Content-Dependent Anti-disclosure Image Watermark

Chun-Shien Lu[*] and Chao-Yong Hsu

Institute of Information Science, Academia Sinica,
Taipei, Taiwan 115, Republic of China
{lcs, cyhsu}@iis.sinica.edu.tw

Abstract. 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 exist watermark-estimation attacks (WEAs), such as the collusion attack and copy attack that are clever at disclosing the hidden watermark for unauthorized purposes while maintaining media's quality. The aim of this study was to deal with the WEA problem. We begin by gaining insight into WEA, leading to formal definitions of optimal watermark estimation and perfect cover data recovery. Subject to these definitions, the content-dependent watermark (CDW) is proposed to resist watermark-estimation attacks. The key is to introduce a media hash as a constituent component of the CDW. Mathematical analyses and experiment results consistently verify the effectiveness of the content-dependent watermarking scheme. To our knowledge, this anti-disclosure watermarking is the first work that takes resistance to both collusion and copy attacks into consideration.

1 Introduction

Digital watermarking is the technology of embedding a piece of information into the cover media data to carry out a specific mission. However, no matter what kinds of missions are considered, robustness is the critical issue affecting the practicability of a watermarking system. Robustness refers to the capability of resistance to attacks that are used to destroy, remove, or disable watermark detection. As previously discussed in [15], attacks can be classified into four categories: (1) removal attacks; (2) geometrical attacks; (3) cryptographic attacks; and (4) protocol attacks. The robustness of current watermarking methods has been examined with respect to either removal attacks or geometrical attacks or both. In particular, removal attacks contain 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 successful removal of the hidden watermarks, the media quality will inevitably be reduced. However, there indeed exists a kind of attacks that can defeat a watermarking

[*] Corresponding author

T. Kalker et al. (Eds.): IWDW 2003, LNCS 2939, pp. 61–76, 2004.

system without certainly sacrificing media quality. Among currently known attacks [15], the collusion attack [10,11], which is a removal attack, and the copy attack [6], which is a protocol attack, are typical examples of attacks that can achieve the aforementioned goal. The common step used to realize a collusion or copy attack is watermark estimation. Consequently, we call both the collusion and copy attacks watermark-estimation attacks (WEAs).

The aim of the collusion attack is to collect and analyze a set of watermarked media data[1] so that unwatermarked copies can be constructed to create the false negative problem. A 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 the frames. This scenario was first addressed in [11]. However, we argue that the collusion attack is not exclusively applied to video watermarking. In the literature, image watermarking with resistance to geometrical attacks has received much attention because even a slight geometrical modification may disorder the hidden watermark bits and disable watermark detection. In view of this, some researches [1,12,16] inserted multiple redundant watermarks into an image in the hope that robustness can be maintained as long as at least one watermark exists. Commonly, various kinds of image units, such as blocks [16], meshes [1], or disks [12], 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; in this way, collusion attacks can be equally applied to those image watermarking methods that employ a multiple redundant watermark embedding strategy.

In contrast to the collusion attack, the copy attack [6] has been developed to create the false positive problem; i.e., one can successfully detect a watermark from an unwatermarked image. Compared with the collusion atatck, the copy attack can be executed on only one media data; thus, it is more flexible. We will also show that the copy attack is rather easier to carry out than the denoising attack (a special case of the collusion attack). Based on the aforementioned reasons, the copy attack must be taken into consideration when the robustness of a watermarking system is to be evaluated.

In this paper, we propose a new scheme to cope with the watermark-estimation attacks (WEAs). After introducing a general watermarking framework in Sec. 2, the WEA will be thoroughly explored in Sec. 3. We will begin by investigating the achievable performance of the denoising attack and the copy attack to show that the copy attack is, in fact, easier to carry out. Then, we analyze to know that both accurate estimation of a watermark's sign and complete subtraction of a watermark's energy are indispensable for achieving effective watermark removal. On the other hand, they also serve as clues to breaking WEA. In order to withstand WEA, we propose the concept of content-dependent watermark (CDW), which is composed of an informative watermark that carries

[1] This set of watermarked media data in fingerprinting [13] 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.

information about an owner and a media hash that represents the cover carrier. The design of the media hash will be addressed in Sec. 4. Based on the presented media hash, in Sec. 5, CDW will be constructed and its properties examined. Furthermore, the validity of resistance to a collusion attack or copy attack using CDW will be analyzed. Finally, experiments and concluding remarks will be summarized in Sec. 6 and Sec. 7, respectively.

