Resistance of Content-dependent Video Watermarking to Watermark-Estimation Attacks

Chun-Shien Lu[†*], Jan-Ru Chen[‡], and Kuo-Chin Fan[‡],

[†]Institute of Information Science, Academia Sinica
Taipei, Taiwan 115, ROC
[‡]Dept. of Computer Sci. and Info. Eng., National Central University
Chung-Li, Taiwan 320, ROC
[*]Email: lcs@iis.sinica.edu.tw

***Abstract*— *One of the key challenges for a watermarking scheme to be mandated in a digital right management (DRM) system is the robustness. This paper is focused on exploring the robustness against the watermark-estimation attacks (WEAs) that are clever at disclosing hidden information for unauthorized purposes without sacrificing media's quality. In WEAs, the collusion attack naturally occurs in video watermarking while the copy attack adapts to any media watermarking. In view of this, the aim of this study is to deal with the WEAs by means of a video frame-dependent watermark (VFDW). We begin by gaining insight into the WEAs, leading to formal definitions of "optimal watermark prediction" and "perfect cover data recovery." Subject to these definitions, the video-frame hash is addressed as a constituent component of the VFDW for anti-estimation of hidden watermarks. Both mathematical analyses and experiment results consistently verify the anti-disclosure capability of the video content-dependent watermarking scheme. Our approach is the first work that takes resistance to both the collusion and copy attacks into consideration.*

Keywords: Video frame-dependent watermark, Collusion attack, Copy attack, Video hash, Watermark estimation

I. Introduction

No matter what kinds of applications are considered, robustness is the critical issue affecting the practicability of a watermarking method in a DRM system. The robustness of the current watermarking methods has been examined with respect to removal attacks or geometrical attacks or both. Especially, removal attacks try to destroy the hidden signal $\mathbf{W}$ (originally embedded into the cover data $\mathbf{I}$) by manipulating the stego data $\mathbf{I^s}$ so that fidelity of the attacked data $\mathbf{I^a}$ is inevitably destroyed (i.e., $PSNR(\mathbf{I}, \mathbf{I^s}) \geq PSNR(\mathbf{I}, \mathbf{I^a})$). However, there indeed exist the attacks that can defeat a watermarking system without sacrificing perceptual quality. Typically, the collusion attack [3], [11], [12], which is a removal attack, can make colluded media data further perceptually similar to its cover version (i.e., $PSNR(\mathbf{I}, \mathbf{I^s}) \leq PSNR(\mathbf{I}, \mathbf{I^a})$). In video watermarking, there are two forms for the collusion attack: Type I collusion (applied to video frames embedded with the same watermark) and Type II collusion (applied to video frames embedded with different watermarks). Type I collusion is conducted first by averaging a set of extracted watermarks (usually obtained using denoising) to better estimate the hidden watermark and then the estimated one is subtracted from all frames in order that the hidden signal can be removed, whereas Type II collusion is operated by averaging those perceptually similar frames in order to directly remove the watermarks. However, Type II collusion is less powerful since it is restricted to operate only on a subset of video frames such that the video watermarks cannot be eliminated entirely. Hence, we will only focus on the Type I collusion in this paper. It should be noted that the conventional denoising-based removal attack [13] merely applied to one single image is a special case of the collusion attack.

On the other hand, the copy attack [4], which is a protocol attack, is developed to create the false positive problem; i.e., a situation in which one can successfully detect a watermark from unwatermarked data. Initially, the copy attack was applied to image watermarking and is carried out 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. Compared with the collusion attack, the copy attack can be executed on only one video frame or an image; thus it is more flexible. In this regard, the copy attack must be taken into consideration when the robustness of a watermarking system is to be evaluated. In addition, the analyses of the achievable performance between the denoising attack and the copy attack should refer to [9]. The common step used to realize a collusion or copy attack is "watermark estimation," which is usually accomplished by means of a denoising procedure. Consequently, we call both them watermark-estimation attacks (WEAs).

Previous collusion-resistant video watermarking methods are either computationally complex [12] or dependent on unstable feature extraction [11]. In addition, copy attack has been ignored. In this paper, we propose an anti-disclosure video watermarking scheme to resist both the collusion and copy attacks. We shall investigate the characteristics of WEAs and find 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 WEAs. Hence, the video-frame hash is addressed and combined with a hidden message to yield the video frame-dependent watermark (VFDW). Properties of the VFDW will be examined and mathematical analyses of the VFDW's resistance to WEAs will be elaborated. Experimental results will be provided to verify our analytic results.

