are constructed from independent, identically distributed (i.i.d.) samples drawn from a Gaussian distribution $\mathcal{N}(0, \rho^2)$, the normalized correlation between $\mathbf{w_1}$ and $\mathbf{w_2}$ will be

$$\begin{aligned} \delta_{nc}(\mathbf{w_1}, \mathbf{w_2}) &= \frac{1}{L\rho^2} E(\mathbf{w_1 w_2}) \\ &= \frac{1}{L\rho^2} E(\mathbf{w_1}) E(\mathbf{w_2}) \\ &= 0, \end{aligned} \tag{9}$$

by exploiting the fact that $Cov(\mathbf{w_1}, \mathbf{w_2}) = 0$ for the independence of $\mathbf{w_1}$ from $\mathbf{w_2}$. For convenient analyses later, both $\mathbf{w_1}$ and $\mathbf{w_2}$ are restricted to have the same variance. It should be noted that the two vectors, $\mathbf{w_1}$ and $\mathbf{w_2}$, have been normalized to have the same energy $\sqrt{L}\rho$. If $\rho = 1$, then normalized correlation will be degenerated as linear correlation. Besides, it can be derived that if $\mathbf{w_1} = \mathbf{w_2}$, then $\delta_{nc}(\mathbf{w_1}, \mathbf{w_2}) = 1$.

4 Media Hash

From the analyses of watermark-estimation attack (WEA) described in Sec. 3, we have found that the success of WEA mainly lies on the fact that the hidden watermark totally behaves like a noise such that anyone can fully and reliably utilize all estimated noise-like watermarks. In order to disguise this prior knowledge that is known and favorable to attackers, a watermark must be designed to carry information relevant to the cover image itself. Meanwhile, the content-dependent information must be secured[¶] by means of a secret key and robust to digital processing [14] in order not to affect watermark detection. To this end, we shall introduce the concept of media hash as a kind of content-dependent information to create the so-called content-dependent watermark (CDW) in the following.

Media hash [3, 25], also known as "digital signature" [11, 14] or "media fingerprint" [4], has been proposed in many applications, including content authentication, copy detection, media recognition. In this paper, we shall employ the essential of either wavelet-based or block DCT-based digital signature to construct the media hash. In the following, the proposed image hash extraction algorithm that can coordinate with block-based image watermarking method (e.g., [28]) and adapt to video watermarking standards is introduced in the 8×8 block-DCT domain. However, we would like to particularly emphasize that depending on different watermarking algorithms (e.g., [1, 23]) the proposed media hash extraction method can be adjusted correspondingly. For a pair of 8×8 blocks, a piece of representative and robust information is created. It is defined as the magnitude relationship between two AC coefficients at blocks i and j:

$$r(k) = \begin{cases} +1, & \text{if } |f_i(p_1)| - |f_j(p_2)| \geq 0 \ (p_1 \neq p_2) \\ -1, & \text{otherwise,} \end{cases} \tag{10}$$

where $r(k)$ is an element of a feature sequence $\mathbf{r}$ and $f_u(v)$ denotes an AC coefficient at position v in block u. The DC coefficients will not be selected because they are not sufficient to represent unique

[¶]This is because either an owner or an attacker can freely derive content-dependent information. Hence, a secret key is required for shuffling. How to combine shuffled content-dependent information and watermark will be described in Sec. 5.

features of image blocks. In addition, the selected AC coefficients should be at lower frequencies because high-frequency coefficients are vulnerable to attacks. In this paper, p_1 and p_2 are selected to be the first two largest AC coefficients from the 64 available frequency subbands. As a result, there are 2 feature values resulted in Eq. (10) by interchanging p_1 and p_2 for each pair of 8×8 blocks. We call this feature value $r(\cdot)$ robust because this magnitude relationship between $f_i(p_1)$ and $f_j(p_2)$ can be mostly preserved under incidental modifications (e.g., compressions, filtering, denoising). Please refer to [11, 14] for similar robustness analyses.

Let $M \times M$ be the block size of a "local region" where a watermark is embedded. Assume M is chosen to be a multiplier of 8. Hence, the number of 8×8 blocks is $(\frac{M}{8})^2$ from which the total pairs of blocks will be $\mathcal{P}\begin{pmatrix} (\frac{M}{8})^2 \\ 2 \end{pmatrix}$, where $\mathcal{P}$ denotes the permutation function. As a result, the total number of feature values at present will be $2\mathcal{P}\begin{pmatrix} (\frac{M}{8})^2 \\ 2 \end{pmatrix}$. In practice, each media hash must be constructed within the range where one watermark is embedded in order that resistance to geometrical distortions still can be preserved. Under this constraint, when the sequence $\mathbf{r}$ is extracted, it is repaired such that a desired image hash with $|\mathbf{r}| = L$ can be created. If $2\mathcal{P}\begin{pmatrix} (\frac{M}{8})^2 \\ 2 \end{pmatrix} > |\mathbf{W}|$, then the extra elements at the tail of $\mathbf{r}$ is deleted; otherwise, $\mathbf{r}$ is cyclically appended. We call the finally created sequence as the media hash $\mathbf{MH}$, which is a bipolar sequence. Next, the media mash $\mathbf{MH}$ of an image is mixed with the watermark $\mathbf{W}$ to generate the content-dependent watermark ($\mathbf{CDW}$) as

$$\mathbf{CDW} = S(\mathbf{W}, \mathbf{MH}), \tag{11}$$

where $S(\cdot, \cdot)$ is a mixing function, which is basically application-dependent and will be used to control the combination of $\mathbf{W}$ and $\mathbf{MH}$. The sequence $\mathbf{CDW}$ is what we will embed into a cover image.

5 Image-Dependent Watermark

In this section, the properties of image-dependent watermark will be first discussed. Then, its resistance to WEA will be analyzed on block-based image watermarking (e.g., [28]).

5.1 Properties

Let an image $\mathbf{I}$ be expressed as $\oplus_{i \in \Omega} \mathbf{B}_i$, where all blocks $\mathbf{B}_i$ are concatenated to form $\mathbf{I}$ and Ω denotes the set of block indices. As far as block-based image watermarking scheme [1, 23, 28] is concerned, each image block $\mathbf{B}_i$ will be embedded with a content-dependent watermark $\mathbf{CDW}_i$ to form a stego

image $\mathbf{I^s}$, i.e.,

$$\mathbf{B^s}_i = \mathbf{B}_i + \mathbf{CDW}_i, \quad \mathbf{I^s} = \oplus_{i\in\Omega}\mathbf{B^s}_i, \tag{12}$$

where $\mathbf{B^s}_i$ is a stego block and $\mathbf{CDW}_i$ similar to Eq. (11) is defined as the mixture of a fixed informative watermark $\mathbf{W}$ and a block-based hash $\mathbf{MH_{B}}_i$, i.e.,