2 Basic Framework of Digital Watermarking

A general digital watermarking scheme is described as follow. In the embedding process, a watermark is a message (author key) that is 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 ECC encoded sequence $\mathbf{S}$ are mapped from $\{0\ \ 1\}$ to $\{-1\ \ 1\}$ such that $\mathbf{S}$ is a Gaussian distribution with zero mean and unit variance. $\mathbf{S}$ is also 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 $\mathbf{I^s}$ by

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

where L denotes the length of $\mathbf{S}$ and $\mathbf{M_I}$ stands for the masking matrix derived from $\mathbf{I}$. We call the finally embedded sequence $\mathbf{S} \cdot \mathbf{M_I}$ as the watermark $\mathbf{W}$. The watermark $\mathbf{W}$ is assumed to be a Gaussian sequence with zero mean and variance ρ^2. Notice that the variance ρ^2 from $\mathbf{M_I}$ determines the watermark's energy. In addition, $\mathbf{S}$ determines the watermark's sign and is secured by the secret key K.

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

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

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 [5,6,14]. It is said that a watermark exists provided that the normalized correlation δ_{nc} between $\mathbf{S}$ and $\mathbf{S^e}$ (with equal energy $\sqrt{L}$) is larger than a threshold T, where

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

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

3 Watermark Estimation Attack

Basically, removal attacks try to vanish the hidden watermark by manipulating a stego image $\mathbf{I^s}$ so that the quality of the attacked image $\mathbf{I^a}$ is further destroyed. Specifically, $PSNR(\mathbf{I}, \mathbf{I^s}) \geq PSNR(\mathbf{I}, \mathbf{I^a})$ always holds. However, a more clever removal attack can achieve $PSNR(\mathbf{I}, \mathbf{I^s}) \leq PSNR(\mathbf{I}, \mathbf{I^a})$. The collusion attack is a typical example of an attack that follows the above scenario. Usually, a collusion attack is applied to video watermarking by averaging of a set of estimated watermarks to obtain the hidden watermark. As for image watermarking, some recent works have been proposed embedding multiple redundant watermarks into local areas [1,12,16] so that global/local geometrical distortions can be resisted. Provided we treat a local region in an image similar to a video frame in a video, then collusion attack can also be applied to region-based image watermarking to create the false negative problem. It should be noted that the conventional denoising-based removal attack [14], which is only applied to a single image, is a special case of the collusion attack.

On the other hand, an estimated watermark can be inserted into unwatermarked media data to produce a counterfeit stego data. This is the so-called copy attack [6], which has been developed to create the false positive problem; i.e., one can successfully detect a watermark from an unwatermarked data. As classified in [15], copy attack belongs to a type of protocol attacks. The copy attack 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 stego image; and (iii) from the counterfeit image, a watermark can be detected that wrongly claims rightful ownership.

To our knowledge, the collusion attack and the copy attack have not been simultaneously taken into consideration when investigating the robustness issue. Owing to watermark estimation is the first step in both attacks, these are called, watermark-estimation attacks (WEAs).

3.1 Analysis of the Achievable Performance of the Denoising Attack and Copy Attack

Two typical examples of watermark-estimation attacks, i.e., the denoising attack [14] (recall that it is a special case of the collusion attack) and the copy attack [6], will be discussed. Without loss of generality, suppose the decision on a watermark's existence will be based on the linear 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 estimated from $\mathbf{X^s}$.

Let $\mathbf{W}$ be a watermark 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 $\delta_{nc}(sgn(\mathbf{W}), sgn(\mathbf{W^a})) < T$ must hold. This result indicates that the ratios of the correctly (C_w) and wrongly (NC_w) decoded watermark bits should, respectively, satisfy

$$C_w \leq \frac{1+T}{2} \text{ and } NC_w \geq \frac{1-T}{2}, \tag{3}$$

where $C_w + NC_w = 1$ and NC_w corresponds to the bit error rate (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 we should set the threshold T 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 required to be destroyed 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 $\mathbf{W}$. The copy attack is claimed to succeed if $\delta_{nc}(sgn(\mathbf{W}), sgn(\mathbf{W^z})) \geq T$ holds. This implies that 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 a bit larger than $\frac{T}{2}$ based on the fact that T is usually small. However, if the denoised results (i.e., more than one estimated watermark) are collected and colluded to generate an estimation that is closer to its original, then the collusion attack will exhibit more powerful performance than the denoising attack, as evidenced in [10,11].

3.2 Optimal Watermark Estimation and Perfect Cover Data Recovery

Mathematical 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, correction estimation of watermark's energy is closely related to the accuracy of removal attack. Several scenarios are shown in Fig. 1, which illustrates the variations of (a) an original watermark; (b) an estimated watermark (in gray-scale); and (c) a residual watermark generated by subtracting the estimated watermark from the original watermark. From Fig. 1, we can