II. Watermark Estimation Attack

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. An estimated watermark's energy is closely related to the accuracy of the removal attack. Several scenarios are shown in Fig. 1, 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. 1(a)∼(c), we can realize that even though the watermark's sign bits are fully obtained (Fig. 1(b)), the residual watermark signal (Fig. 1(c)) still suffices to reveal the encoded message due to the original watermark's energies cannot be completely discarded. 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 be helpful in improving removal efficiency. On the contrary, watermark removal in terms of energy subtraction operated in the opposite (wrong) polarity will undesirably damage the media data's fidelity. Actually, this corresponds to adding a watermark with higher energy into cover data without satisfying the masking constraint, as shown in Fig. 1(d). After Fig. 1(d) is subtracted from Fig. 1(a), the resultant residual watermark is illustrated in Fig. 1(e). By correlating Figs. 1(a) and (e), it is highly possible to reveal the existence of a watermark.

The above observations have inspired us to derive formal definitions of "optimal watermark estimation" and "perfect cover data recovery" as follows for use in further analyses.

Definition 1 (Optimal Watermark Estimation): Given an original watermark $\mathbf{W}$ and its approximate version $\mathbf{W^e}$ estimated from $\mathbf{I^s}$, the necessary condition for the optimal estimation of $\mathbf{W}$ as $\mathbf{W}^e$ is defined as

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

where $sgn(\mathbf{v})$ represents the signs of the elements in a vector $\mathbf{v}$. In Eq. (1), let Θ denotes the set of indices satisfying $sgn(W^e(i)) = sgn(W(i))$. The proof of collusion attack that tends to satisfy Eq. (1) will be provided in Appendix. Beyond this step, however, to avoid leaving a residual watermark (Fig. 1(c)) that can reveal the hidden watermark, accurate estimation of the energy of $\mathbf{W^e}$ is absolutely indispensable. Watermark removal can be achieved if the watermark energy to be subtracted is also larger than or equal to the added energy, i.e., $mag(W^e(i)) \geq mag(W(i))$, with $mag(t)$ being $|t|$. Therefore, the sufficient and necessary condition for complete watermark removal can be defined $\forall i \in \Theta$ as

$$mag(W^e(i)) \geq mag(W(i)) \quad \text{and} \quad sgn(W^e(i)) = sgn(W(i)).$$

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

$$PSNR(\mathbf{I}, \mathbf{I^r}) \approx \infty, \;\; \mathbf{I^r} = \mathbf{I} - sgn(\mathbf{W^e})mag(\mathbf{W^e}).$$

Of course, it is best to get $mag(W^e(i))$ as the upper bound of $mag(W(i))$; otherwise, even if the watermarks have been completely removed the qualities of attacked video frames/images will be poor. Typically, evaluation of $mag(\mathbf{W^e})$ can be achieved by means of either averaging [12] or remodulation [13].