$$\mathbf{CDW}_i = S(\mathbf{W}, \mathbf{MH_{B}}_i). \tag{13}$$

Recall that $\mathbf{W}$ previously defined in Sec. 2 is equal to $\mathbf{S}\cdot\mathbf{MH_{B}}_i$. As $\mathbf{MH_{B}}_i$ can be easily obtained by either owners or attackers, the mixing function $S(\cdot,\cdot)$ will be designed as a procedure of shuffling the media hash $\mathbf{MH_{B}}_i$ by the same secret key K (used to generate the watermark $\mathbf{W}$) and followed by shuffling the watermark to enhance security. Specifically, it is expressed as

$$S(\mathbf{W}, \mathbf{MH_{B}}_i)(k) = W(k)PT(\mathbf{MH_{B}}_i, K)(k), \tag{14}$$

where PT denotes a shuffling function controlled using the secret key K with the aim of achieving uncorrelated crosscorrelation:

$$\delta_{nc}(PT(\mathbf{MH_{B}}_i, K), \mathbf{MH_{B}}_i) = 0,$$

and autocorrelation:

$$\delta_{nc}(\mathbf{MH_{B}}_i, \mathbf{MH_{B}}_i) = \delta_{nc}(PT(\mathbf{MH_{B}}_i, K), PT(\mathbf{MH_{B}}_i, K)).$$

Using the strategy specified in Eq. (14), only one secret key is used in our watermarking scheme. The advantage is that no more secret keys are required to increase the burden of key management problem.

The proposed content-dependent watermark possesses the characteristic described as follows. They are useful to prove resistance to WEA.

Definition 3 Given two image blocks $\mathbf{B}_i$ and $\mathbf{B}_j$, their degree of similarity depends on the correlation between $\mathbf{MH_{B}}_i$ and $\mathbf{MH_{B}}_j$, i.e.,

$$\delta_{nc}(\mathbf{B}_i, \mathbf{B}_j) = \delta_{nc}(\mathbf{MH_{B}}_i, \mathbf{MH_{B}}_j). \tag{15}$$

Accordingly, we have: (i) if $\mathbf{B}_i = \mathbf{B}_j$, then $\delta_{nc}(\mathbf{B}_i, \mathbf{B}_j) = 1$; (ii) if $\mathbf{B}_i$ and $\mathbf{B}_j$ look visually dissimilar, then $\delta_{nc}(\mathbf{B}_i, \mathbf{B}_j) \approx 0$; and (iii) otherwise, $\delta_{nc}(\mathbf{B}_i, \mathbf{B}_j) < 1$.

Proposition 1 Given two image blocks $\mathbf{B}_i$ and $\mathbf{B}_j$, $\delta_{nc}(\mathbf{B}_i, \mathbf{B}_j)$, and their respectively embedded content-dependent watermarks $\mathbf{CDW}_i$ and $\mathbf{CDW}_j$ that are assumed to be i.i.d. Gaussian distributions$\sim\mathcal{N}(0, \rho^2)$, the following properties can be established: (i) $\delta(\mathbf{CDW}_i, \mathbf{CDW}_j)$ is linearly

proportional to $\delta(\mathbf{B}_i, \mathbf{B}_j)$; (ii) $\delta_{nc}(\mathbf{CDW}_i, \mathbf{CDW}_j) \leq \delta(\mathbf{W}^2)$; (iii) $\delta(\mathbf{W}, \mathbf{CDW}) = 0$.

Proof (i): Substitution of Eqs. (13)$\sim$(15) into $\delta_{nc}(\mathbf{CDW}_i, \mathbf{CDW}_j)$ will lead to

$$\begin{aligned}
\delta_{nc}(\mathbf{CDW}_i, \mathbf{CDW}_j) &= \delta_{nc}(S(\mathbf{W}, \mathbf{MH}_{\mathbf{B}_i}), S(\mathbf{W}, \mathbf{MH}_{\mathbf{B}_j})) \\
&= \delta_{nc}(\mathbf{W}PT(\mathbf{MH}_{\mathbf{B}_i}, K), \mathbf{W}PT(\mathbf{MH}_{\mathbf{B}_j}, K)) \\
&= \frac{1}{L\rho^2}\sum_k W(k)^2 PT(\mathbf{MH}_{\mathbf{B}_i}, K)(k) PT(\mathbf{MH}_{\mathbf{B}_j}, K)(k) \\
&= \frac{1}{L\rho^2}\sum_k W(k)^2 MH_{B_i}(k) MH_{B_j}(k) \\
&\equiv \delta_{nc}(\mathbf{MH}_{\mathbf{B}_i}, \mathbf{MH}_{\mathbf{B}_j}) = \delta_{nc}(\mathbf{B}_i, \mathbf{B}_j), \qquad (16)
\end{aligned}$$

where the symbol "$\equiv$" is used to designate "linear proportion." Hence, property (i) implies that similar/dissimilar image blocks will be embedded with similar/dissimilar watermarks that are content-dependent. This result contrasts with the one pointed out in [22] but the novelty of our scheme is that the concept of content-dependent watermark has been employed.

Proof (ii): Follows by using Eqs. (14) and (16), we can obtain

$$\begin{aligned}
\delta_{nc}(\mathbf{CDW}_i, \mathbf{CDW}_j) &= \frac{1}{L\rho^2}\sum_k W(k)^2 [PT(\mathbf{MH}_{\mathbf{B}_i}, K)(k) PT(\mathbf{MH}_{\mathbf{B}_j}, K)(k)] \\
&\leq \delta_{nc}(\mathbf{W}^2), \qquad (17)
\end{aligned}$$

by exploiting the fact that $[PT(\mathbf{MH}_{\mathbf{B}_i}, K)(k) PT(\mathbf{MH}_{\mathbf{B}_j}, K)(k)]$ is either $+1$ or -1. If i is equal to j, then the equality in Eq. (17) holds. Hence, property (ii) implies that media hash is feasible to randomize the hidden watermark. In particular, this effect is remarkable for those visually dissimilar blocks, i.e., $\delta_{nc}(\mathbf{CDW}_i, \mathbf{CDW}_j) \approx 0$ if $\delta_{nc}(\mathbf{B}_i, \mathbf{B}_j) \approx 0$ (as specified in Definition 3).

Proof (iii): Since both $\mathbf{W}$ and $\mathbf{CDW}$ are i.i.d. Gaussian distributions with zero mean and the same variance ρ^2, it is easy to obtain

$$\delta_{nc}(\mathbf{W}, \mathbf{CDW}) = 0,$$

according to Lemma 1. Hence, property (iii) implies that watermark shuffling using the shuffled media hash can increase the difficulty of estimating the hidden watermark. If $\mathbf{W}$ is replaced with any random zero-mean Gaussian noise $\mathbf{n}$, zero correlation still can be obtained.