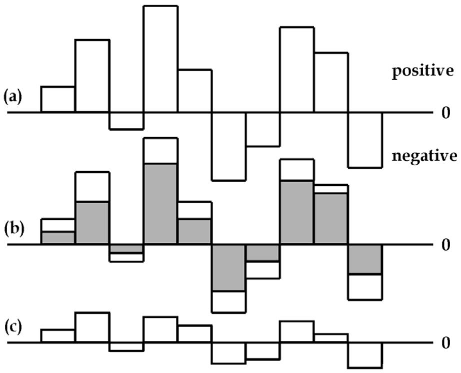


Fig. 1. Watermark estimation/removal illustrated by energy variations: (a) original embedded watermark with each white bar indicating the energy of each watermark bit; (b) gray bars show the energies of an estimated watermark; (c) the residual watermark obtained after removing the estimated watermark. A sufficiently large correlation (Eq. (2)) between (a) and (c) exists to indicate the presence of a watermark.

realize that if the energy of a hidden watermark cannot be completely removed, 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., $sgn(W(i)) \neq sgn(W^e(i))$), then any additional energy subtraction will not helpful in improving removal efficiency. On the contrary, watermark removal by energy subtraction operated in the opposite polarity will severely damage the media data's fidelity. Actually, this corresponds to adding a watermark with higher energy into cover data without satisfying the masking constraint. The importance of polarities of watermark bits has been previously emphasized in [8] by embedding two complementary watermarks that are modulated in different ways to resist different sets of attacks. With this understanding, we shall define "optimal watermark estimation" and "perfect cover data recovery," respectively, as follows for use in further analyses.

Definition 1 (Optimal Watermark Estimation): 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 for optimal estimation of $\mathbf{W}$ as $\mathbf{W}^e$ is defined as

$$\delta_{nc}(sgn(\mathbf{W}), sgn(\mathbf{W^e})) = 1, \tag{6}$$

where $sgn(\mathbf{v}) = \{sgn(v_1), sgn(v_2), ..., sgn(v_L)\}$ represents the signs of elements in a vector $\mathbf{v} = \{v_1, v_2, ..., v_L\}$. This is the first step, where a watermark may be undetected by an owner if more than $\frac{L(1+T)}{2}$ sign bits of the watermark can be obtained by attackers. Beyond this, however, to avoid leaving a residual watermark (as illustrated in Fig. 1(c)) that can reveal the hidden watermark, accurate estimation of the energy of $\mathbf{W^e}$ is absolutely indispensable. In addition to Eq. (6), watermark removal can be achieved only if the watermark energy to be subtracted is larger than or equal to the added energy, i.e., $mag(W^e(i)) \geq$

$mag(W(i))$, where $mag(t)$ denotes the magnitude $|t|$ of t. Therefore, the sufficient and necessary condition for complete watermark removal can be defined $\forall i$ as

$$mag(W^e(i)) \geq mag(W(i)), |mag(W^e(i)) - mag(W(i))| < JND(i), sgn(W^e(i)) = sgn(W(i)), \tag{7}$$

where $JND(i)$ denotes a masking threshold. After the optimal watermark estimation scheme defined in Eq. (7) is employed, the extracted watermark would behave like a random signal so that no trace of the watermark can be observed. **Definition 2 (Perfect Cover Data Recovery):** Under the prerequisite that Definition 1 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 if watermarks have been completely removed, the quality of the attacked images will be poor. Typically, evaluation of $mag(\mathbf{W^e})$ can be achieved either by means of averaging [11] or remodulation [14].

In summary, 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 undoubtedly hold. Unlike other watermark removal attacks that reduce the quality of the media data, the collusion attack may improve the quality of colluded data.

4 Image Hash

From the analyses of watermark-estimation attack (WEA) described in Sec. 3, we have found that the success of WEA mainly depends on the fact that the hidden watermark totally behaves like a noise, and can be easily and reliably obtained by means of a denoising process. In order to disguise this prior knowledge and hide it from attackers, a watermark must be designed to carry information relevant to the cover image itself. Meanwhile, the content-dependent information must be secured[2] by a secret key and be robust to digital processing [9] in order not to affect watermark detection. To this end, we shall introduce the concept of the image hash as a kind of content-dependent information used to create the so-called content-dependent watermark (CDW).

The image hash [3], also known as the "digital signature" [9] or "media fingerprint" [4], has been used in many applications, including content authentication, copy detection, and media recognition. In this paper, the proposed image hash extraction procedure is operated in the 8 × 8 block-DCT domain. For each block,

[2] 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 discussed in Sec. 5.

a piece of representative but robust information is created. It is defined as the magnitude relationship between two AC coefficients:

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

where $r(i)$ is a robust feature value in a sequence $\mathbf{r}$, and $f_i(p_1)$ and $f_i(p_2)$ are two AC coefficients at positions p_1 and p_2 in block i. The length of $\mathbf{r}$, $|\mathbf{r}|$, is equal to the number of blocks. The DC coefficient will not be selected because it is positive and, thus, not random. In addition, the two 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. We call this feature value $r(\cdot)$ robust because this magnitude relationship between $f_i(p_1)$ and $f_i(p_2)$ can be mostly preserved under incidental modifications. Please refer to [9] for similar robustness analyses. It should be noted that depending on different watermarking algorithms the proposed media hash extraction method can be adjusted correspondingly.