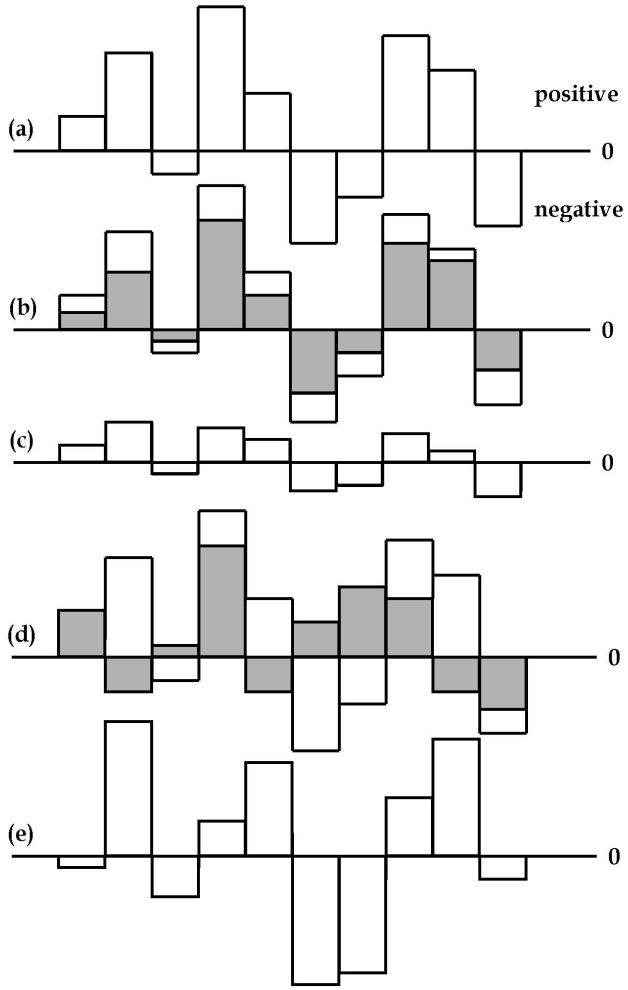


Fig. 1. Watermark estimation/removal illustrated with energy variation: (a) original embedded watermark with each white bar indicating the energy (determined using perceptual masking) of each watermark bit; (b) gray bars show the energies of an estimated watermark with all the signs being the same as the originals (a); (c) the residual watermark obtained after removing the estimated watermark (b); (d) the energies of an estimated watermark with most the signs being opposite to those in (a); (e) the residual watermark derived from (d). In the above examples, sufficiently large linear correlations between (a) and (c), and between (a) and (e) exist, indicating the presence of a watermark. The importance of polarities of watermark bits has also been previously emphasized in [6].

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 the colluded data.

III. Video Frame-Dependent Watermark

A. Frame Hash

From Sec. II, we have found that WEAs are achievable mainly based on the fact that the hidden watermark behaves like a noise, so anyone can reliably utilize all estimated noise-like watermarks. In order to disguise this prior knowledge and hide it from attackers, the key is to reduce the confidence of watermark estimation achieved by the collusion attack (by making $P_s(\beta \leq \frac{|\mathcal{C}|}{2})$ possible in Eq. (8)). To this end, we shall introduce the idea of the video frame-dependent hash as a kind

of content-dependent information used to create the so-called video frame-dependent watermark (VFDW). Meanwhile, the content-dependent information (called frame hash herein) must be secured by means of a secret key for anti-forgery and be robust to digital processing [8] in order not to affect watermark detection.

Here, the proposed video frame hash extraction procedure is operated in the VLC domain. For each macroblock, a piece of representative and robust information is created. It is defined in each macroblock i as the magnitude relationship between two energies computed from level values:

$$h(i) = \begin{cases} +1, & \text{if } \sum_j |f_j(p_1)| \geq \sum_j |f_j(p_2)|, \\ -1, & \text{otherwise,} \end{cases}$$

where $h(i)$ is an element of a frame hash $\mathbf{FH}$, j is the index that indicates a block belonging to a macroblock i, and $f_j(p_1)$ and $f_j(p_2)$ denote level values at zig-zaged positions p_1 and p_2 in a block j, respectively. The length of a $\mathbf{FH}$ is exactly equal to the number of macroblocks. In addition, the selected level values should be at lower frequencies because level-run pairs located at high-frequency positions are vulnerable to attacks. We call this feature value $h(\cdot)$ robust because this magnitude relationship can be mostly preserved under incidental modifications. Since the robustness issue of frame hash is beyond the score of this paper, the reader may refer to [10] for more robustness analyses. Next, the frame mash, $\mathbf{FH}$, is merged with the watermark, $\mathbf{W}$, to generate the video frame-dependent watermark ($\mathbf{VFDW}$) as

$$\mathbf{VFDW} = S(\mathbf{W}, \mathbf{FH}), \tag{2}$$

where $S(\cdot,\cdot)$ is a mixing function, which is operated based on a secret key (will be described in the following section) and is used to prevent attackers from forging the $\mathbf{VFDW}$. The sequence $\mathbf{VFDW}$ is what we will embed into a video frame.

B. Properties of VFDW

Let a video $\mathbf{V}$ be expressed as $\oplus_{i\in\Omega}\mathbf{F}_i$, where all frames $\mathbf{F}_i$ are concatenated to form $\mathbf{V}$ and Ω denotes the set of frame indices. In our video watermarking method, each frame $\mathbf{F}_i$ will be embedded with a content-dependent watermark $\mathbf{VFDW}_i$ to form a stego video $\mathbf{V^s}$, i.e.,

$$\mathbf{F^s}_i = \mathbf{F}_i + \mathbf{VFDW}_i, \qquad \mathbf{V^s} = \oplus_{i\in\Omega}\mathbf{F^s}_i,$$

where $\mathbf{F^s}_i$ is a stego frame and $\mathbf{VFDW}_i$, similar to Eq. (2), is defined as

$$\mathbf{VFDW}_i = S(\mathbf{W}, \mathbf{FM}_{\mathbf{F}_i}). \tag{3}$$

In Eq. (3), the mixing function $S(\cdot,\cdot)$ will be designed as a procedure of shuffling the frame hash $\mathbf{FM}_{\mathbf{F}_i}$ using the same secret key K (used to generate the watermark $\mathbf{W}$), followed by shuffling of the watermark to enhance security. Specifically, it is expressed as