5.2 Resistance to Collusion Attacks

Now, resistance to collusion attack regarding the incorporation of the CDW and a block-based image watermarking method will be analyzed. By collusion attack, averaging operation is performed on stego blocks $\mathbf{B}^{\mathbf{s}}{}_i$'s of a stego image $\mathbf{I}^{\mathbf{s}}$. From an attacker's perspective, each hidden watermark has to

be estimated by a denoising operation so that deviations of estimation occur inevitably. Let $\mathbf{W^e}_i$ be an estimated watermark from $\mathbf{B^s}_i$. Without loss of generality, it is assumed to have the zero mean. In fact, $\mathbf{W^e}_i$ can modeled as a partial hidden watermark plus a noise component, i.e.,

$$\mathbf{W^e}_i = \alpha_i \mathbf{CDW}_i + \mathbf{n}_i, \tag{18}$$

where $\mathbf{n}_i$ represents an image block-dependent Gaussian noise with zero mean, α_i denotes the weight that the watermark has been extracted, and $\mathbf{W^e}_i \sim \mathcal{N}(0, \rho^2)$ is enforced to maintain the estimated watermark and the hidden watermark to have the same energy. Under these circumstances, $1 \geq \alpha_i = \delta_{nc}(\mathbf{W^e}_i, \mathbf{CDW}_i) > T$ always holds under the fact that a watermark is a high-frequency signal and can be efficiently estimated by denoising [7, 9, 26]. This factor α_i plays a crucial role in two aspects: (i) on one hand, from an attacker's viewpoint, α_i should be adjusted in a pixel/coefficient-wise manner so that perceptual fidelity can be maintained [26]; (ii) on the other hand, from an owner's viewpoint, a watermarking system should be able to allow large α_i in order that strong attacks can be tolerated. Let $\mathcal{C}$ ($\subset \Omega$) denote the set of blocks used for collusion. By employing the Central Limit Theorem, the average [‖] of all the estimated watermarks can be expressed as

$$\bar{\mathbf{W}}^{\mathbf{e}} = \frac{\sqrt{|\mathcal{C}|}}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} \mathbf{W^e}_i = \frac{1}{\sqrt{|\mathcal{C}|}} \sum_{i \in \mathcal{C}} (\alpha_i \mathbf{CDW}_i + \mathbf{n}_i), \tag{19}$$

because $\mathbf{W^e}_i$'s are obtained from (nearly) visually dissimilar image blocks, which can be regarded to be i.i.d. approximately. Now, we are ready to derive a sufficient and necessary condition of resisting collusion attack in Proposition 2.

Proposition 2 As far as collusion atatck is concerned, an attacker first estimates $\bar{\mathbf{W}}^{\mathbf{e}}$ from a set $\mathcal{C}$ of image blocks. Then, a counterfeit unwatermarked image $\mathbf{I^u}$ is generated from a watermarked image $\mathbf{I^s} = \oplus_{i \in \Omega} \mathbf{B^s}_i$ as

$$\mathbf{B^u}_i = \mathbf{B^s}_i - \bar{\mathbf{W}}^{\mathbf{e}}, \qquad \mathbf{I^u} = \oplus_{i \in \Omega} \mathbf{B^u}_i. \tag{20}$$

It is said that the collusion attack fails in an image block $\mathbf{B^u}_k$,$k \in \Omega$, i.e., $\delta_{nc}(\mathbf{B^u}_k, \mathbf{CDW}_k) > T$ if and only if $\delta_{nc}(\bar{\mathbf{W}}^e, \mathbf{CDW}_k) = \frac{\alpha_k}{\sqrt{|\mathcal{C}|}} < 1 - T$.

Proof: First of all, we need to derive $\delta_{nc}(\bar{\mathbf{W}}^{\mathbf{e}}, \mathbf{CDW}_k)$. Upon making use of Lemma 1, Eq. (19) and Proposition 1, we have the following derivation:

$$\delta_{nc}(\bar{\mathbf{W}}^{\mathbf{e}}, \mathbf{CDW}_k) \quad = \quad \frac{\sqrt{|\mathcal{C}|}}{|\mathcal{C}|} \delta_{nc}(\sum_{i \in \mathcal{C}} (\alpha_i \mathbf{CDW}_i + \mathbf{n}_i), \mathbf{CDW}_k)$$

[‖]Recall that the final estimation of the hidden watermark can also be conducted using perceptual masking to maintain fidelity of a colluded image, as previously discussed in Definition 2.

$$
\begin{aligned}
&= \frac{1}{\sqrt{|\mathcal{C}|}} \sum_{i \in \mathcal{C}} \alpha_i \delta_{nc}(\mathbf{CDW}_i, \mathbf{CDW}_k) + \frac{1}{\sqrt{|\mathcal{C}|}} \sum_{i \in \mathcal{C}} \delta_{nc}(\mathbf{n}_i, \mathbf{CDW}_k) \\
&= \frac{\alpha_k}{\sqrt{|\mathcal{C}|}}, \qquad (21)
\end{aligned}
$$

where $\mathbf{CDW}_k$ represents the content-dependent watermark embedded in $\mathbf{B}_k$. According to Eq. (21), our derivations are further explained as follows: the first row is resulted from Eq. (19) while the second term of the second row is zero by employing the independence of $\mathbf{n}_i$ from $\mathbf{CDW}_k$. Consequently, given property (ii) of Proposition 1, Eqs. (20) and (21), we get:

$$
\begin{aligned}
\delta_{nc}(\mathbf{B^u}_k, \mathbf{CDW}_k) > T \quad &\text{iff} \quad \delta_{nc}(\mathbf{B}_k + \mathbf{CDW}_k - \bar{\mathbf{W}}^{\mathbf{e}}, \mathbf{CDW}_k) > T \\
&\text{iff} \quad \delta_{nc}(\mathbf{CDW}_k, \mathbf{CDW}_k) - \delta_{nc}(\bar{\mathbf{W}}^{\mathbf{e}}, \mathbf{CDW}_k) > T \\
&\text{iff} \quad \delta_{nc}(\bar{\mathbf{W}}^{\mathbf{e}}, \mathbf{CDW}_k) = \frac{\alpha_k}{\sqrt{|\mathcal{C}|}} < 1 - T. \qquad (22)
\end{aligned}
$$

Remarks (Further interpretation of $|\mathcal{C}|$): If $|\mathcal{C}| = 1$ (we mean collusion attack is only applied on one block), then the collusion attack is degenerated into a denoising-based removal attack. Under this circumstance, the success of a collusion attack depends on the accuracy of estimation or the factor α_k (as pointed out previously, this factor plays a trade-off role between fidelity and robustness). By substituting $|\mathcal{C}| = 1$ into Eq. (22) and using $T < \alpha_k$, we get $T < 0.5$. In other words, α_k must be larger than or equal to 0.5 to guarantee the success of collusion attack when $|\mathcal{C}| = 1$. This result totally depends on the effectiveness of a denoising processing in estimating an added signal. Provided $|\mathcal{C}|$ becomes infinite, i.e., $|\mathcal{C}| = |\Omega| \to \infty$, then $\delta_{nc}(\bar{\mathbf{W}}^{\mathbf{e}}, \mathbf{CDW}_k) \to 0$ is obtained such that T can be a arbitrarily small but positive value, which means that the incorrectly estimated watermarks dominate the correctly estimated ones. On the other hand, the proposed content-dependent watermarking scheme is unfavorable to the collusion attack conducted from more and more image blocks. It is interesting to note that this result contradicts from the expected characteristic of collusion attacks. Particularly, performance degradation of the proposed method can be interpreted to be lower bounded by the denoising-based watermark removal attack (e.g., for $|\mathcal{C}| = 1$), as proved in Proposition 2 and later verified in experiments.