In practice, each media hash must be constructed within the range where one watermark is embedded so that resistance to geometrical distortions can still be preserved. Under this constraint, when the sequence $\mathbf{r}$ is extracted, it is repaired to form an image hash with $|\mathbf{r}| = L$. If $|\mathbf{r}| > |\mathbf{W}|$, then the extra elements at the tail of $\mathbf{r}$ are deleted; otherwise, $\mathbf{r}$ is cyclically appended. We call the finally created sequence media hash **MH**, which is a bipolar sequence. Next, media hash **MH** of an image is mixed with the watermark, **W**, to generate the content-dependent watermark (**CDW**) as

$$\mathbf{CDW} = S(\mathbf{W}, \mathbf{MH}), \tag{9}$$

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

5 Image-Dependent Watermark

The properties of the image-dependent watermark will be discussed first. Then, its resistance to WEA will be analyzed based on block-based image watermarking.

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 the block-based image watermarking scheme [1,12,16] 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.,

$$\mathbf{B^s}_i = \mathbf{B}_i + \mathbf{CDW}_i, \qquad \mathbf{I^s} = \oplus_{i \in \Omega} \mathbf{B^s}_i, \tag{10}$$

where $\mathbf{B^s}_i$ is a stego block and $\mathbf{CDW}_i$, similar to Eq. (9), is defined as the mixture of a fixed informative watermark $\mathbf{W}$ and a block-based hash $\mathbf{MH}_{\mathbf{B}_i}$, i.e.,

$$\mathbf{CDW}_i = S(\mathbf{W}, \mathbf{MH}_{\mathbf{B}_i}). \tag{11}$$

In Eq. (11), the mixing function $S(\cdot,\cdot)$ will be designed as a procedure of permuting the media hash $\mathbf{MH}_{\mathbf{B}_i}$ using the same secret key K, followed by shuffling the watermark to enhance security. Specifically, it is expressed as

$$S(\mathbf{W}, \mathbf{MH}_{\mathbf{B}_i})(k) = W(k)PT(\mathbf{MH}_{\mathbf{B}_i}, K)(k),$$

where PT denotes a permutation function controlled by the secret key K with the aim of achieving uncorrelated crosscorrelation,

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

and autocorrelation:

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

The proposed content-dependent watermark possesses the characteristics described as follows. They are useful for proving 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 $\mathbf{MH}_{\mathbf{B}_i}$ and $\mathbf{MH}_{\mathbf{B}_j}$, i.e.,

$$\delta_{nc}(\mathbf{B}_i, \mathbf{B}_j) = \delta_{nc}(\mathbf{MH}_{\mathbf{B}_i}, \mathbf{MH}_{\mathbf{B}_j}). \tag{12}$$

Accordingly, we have two extreme cases: (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 $\delta_{nc}(\mathbf{B}_i, \mathbf{B}_j) \approx 0$.

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. with Gaussian distributions $\mathcal{N}(0, \rho^2)$, the following properties can be established: (i) $\delta_{nc}(\mathbf{CDW}_i, \mathbf{CDW}_j)$ is linearly proportional to $\delta_{nc}(\mathbf{B}_i, \mathbf{B}_j)$; (ii) $\delta_{nc}(\mathbf{CDW}_i, \mathbf{CDW}_j) \leq \delta_{nc}(\mathbf{W}^2)$; (iii) $\delta_{nc}(\mathbf{n}, \mathbf{CDW}) = 0$ ($\mathbf{n}$ is generally a Gaussian noise with zero mean). Due to limits of space, proofs of Proposition 1 by exploiting the above properties will not be shown here.

5.2 Resistance to Collusion Attacks

By means of a collusion attack, the averaging operation is performed on stego blocks $\mathbf{B^s}_i$'s of a stego image $\mathbf{I^s}$. From an attacker's perspective, each hidden watermark has to be estimated by means of a denoising operation (e.g., Wiener filtering), so deviations of estimation will inevitably occur. Let $\mathbf{W^e}_i$ be an estimated watermark from $\mathbf{B^s}_i$. Without loss of generality, it is assumed to have zero mean. In fact, $\mathbf{W^e}_i$ can be modeled as a partial hidden watermark plus a noise component, i.e.,

$$\mathbf{W^e}_i = \alpha_i \mathbf{CDW}_i + \mathbf{n}_i, \tag{13}$$

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 ensure that the estimated watermark and the hidden watermark have the same energy. Under these circumstances, $1 \geq \alpha_i = \delta_{nc}(\mathbf{W^e}_i, \mathbf{CDW}_i) > T$ always holds based on the fact that a watermark is a high-frequency signal and can be efficiently estimated by means of denoising [5,6,14]. This factor α_i plays a crucial role in two ways: (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 [14]; (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

$$\bar{\mathbf{W}}^{\mathbf{e}} = \frac{\sqrt{|\mathcal{C}|}}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} \mathbf{W^e}_i = \frac{1}{\sqrt{|\mathcal{C}|}} \sum_{i \in \mathcal{C}} (\alpha_i \mathbf{CDW}_i + \mathbf{n}_i) \tag{14}$$

because $\mathbf{W^e}_i$'s are obtained from (nearly) visually dissimilar image blocks, which can be regarded as i.i.d. approximately.