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

where PT denotes a shuffling function controlled using the secret key K to achieve uncorrelated cross-correlation,

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

and auto-correlation:

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

The proposed content-dependent watermark possesses the characteristics described in the following. They are useful for proving resistance to WEAs.

Definition 3 Given two frames $\mathbf{F}_i$ and $\mathbf{F}_j$, their degree of similarity depends on the correlation between $\mathbf{FM}_{\mathbf{F}_i}$ and $\mathbf{FM}_{\mathbf{F}_j}$, i.e., $\delta_{nc}(\mathbf{F}_i, \mathbf{F}_j) = \delta_{nc}(\mathbf{FM}_{\mathbf{F}_i}, \mathbf{FM}_{\mathbf{F}_j})$. Two extreme cases exist: (i)if $\mathbf{F}_i = \mathbf{F}_j$, then $\delta_{nc}(\mathbf{F}_i, \mathbf{F}_j) = 1$; (ii)if $\mathbf{F}_i$ and $\mathbf{F}_j$ are visually dissimilar, then $\delta_{nc}(\mathbf{F}_i, \mathbf{F}_j) \approx 0$.

Proposition 1 Given two frames $\mathbf{F}_i$ and $\mathbf{F}_j$, $\delta_{nc}(\mathbf{F}_i, \mathbf{F}_j)$, and their respectively embedded content-dependent watermarks $\mathbf{VFDW}_i$ and $\mathbf{VFDW}_j$ that are assumed to be i.i.d. Gaussian distribution, the following properties can be established: (i) $\delta_{nc}(\mathbf{VFDW}_i, \mathbf{VFDW}_j)$ is linearly proportional to $\delta_{nc}(\mathbf{F}_i, \mathbf{F}_j)$; (ii) $\delta_{nc}(\mathbf{VFDW}_i, \mathbf{VFDW}_j) \leq \delta_{nc}(\mathbf{W}^2)$; (iii) $\delta_{nc}(\mathbf{W}, \mathbf{VFDW}) = 0$. Due to limits of space, proofs of Proposition 1 by exploiting the above properties can be found in [9]. It is essential to emphasize that property (i) of Proposition 1 contrasts with the one pointed out in [12], but the novelty of our scheme is that the concept of the content-dependent watermark has been employed.

C. Resistance to WEAs

Note that in order to better explain resistance of the VFDW to WEAs, our analyses are conducted in the spatial domain. This is reasonable because a signal embedded in the transformed domain can be transferred to another equivalent signal in the spatial domain and watermark estimation by means of denoising [4], [13] is intuitively applied in the spatial domain. Assume that by means of a collusion attack, the averaging operation is performed on stego frames $\mathbf{F^s}_i$'s of a stego video $\mathbf{V^s}$. From an attacker's perspective, each hidden watermark has to be estimated using a denoising operation, so deviation in estimation will inevitably occur. Let $\mathbf{W^e}_i$ be a watermark extracted from $\mathbf{F^s}_i$. In fact, $\mathbf{W^e}_i$ can be modeled as a partial hidden watermark plus a noise component, i.e.,

$$\mathbf{W^e}_i = \alpha_i \mathbf{VFDW}_i + \mathbf{n}_i,$$

where $\mathbf{n}_i$ represents a frame-dependent Gaussian noise with zero mean, α_i denotes the proportion 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, $1 \geq \alpha_i = \delta_{nc}(\mathbf{W^e}_i, \mathbf{VFDW}_i) > T$ always holds based on the fact that a watermark is a high-frequency signal, which can be efficiently extracted by means of denoising [2], [4], [13]. Let $\mathcal{C}$ ($\subset \Omega$) denote the set of frames used for collusion. By employing the Central Limit Theorem, the average of all the estimated watermarks (by means of collusion) can be expressed as

$$\bar{\mathbf{W}}^{\mathbf{e}} = \frac{\sqrt{|\mathcal{C}|}}{|\mathcal{C}|} \sum_{i\in\mathcal{C}} \mathbf{W^e}_i = \frac{1}{\sqrt{|\mathcal{C}|}} \sum_{i\in\mathcal{C}} (\alpha_i \mathbf{VFDW}_i + \mathbf{n}_i). \tag{4}$$

Now, we are ready to derive a sufficient and necessary condition for resisting a collusion attack as described in Proposition 2.