5.3 Resistance to Copy Attack

Next, we proceed to justify why the presented content-dependent watermark can be immune from the copy attack. Let $\mathbf{MH_X}$ and $\mathbf{MH_Z}$ denote the hash sequences generated from two different image blocks, $\mathbf{X}$ and $\mathbf{Z}$, respectively. In addition, let $\mathbf{CDW_X}$ denote the content-dependent watermark to be hidden into the cover image block $\mathbf{X}$. As has been stated previously, let the watermark estimated

from $\mathbf{X^s}$ be $\mathbf{W^x}$, which will contain partial information from $\mathbf{CDW_X}$. By directing the copy attack at the target block $\mathbf{Z}$, we can get the counterfeit watermarked block $\mathbf{Z^s}$ as defined in Eq. (4). Later, in the detection process, the content-dependent watermark, $\mathbf{W^z}$ estimated from the block $\mathbf{Z^s}$ will be

$$\mathbf{W^z} = (\alpha \times \mathbf{CDW_X} + \mathbf{n}), \tag{23}$$

according to Eq. (18), where $\mathbf{n}$ indicates the noise sequence (which is irrelevant to watermarks) generated by means of denoising $\mathbf{Z^s}$. Under the evidence that denoising is an efficient way to estimate watermarks [9, 26, 28], $||\alpha\mathbf{CDW_X}||_2 > ||\mathbf{n}||_2$ can undoubtedly hold, with $||\cdot||_2$ being the energy. Given Eqs. (13) and (23), Proposition 1, and Definition 3, normalized correlation between $\mathbf{CDW_Z}$ and $\mathbf{W^z}$ can be derived as follows based on blocks $\mathbf{X}$ and $\mathbf{Z}$ that are dissimilar:

$$\begin{aligned}
\delta_{nc}(\mathbf{CDW_Z}, \mathbf{W^z}) &= \frac{1}{|\mathbf{W}|\rho^2}\sum_{i=1}^{|\mathbf{W}|} CDW_Z(i)W^z(i) \\
&= \frac{1}{|\mathbf{W}|\rho^2}\sum_{i=1}^{|\mathbf{W}|}(\alpha CDW_Z(i)CDW_X(i) + CDW_Z(i)n(i)) \\
&= \alpha\delta_{nc}(\mathbf{CDW_Z}, \mathbf{CDW_X}) + \delta_{nc}(\mathbf{CDW_Z}, \mathbf{n}) \\
&\approx \alpha\delta_{nc}(\mathbf{CDW_Z}, \mathbf{CDW_X}) \approx 0,
\end{aligned} \tag{24}$$

where the term $\delta_{nc}(\mathbf{CDW_Z}, \mathbf{n})$ in the 2nd row approximates zero based on property (iii) of Proposition 1.

6 Experimental Results

In our experiments, ten varieties of gray-scale cover images of size 512×512, as shown in Fig. 5, were used for watermarking. In this study, Voloshynovskiy *et al.*'s block-based image watermarking approach [28] was chosen as the benchmark, denoted as Method I, due to its strong robustness and computational simplicity. Each watermark was embedded into an image block of size 32×32 so that the watermark's length is 1024 and the number of blocks is $|\Omega| = 256$. The incorporation of our CDW and Voloshynovskiy *et al.*'s scheme was denoted as Method II. Both methods were implemented in MATLAB. The advantage of using CDW will be manifested by comparing Methods I and II when WEA was imposed. However, we would like to particularly emphasize that the proposed CDW is readily applied to other watermarking algorithms. On the other hand, the Lee's Wiener filter [10] was used to perform denoising-based blind watermark extraction. The threshold T used to determine the existence of a watermark was selected as 0.12 if the false positive probability is desired to approximate 10^{-8} [2].

6.1 CDW Resistance to Collusion Attack

The collusion attack (by colluding $|\mathcal{C}| = |\Omega| = 256$ blocks) was applied to Method I and Method II, respectively, on ten cover images. The resistance of CDW to the collusion attack will be examined with respect to three scenarios: (s1) BER of estimated watermark's sign bits from an owner's perspective; (s2) quality of a colluded image, and (s3) watermark detection after performing collusion. All experimental results are shown in Figs. 6∼9, respectively.

As for (s1), Fig. 6(a) shows that when CDW is not involved in watermark embedding, most watermark bits are removed using collusion attack. However, once CDW is used, Fig. 6(b) shows that an owner can extract most watermark bits no matter collusion is conducted or not. This experiment confirms that the CDW is able to prevent more than half of watermark's sign bits from being correctly estimated by attackers. Namely, CDW neither obeys the rule of optimal watermark estimation (Eq. (6)) nor provides sufficient confidence (see Appendix) about watermark's sign estimation.

As for (s2), it can be found in Fig. 7(a) that collusion demonstrates its capability in improving the qualities of colluded images in terms of MSE. However, CDW can force collusion undesirably degrade the qualities of colluded images, as shown in Fig. 7(b). Two examples of colluded images are demonstrated in Fig. 8 for visible inspections. It is obvious that perceptual defects appear in the colluded images produced from the watermarking method by embedding the CDW. This experiment demonstrates that the CDW is efficient in prohibiting the collusion attack from achieving perfect cover data recovery (Eq. (8)).

As for (s3), watermark detection from colluded images are demonstrated in Fig. 9. There were 256 correlations resulted in an image. Only the minimum and the maximum correlations are plotted for each image. Fig. 9(a) shows that when CDW is not employed almost all watermarks cannot be extracted from colluded blocks. On the contrary, once CDW is involved in embedding, Fig. 9(b) shows that watermarks can be detected from most image blocks. This experiment verifies the resistance of CDW to collusion attack (Proposition 2). On the other hand, if the watermarks extracted (using Method II) from all image blocks are integrated to reconstruct the hidden watermark, we can find that the integrated watermark is very close to its origin in terms of correlation. We exemplify the case of $|\mathcal{C}| = 1$ to demonstrate the integrated result, as shown with the "diamond curve" in Fig. 9(b). Apparently, the integrated results show high robustness. This explains why some work proposes to embed multiple watermarks to enhance robustness [19].

In summary, as long as media hash is involved in constructing the watermark, the qualities of colluded images will be degraded but watermarks still can be extracted. Therefore, the merits of CDW

in helping block-based watermarking methods to resist collusion has been thoroughly demonstrated.