Proposition 2 In a collusion atatck, 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 $\mathbf{I^s} = \oplus_{i \in \Omega} \mathbf{B^s}_i$ by

$$\mathbf{B^u}_i = \mathbf{B^s}_i - \bar{\mathbf{W}}^{\mathbf{e}}, \qquad \mathbf{I^u} = \oplus_{i \in \Omega} \mathbf{B^u}_i. \tag{15}$$

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: By making use of Eq. (14) and Proposition 1, we get:

$$\begin{aligned} \delta_{nc}(\bar{\mathbf{W}}^{\mathbf{e}}, \mathbf{CDW}_k) &= \frac{\sqrt{|\mathcal{C}|}}{|\mathcal{C}|} \delta_{nc}(\sum_{i \in \mathcal{C}} (\alpha_i \mathbf{CDW}_i + \mathbf{n}_i), \mathbf{CDW}_k) \\ &= \frac{1}{\sqrt{|\mathcal{C}|}} \sum_{i \in \mathcal{C}} \alpha_i \delta_{nc}(\mathbf{CDW}_i, \mathbf{CDW}_k) + \frac{1}{\sqrt{|\mathcal{C}|}} \sum_{i \in \mathcal{C}} \delta_{nc}(\mathbf{n}_i, \mathbf{CDW}_k) \\ &= \frac{\alpha_k}{\sqrt{|\mathcal{C}|}}, \end{aligned} \tag{16}$$

where $\mathbf{CDW}_k$ represents the content-dependent watermark embedded in $\mathbf{B}_k$. According to Eq. (16), our derivations are further explained as follows: the first row is resulted from Eq. (14) 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 2 of Proposition 1 and Eqs. (15) and (16), we get:

$$\begin{aligned} \delta_{nc}(\mathbf{B^u}_k, \mathbf{CDW}_k) > T \text{ iff } & \delta_{nc}(\mathbf{B}_k + \mathbf{CDW}_k - \bar{\mathbf{W}}^{\mathbf{e}}, \mathbf{CDW}_k) > T \\ \text{iff } & \delta_{nc}(\mathbf{CDW}_k, \mathbf{CDW}_k) - \delta_{nc}(\bar{\mathbf{W}}^{\mathbf{e}}, \mathbf{CDW}_k) > T \\ \text{iff } & \delta_{nc}(\bar{\mathbf{W}}^{\mathbf{e}}, \mathbf{CDW}_k) = \frac{\alpha_k}{\sqrt{|\mathcal{C}|}} < 1 - T. \end{aligned} \tag{17}$$

Remarks (Further interpretation of $|\mathcal{C}|$): If $|\mathcal{C}| = 1$ (we mean that the collusion attack is only applied to one block), then the collusion attack degenerates into a denoising-based removal attack. Under this circumstance, the success of the 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. (17) 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 success of the collusion attack when $|\mathcal{C}| = 1$. This result totally depends on the effectiveness of denoising in estimating an added signal. Provided that $|\mathcal{C}|$ becomes infinite, i.e., $|\mathcal{C}| = |\Omega| \to \infty$, $\delta_{nc}(\bar{\mathbf{W}}^{\mathbf{e}}, \mathbf{CDW}_k) \to 0$ is obtained such that T can be an 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, which is by definition applied to more than one image block. It is interesting to note that this result contradicts the expected characteristic of a collusion attack. In particular, the performance degradation of the proposed method can be interpreted as being 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 will proceed to show why the presented content-dependent watermark can be immune to a 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 block $\mathbf{Z^s}$ will be

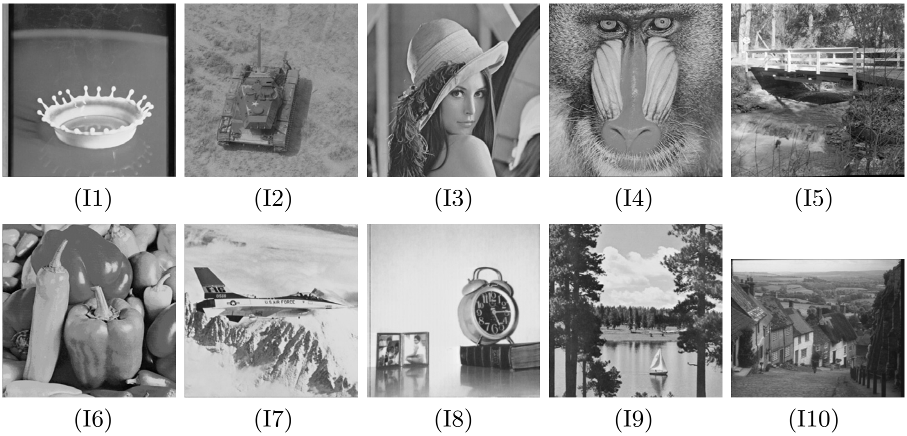
$$\mathbf{W^z} = (\alpha \times \mathbf{CDW_X} + \mathbf{n}), \tag{18}$$