6.2 CDW Resistance to Copy Attack

The copy attack was conducted on Method I and Method II to compare their capability of resistance. One of the ten images was first watermarked and then the watermark was estimated and copied to the other nine unwatermarked images to form nine counterfeit stego images. By repeating the above procedure, there was in total 90 counterfeit stego images obtained. The PSNR values of the 90 attacked images are in the range of $26 \sim 36$dB (no perceptual masking is particularly considered). Our another goal is to verify whether strong watermark's energy can defeat the CDW. The 90 correlation values obtained after employing copy attack on Method I fall within the interval $[0.474 \quad 0.740]$ (all are sufficiently larger than $T = 0.12$), which indicates the presence of watermarks. However, when CDW is introduced these correlations are significantly decreased into the interval of $[-0.090 \quad 0.064]$, which indicates the absence of watermarks. The experimental results are consistent with the analytic result, derived in Eq. (24). Obviously, the proposed CDW is able to deter the detection of copied watermarks.

7 Conclusions and Future Work

Although multiple watermarks can be embedded into an image to withstand geometrical distortions, they are vulnerable to the collusion and copy attacks such that the desired functionality is lost. To cope with this problem, anti-disclosure watermark with resistance to watermark-estimation attack (WEA) has been investigated in this paper. Notably, we point out that both accurate estimation of watermark's sign and complete subtraction of watermark's energy constitute the sufficient and necessary condition in achieving effective watermark removal. As a consequence, we introduce the concept of media hash and combine it with the hidden information to create the so-called content-dependent watermark (CDW) to prevent from unauthorized watermark estimation. Potential characteristics of CDW have been analyzed to justify its resistance to WEA. Overall, experimental results have confirmed our mathematical analyses of WEA and CDW.

Moreover, it is straightforward to extend our content-dependent watermark to other multiple watermark embedding techniques or other media watermarking methods. In fact, the content-dependent watermark also has been successfully applied to video watermarking, wherein experimental results (can be found in http://www.iis.sinica.edu.tw/~lcs/ADVW) are similar to those obtained in image

watermarking. To our knowledge, the proposed anti-disclosure content-dependent watermark is the first to enable both resistance to the collusion and the copy attacks. At present, the proposed image hashing is variant to geometric distortions and thus potentially affects the resistance of the CDW to them. Our future work will continue to study the geometrical invariance of media hash [17].

Appendix: Confidence of Watermark's Sign Estimation under Collusion Attack

We verify how we are confident about the collusive estimation of watermark's sign using binomial probability distribution. Suppose each $sgn(W_b^e(i)), b \in \mathcal{C}$ is regarded as a trial and the trials are independent. This kind of trial will result in one of two outcomes: $+1$ and -1. Each outcome occurs with equal probability 0.5. Now, the confidence about the occurrence of $sgn(W_b^e(i))$ is formulated as the probability P_s of $sgn(W_b^e(i))$ mostly observed in $|\mathcal{C}|$ samples, $\Delta = \{sgn(W_1^e(i)), sgn(W_2^e(i)), ..., sgn(W_{|\mathcal{C}|}^e(i))\}$. Let β be the random variable denoting the number of $sgn(W_b^e(i))$ observed in Δ. As a consequence, P_s can be expressed as

$$P_s(\beta > \frac{|\mathcal{C}|}{2}) = \sum_{n=\frac{|\mathcal{C}|}{2}+1}^{|\mathcal{C}|} \left(\begin{array}{c} |\mathcal{C}| \\ n \end{array} \right) 0.5^{|\mathcal{C}|}. \tag{25}$$

By looking at the table of binomial probabilities, (pp. 685-688 of [16]), we can find that P_s will increase rapidly (usually it is larger than 0.8) as long as n is slightly larger than $\frac{|\mathcal{C}|}{2}$. The large the P_s is, the more confident we are. Hence, we can conclude that we are sufficiently confident to rely on the collusion attack in determining the sign of a hidden watermark.

Acknowledgment: This paper was supported by the National Science Council under NSC grants 91-2213-E-001-037 and 92-2422-H-001-004.

References

[1] P. Bas, J. M. Chassery, and B. Macq, "Geometrically Invariant Watermarking Using Feature Points," *IEEE Trans. on Image Processing*, Vol. 11, No. 9, pp. 1014 1028, 2002.

[2] I. J. Cox, M. L. Miller, and J. A. Bloom, "Digital Watermarking," *Morgan Kaufmann*, 2002.

[3] J. Fridrich, "Visual Hash for Oblivious Watermarking," *Proc. SPIE: Security and Watermarking of Multimedia Contents II*, Vol. 3971, 2000.

[4] IEEE Int. Workshop on Multimedia Signal Processing, special session on Media Recognition, 2002.

[5] *IEEE Journal on Selected Areas in Communications: special issue on Copyright and Privacy Protection*, Vol. 16, No. 4, 1998.

[6] *IEEE Trans. on Signal Processing: special issue on Signal Processing for Data Hiding in Digital Media and Secure Content Delivery*, Vol. 51, No. 4, 2003.

[7] T. Kalker, G. Depovere, J. Haitsma, and M. Maes, "A Video Watermarking System for Broadcast Monitoring," *Proc. of the SPIE*, Vol. 3657, pp. 103-112, 1999.

[8] D. Kirovski, H.S. Malvar, and Y. Yacobi, "A Dual Watermarking and Fingerprinting System," *Technical Report No. MSR-TR-2001-57*, Microsoft Research, 2001.

[9] M. Kutter, S. Voloshynovskiy, and A. Herrigel, "The Watermark Copy Attack", *Proc. SPIE: Security and Watermarking of Multimedia Contents II*, Vol. 3971, 2000.

[10] J. S. Lee, "Digital Image Enhancement and Noise Filtering by Use of Local Statistics", *IEEE Trans. on Pattern Anal. and Machine Intell.*, Vol. 2, No. 2, pp. 165-168, 1980.

[11] C.Y. Lin and S. F. Chang, "A Robust Image Authentication Method Distinguishing JPEG Compression from Malicious Manipulation," *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 11, No. 2, pp. 153-168, 2001.

[12] C. S. Lu, S. K. Huang, C. J. Sze, and H. Y. Mark Liao, "Cocktail Watermarking for Digital Image Protection", *IEEE Trans. on Multimedia*, Vol. 2, No. 4, pp. 209-224, 2000.

[13] C. S. Lu, H. Y. Mark Liao, and M. Kutter, "Denoising and Copy Attacks Resilient Watermarking by Exploiting Knowledge at Detector", *IEEE Trans. on Image Processing*, Vol. 11, No. 3, pp. 280-292, 2002.

[14] C. S. Lu and H. Y. Mark Liao, "Structural Digital Signature for Image Authentication: An Incidental Distortion Resistant Scheme", *IEEE Trans. on Multimedia*, Vol. 5, No. 2, 2003.

[15] A. R. Manuel and P. G. Fernando, "Analysis of Pilot-based Synchronization Algorithm for Watermarking of Still Images," *Signal Processing: Image Communications*, Vol. 17, pp. 611-633, 2002.

[16] W. Mendenhall, R. L. Scheaffer, and D. D. Wackerly, "Mathematical Statistica with Applications,", Duxbury Press, Boston, 1986.

[17] M. K. Mihcak and R. Venkatesan, "New Iterative Geometric Methods for Robust Perceptual Image Hashing," *Proc. Digital Right Management*, LNCS 2320, pp. 13-21, Springer-Verlag, 2002.

[18] *Proceedings of the IEEE*, Vol. 87, No. 7, 1999.

[19] M. Ramkumar and A. N. Akansu, "A Robust Scheme for Oblivious Detection of Watermarks/Data Hiding in Still Images", *Proc. SPIE Multimedia Systems and Applications*, Vol. 3528, pp. 474-481, 1998.

[20] *Signal Processing: special issue on Information Theoretic Aspects of Digital Watermarking*, Vol. 81, No. 6, 2001.

[21] K. Su, D. Kundur, D. Hatzinakos, "A Content-Dependent Spatially Localized Video Watermark for Resistance to Collusion and Interpolation Attacks," *Proc. IEEE Int. Conf. on Image Processing*, 2001.

[22] M. D. Swanson, B. Zhu, and A. H. Tewfik, "Multiresolution Scene-Based Video Watermarking Using Perceptual Models," *IEEE Journal on Selected Areas in Communications*, Vol. 16, No. 4, pp. 540-550, 1998.

[23] C. W. Tang and H. M. Hang, "A Feature-based Robust Digital Watermarking Scheme," *IEEE Trans. on Signal Processing*, Vol. 51, No. 4, pp. 950-959, 2003.

[24] W. Trappe, M. Wu, J. Wang, and K. J. Ray Liu, "Anti-collusion Fingerprinting for Multimedia", *IEEE Trans. on Signal Processing*, Vol. 51, No. 4, pp. 1069-1087, 2003.

[25] R. Venkatesan, S. M. Koon, M. H. Jakubowski, and P. Moulin, "Robust Image Hashing," *Proc. IEEE Int. Conf. Image Processing*, 2000.

[26] S. Voloshynovskiy, S. Pereira, A. Herrigel, N. Baumgartner, and T. Pun, "Generalized Watermarking Attack Based on Watermark Estimation and Perceptual Remodulation", *Proc. SPIE: Security and Watermarking of Multimedia Contents II*, Vol. 3971, San Jose, CA, USA, 2000.

[27] S. Voloshynovskiy, S. Pereira, V. Iquise, and T. Pun, "Attack Modelling: Towards a Second Generation Watermarking Benchmark," *Signal Processing*, Vol. 81, No. 6, pp. 1177-1214, 2001.

[28] S. Voloshynovskiy, F. Deguillaume, and T. Pun, "Multibit Digital Watermarking Robust against Local Nonlinear Geometrical Distortions," *Proc. IEEE Int. Conf. on Image Processing*, pp. 999 1002, 2001.

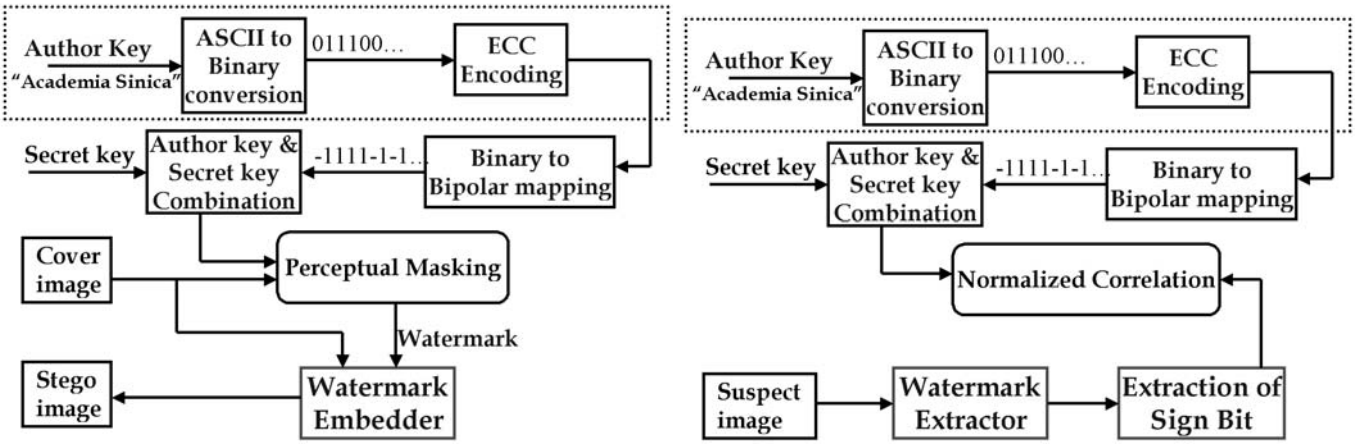


Figure 1: A general framework of digital watermarking system: watermark embedding (left) and watermark extraction (right).

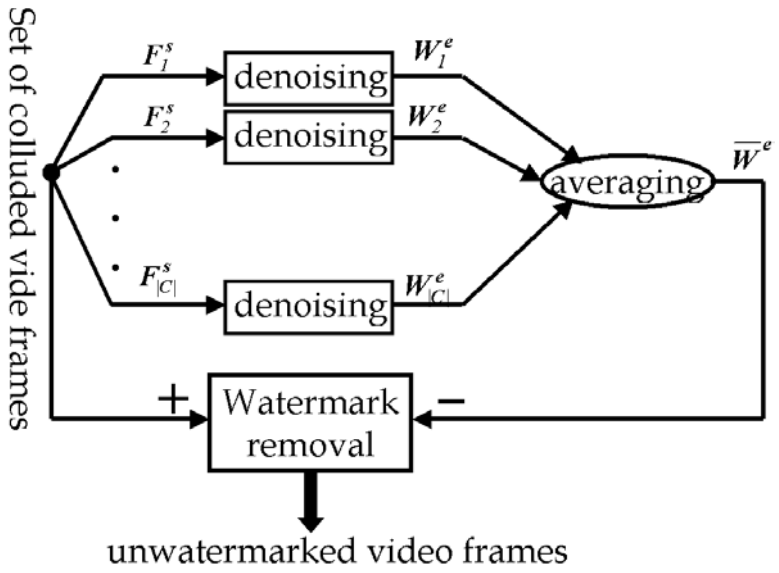


Figure 2: Collusion attack on video frames: (i) all video frames are embedded with the same watermark to form stego frames; (ii) denoising is used to estimate the hidden watermarks for those visually dissimilar frames; (iii) the final watermark is determined by averaging those watermarks obtained in (ii); and then subtracted from stego frames to construct unwatermarked frames.