according to Eq. (13), where $\mathbf{n}$ indicates the noise sequence (which is irrelevant to watermarks) generated by means of denoising $\mathbf{Z^s}$. Based on the evidence that denoising is an efficient way to estimate watermarks [6,14,16], $||\alpha\mathbf{CDW_X}||_2 > ||\mathbf{n}||_2$ can undoubtedly hold, with $||\cdot||_2$ being the energy. Given Eqs. (11) and (18), 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:

$$\begin{aligned}\delta_{nc}(\mathbf{CDW_Z}, \mathbf{W^z}) &= \frac{1}{|\mathbf{W}|\rho^2}\sum_{i=1}^{|\mathbf{W}|} CDW_Z(i)W^z(i)\\ &\approx \frac{\alpha}{|\mathbf{W}|\rho^2}\sum_{i=1}^{|\mathbf{W}|} CDW_Z(i)CDW_X(i) \approx 0.\end{aligned} \tag{19}$$

6 Experimental Results

In our experiments, ten varieties of gray-scale cover images of size 512 × 512, as shown in Fig. 2, were used for watermarking. In this study, Voloshynovskiy *et al.*'s block-based image watermarking approach [16] was chosen as the benchmark, denoted as Method I, due to its strong robustness and computational simplicity. Each image block is of size 32 × 32 so that the watermark's length was 1024 and the number of blocks was $|\Omega| = 256$. The combination of our CDW and Voloshynovskiy *et al.*'s scheme is denoted as Method II. We would like to manifest the advantage of using CDW by comparing the results obtained using Methods I and II when WEA is imposed. However, we would like to particularly emphasize that the proposed CDW can be readily applied to other watermarking algorithms. On the other hand, Lee's Wiener filter [7] was used to perform denoising-based blind watermark extraction.



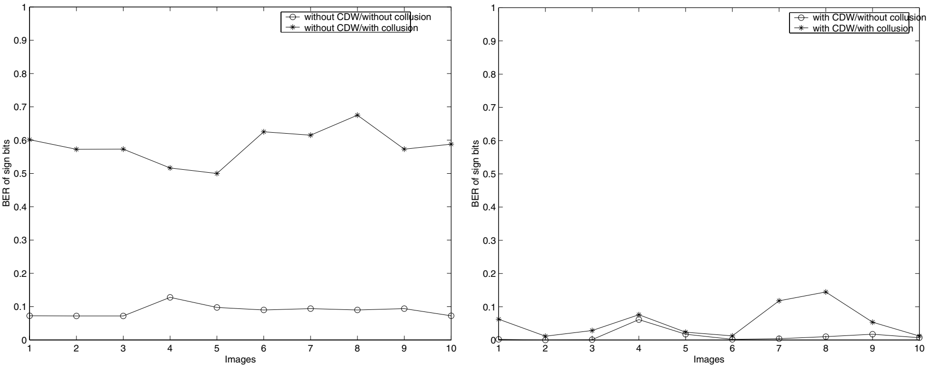
(I1) (I2) (I3) (I4) (I5)

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

Fig. 2. Cover images.

6.1 CDW Resistance to Collusion Attack

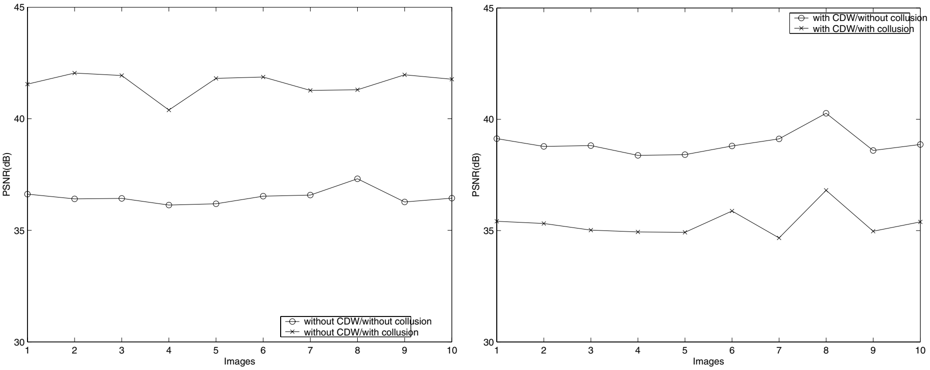
The collusion attack (operated by colluding $|\mathcal{C}| = |\Omega| = 256$ blocks) was applied to Method I and Method II, respectively, on ten cover images. The impacts of collusion attack and CDW will be examined with respect to the three scenarios: (s1) the BER of the estimated watermark's sign bits from an owner's perspective; (s2) the quality of a colluded image; and (s3) watermark detection after collusion. For (s3), there are 256 correlations resulted in an image. Only the minimum and the maximum correlations are plotted for each image. All the numerical results are depicts in Figs. $3\sim5$, respectively. Some colluded images are illustrated in Fig. 6 for visual inspection. In summary, as long as an image hash is involved in constructing a watermark, the quality of the colluded images will be degraded,