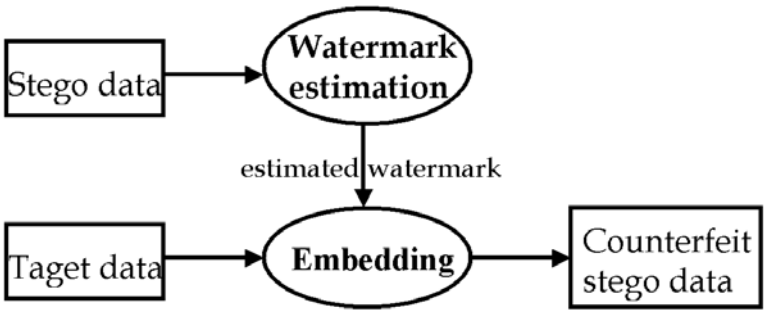


Figure 3: Copy attack: watermark estimation can be accomplished by denoising without needing to know any secret key.

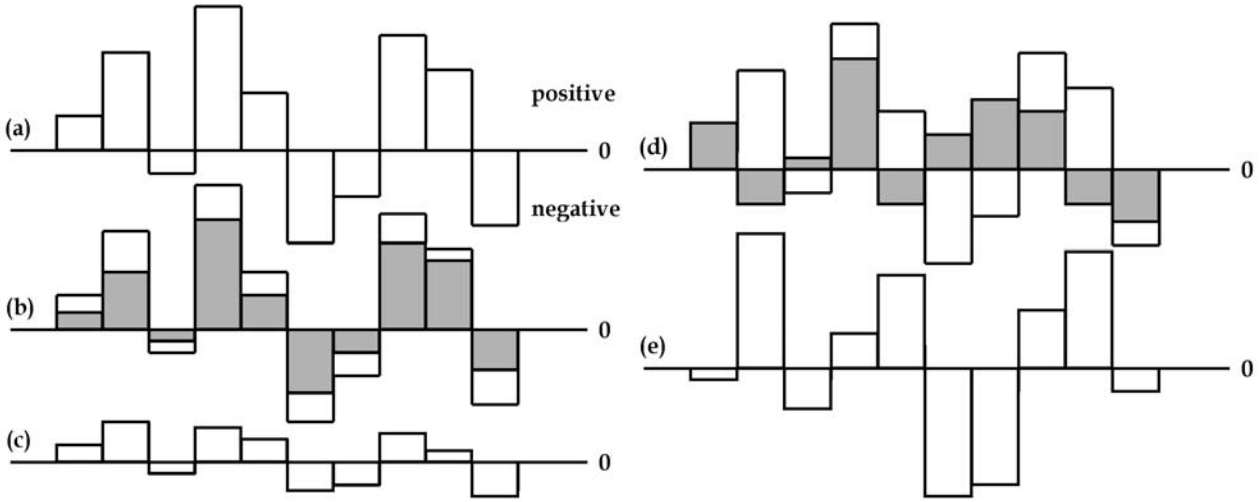
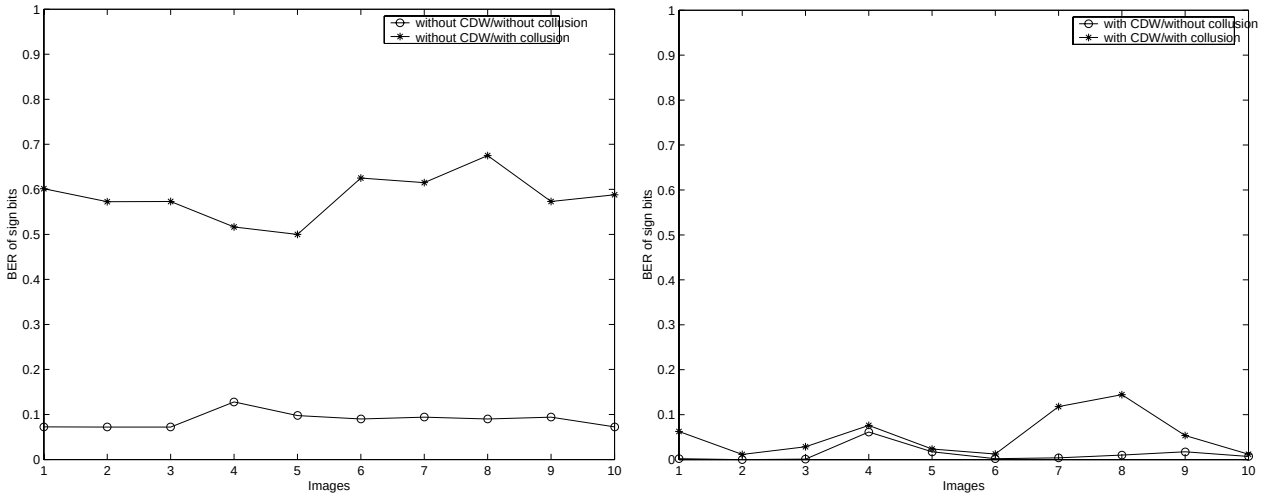


Figure 4: Watermark estimation/removal illustrated with energy variations: (a) original embedded watermark with each white bar describing the energy (determined using perceptual masking) of each watermark bit; (b) gray bars show the energies of an estimated watermark with all signs the same to the origin (a); (c) the residual watermark obtained after removing the estimated watermark (b); (d) the energies of an estimated watermark with most signs opposite to (a); (e) the residual watermark derived from (d). In the above examples, sufficiently large linear correlations (Eq. (2)) between (a) and (c), and (a) and (e), exist to indicate the presence of a watermark.

(I1) (I2) (I3) (I4) (I5)

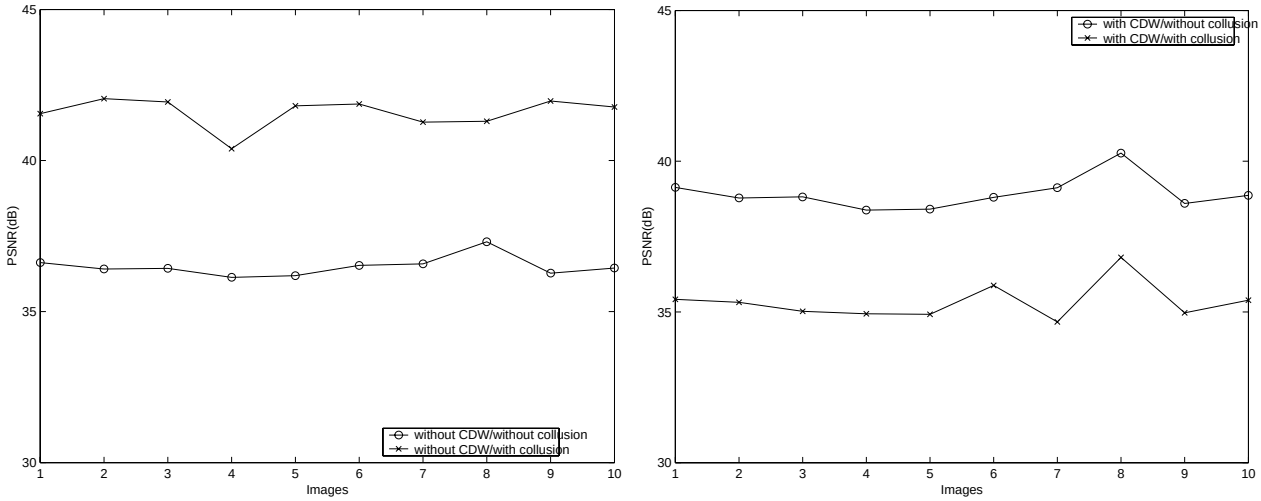
(I6) (I7) (I8) (I9) (I10)

Figure 5: Cover images.



(a) Method I (b) Method II

Figure 6: Scenario 1 (BER of estimated watermark's sign bits from an owner's perspective): (a) most watermark's sign bits are correctly estimated using collusion attack; (b) once CDW is introduced, watermark's sign bits mostly remain unchanged. This experiment confirms that CDW is efficient in randomizing watermarks to disable collusion.

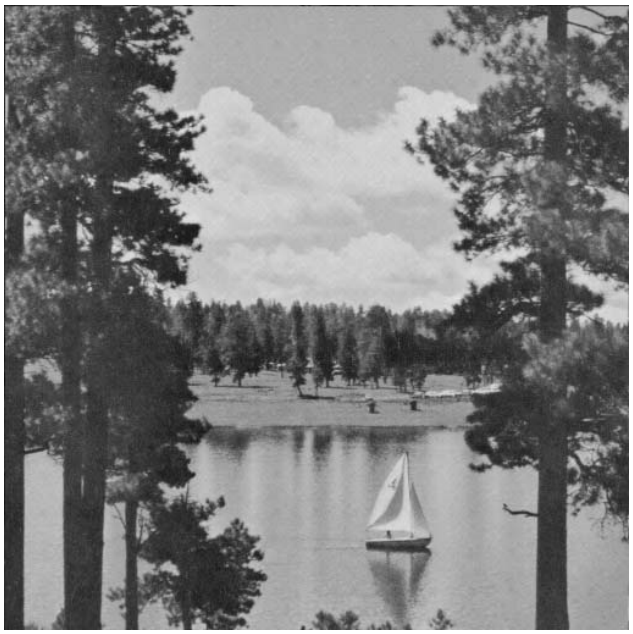


(a) Method I (b) Method II

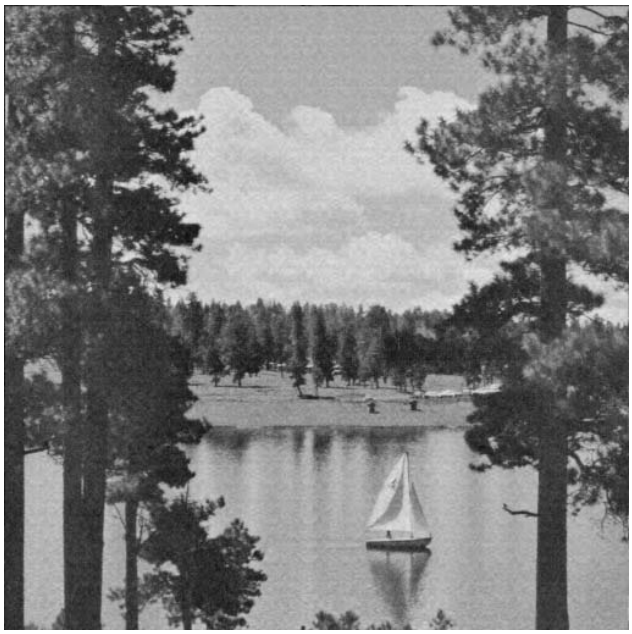
Figure 7: Scenario 2 (quality of a colluded image): (a) PSNR values of colluded images are higher than those of stego images; (b) when CDW is applied, PSNR values of colluded images are lower than those of stego images. This experiment reveals that collusion attack fails to improve fidelity of a colluded image when CDW is involved.

(a) colluded Lenna (Method I)

(b) colluded Lenna (Method II)

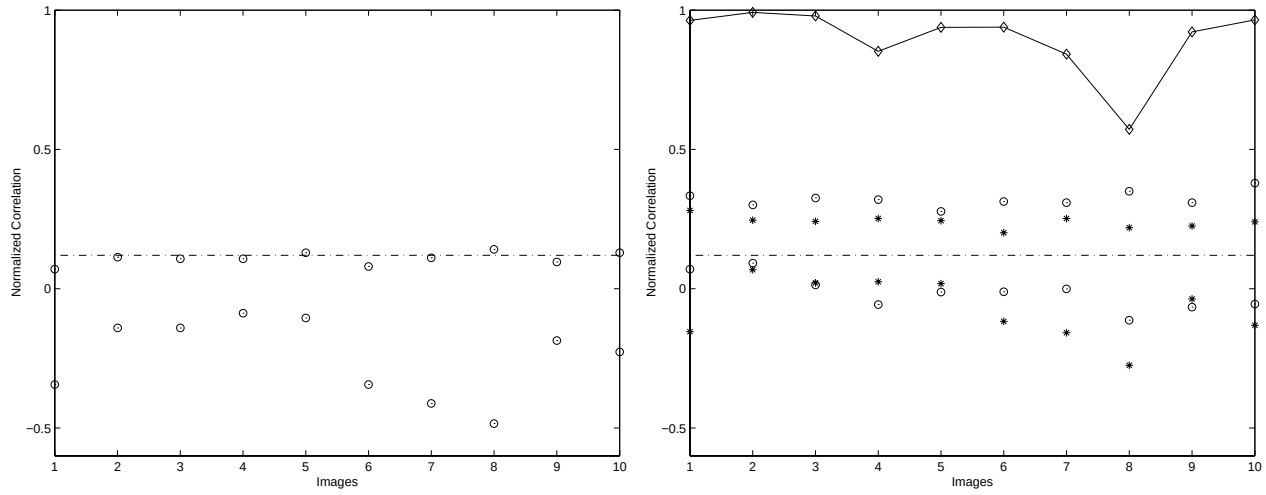


(c) colluded Sailboat (Method I)



(d) colluded Sailboat (Method II)

Figure 8: Perceptual illustrations of colluded images obtained from Method I (without using CDW) and Method II (using CDW). By comparing these two examples, it can be found that when collusion attack is encountered the CDW is able to make the colluded image noisy perceptually.



(a) Method I (b) Method II

Figure 9: Scenario 3 (watermark detection under collusion): (a) without using CDW, normalized correlations nearly show the absence of hidden watermarks; (b) using CDW, normalized correlations mostly show the presence of hidden watermarks. In (b), 'o' denotes the results obtained from colluding all blocks ($|\mathcal{C}| = 256 = |\Omega|$) while '*' denotes those obtained from colluding only one block ($|\mathcal{C}| = 1$). The dashdot line indicates the threshold $T = 0.12$. Definitely, the result of (b) has verified Proposition 2. Furthermore, when the watermark extracted from all image blocks are integrated to determine the final watermark, Method II produces most normalized correlations as high as 0.9 (as illustrated with the diamond curve in (b)) while Method I produces normalized correlations near 0.