(a) Method I

(b) Method II

Fig. 3. Scenario 1 (BER of the estimated watermark's sign bits): (a) most of the watermark's sign bits are correctly estimated by a collusion attack; (b) when CDW is introduced, the watermark's sign bits mostly remain unchanged. This experiment confirms that CDW is efficient in randomizing watermarks in order to disable collusion.



(a) Method I

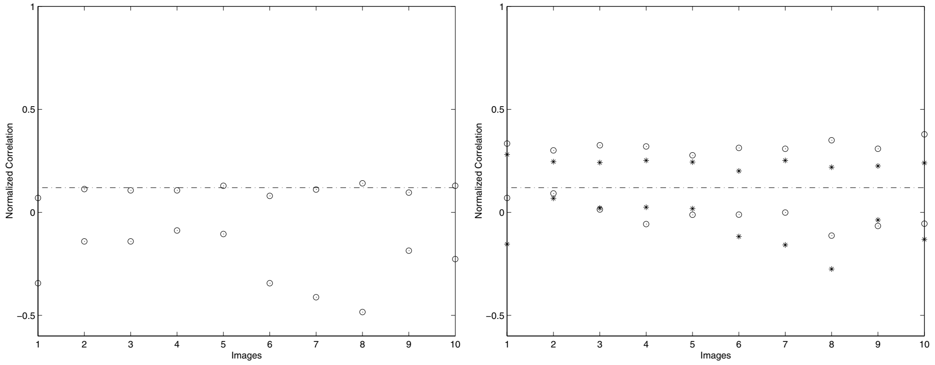
(b) Method II

Fig. 4. Scenario 2 (quality of a colluded image): (a) the PSNR values of the colluded images are higher than those of the stego images; (b) when CDW is applied, the PSNR values of the colluded images are lower than those of the stego images. This experiment reveals that a collusion attack will fail to improve the fidelity of a colluded image when CDW is involved.

but it will still be possible for the watermarks to be extracted. Therefore, the merits of CDW in resisting collusion have been thoroughly demonstrated.

6.2 CDW Resistance to the Copy Attack

The copy attack was applied to Method I and Method II to compare their capability of resistance. One of the ten images was watermarked, estimated, and copied to the other nine unwatermarked images to form nine counterfeit stego



(a) Method I

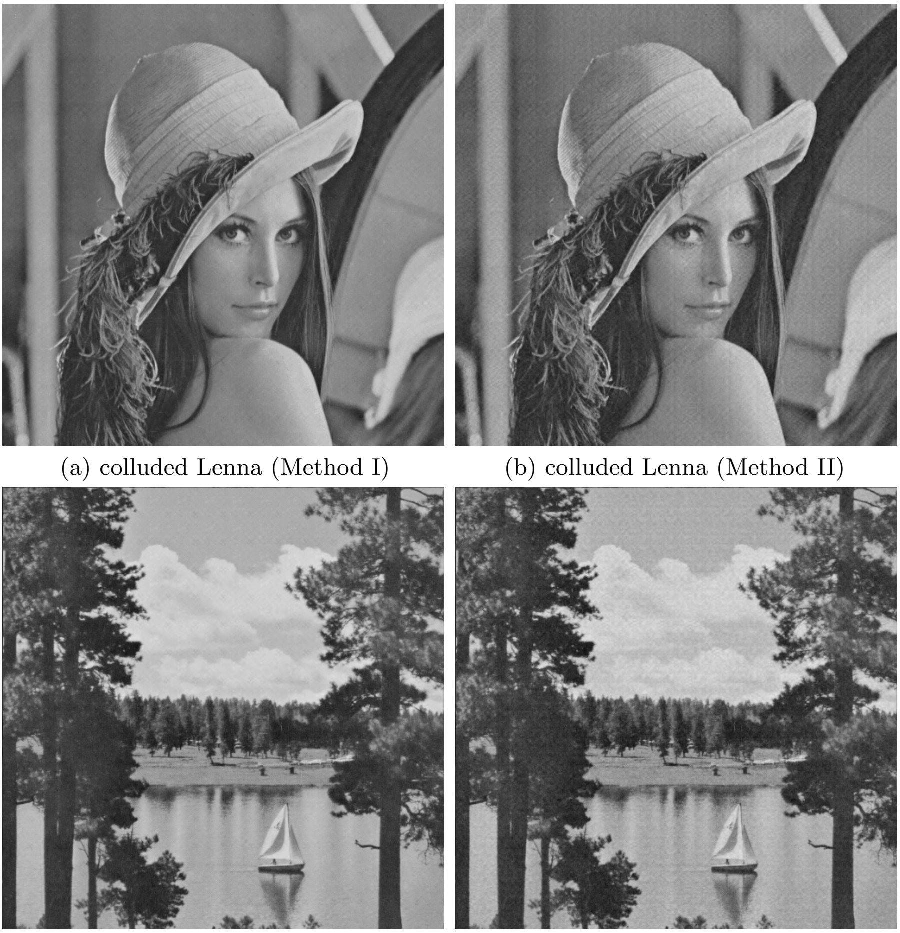
(b) Method II

Fig. 5. Scenario 3 (watermark detection after collusion): (a) without using CDW, normalized correlations show the almost complete absence of hidden watermarks; (b) using CDW, normalized correlations mostly show the presence of hidden watermarks. In (b), 'o' denotes the results obtained by colluding all blocks ($|\mathcal{C}| = 256 = |\Omega|$), while '*' denotes those obtained by colluding only one block ($|\mathcal{C}| = 1$). The dashdot line indicates the threshold $T = 0.12$. Definitely, the result of (b) verifies Proposition 2. Furthermore, when the watermarks extracted from all the image blocks are integrated (a kind of collusion estimation) to obtain the final watermark, Method II produces normalized correlations as high as 0.9, while Method I produces normalized correlations close to 0.

images. By repeating the above procedure, a total of 90 counterfeit stego images were obtained. The PSNR values of the 90 attacked images were in the range of $26 \sim 36$dB (no masking was used). The 90 correlation values obtained by applying the copy attack to Method I fell within the interval [0.474 0.740] (all were sufficiently larger than $T = 0.12$), which indicates the presence of watermarks. However, when CDW was introduced, these correlations decreased significantly to [−0.090 0.064], which indicates the absence of watermarks. The experimental results are consistent with the analytic result, derived in Eq. (19). Obviously, the proposed CDW is able to deter the detection of copied watermarks.

7 Concluding Remarks

Although multiple watermarks can be embedded into an image to withstand geometrical distortions, they are vulnerable to be colluded or copied, and the desired functionality is lost. To cope with this problem, an anti-disclosure watermark with resistance to watermark-estimation attack (WEA) has been investigated in this paper. We have pointed out that both accurate estimation of a watermark's sign and complete subtraction of a watermark's energy constitute the sufficient and necessary conditions to achieve complete watermark removal. We have introduced the concept of the media hash and combined it with hidden information to create the so-called content-dependent watermark (CDW).



(a) colluded Lenna (Method I)

(b) colluded Lenna (Method II)

(c) colluded Sailboat (Method I)

(d) colluded Sailboat (Method II)

Fig. 6. Perceptual illustrations of colluded images obtained using Method I (without using CDW) and Method II (using CDW). By comparing these two examples, it can be found that when a collusion attack is encountered, CDW is able to make the colluded image perceptually noisy.

The characteristics of CDW have been analyzed to justify its resistance to WEA. The experimental results have confirmed our mathematical analyses of WEA and CDW. Extensions of our content-dependent watermark to other multiple watermark embedding techniques or other media watermarking are straightforward. To our knowledge, the proposed content-dependent anti-disclosure watermark is the first to enable both resistance to the collusion and the copy attacks. Media hash with geometric-invariance is a worthy direction for further studying.

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, 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 (MMSP), special session on Media Recognition, Virgin Islands, USA, 2002.
5. 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.
6. M. Kutter, S. Voloshynovskiy, and A. Herrigel, "The Watermark Copy Attack", *Proc. SPIE: Security and Watermarking of Multimedia Contents II*, Vol. 3971, 2000.
7. J. S. Lee, "Digital Image Enhancement and Noise Filtering by Use of Local Statistics", *IEEE Trans. on Pattern Anal. and Machine Intell.*, Vol. 2, pp. 165–168, 1980.
8. 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, pp. 209–224, 2000.
9. 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, pp. 161–173, 2003.
10. 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.
11. 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.
12. 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.
13. W. Trappe, M. Wu, J. Wang, and K. J. Ray Liu, "Anti-collusion Fingerprinting for Multimedia", *IEEE Trans. on Signal Processing*, Vol. 51, pp. 1069–1087, 2003.
14. 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.
15. 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.
16. 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.