Proposition 2 In a collusion atatck, an attacker first estimates $\bar{\mathbf{W}}^{\mathbf{e}}$ from a set, $\mathcal{C}$, of stego frames. Then, a counterfeit unwatermarked video $\mathbf{V}^{\mathbf{u}}$ is generated from a stego video $\mathbf{V}^{\mathbf{s}} = \oplus_{i\in\Omega}\mathbf{F}^{\mathbf{s}}{}_i$ as

$$\mathbf{F}^{\mathbf{u}}{}_i = \mathbf{F}^{\mathbf{s}}{}_i - \bar{\mathbf{W}}^{\mathbf{e}}, \qquad \mathbf{V}^{\mathbf{u}} = \oplus_{i\in\Omega}\mathbf{F}^{\mathbf{u}}{}_i. \tag{5}$$

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

Proof: First of all, we need to derive $\delta_{nc}(\bar{\mathbf{W}}^{\mathbf{e}}, \mathbf{VFDW}_k)$. By making use of Eq. (4) and Proposition 1, we have the following brief derivation:

$$\begin{aligned}\delta_{nc}(\bar{\mathbf{W}}^{\mathbf{e}}, \mathbf{VFDW}_k) &= \frac{\sqrt{|\mathcal{C}|}}{|\mathcal{C}|}\delta_{nc}(\sum_{i\in\mathcal{C}}(\alpha_i\mathbf{VFDW}_i + \mathbf{n}_i), \mathbf{VFDW}_k) \\ &= \frac{\sum_{k\in\mathcal{C}}\alpha_k}{\sqrt{|\mathcal{C}|}}, \end{aligned} \tag{6}$$

where $\mathbf{VFDW}_k$ represents the content-dependent watermark embedded in $\mathbf{F}_k$. Consequently, given property (ii) of Proposition 1, and Eqs. (5) and (6), we get

$$\begin{aligned} &\delta_{nc}(\mathbf{F}^{\mathbf{u}}{}_k, \mathbf{VFDW}_k) > T \\ \text{iff } &\delta_{nc}(\mathbf{F}_k + \mathbf{VFDW}_k - \bar{\mathbf{W}}^{\mathbf{e}}, \mathbf{VFDW}_k) > T \\ \text{iff } &\delta_{nc}(\mathbf{VFDW}_k, \mathbf{VFDW}_k) - \delta_{nc}(\bar{\mathbf{W}}^{\mathbf{e}}, \mathbf{VFDW}_k) > T \\ \text{iff } &\delta_{nc}(\bar{\mathbf{W}}^{\mathbf{e}}, \mathbf{VFDW}_k) = \frac{\sum_{k\in\mathcal{C}}\alpha_k}{\sqrt{|\mathcal{C}|}} < 1 - T. \end{aligned} \tag{7}$$

Resistance to the copy attack can be similarly derived. To save space, please refer to [9] for more details.

IV. Experimental Results

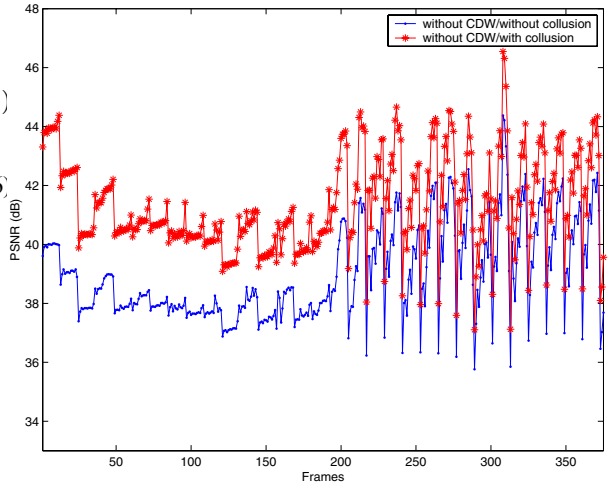
Three video sequences, including Flower garden, Table tennis, and Football, were used for watermarking. In this study, the side-informed real-time video watermarking scheme [7] was chosen for test, denoted as Method I, due to its simple implementation and robustness. The combination of our VFDW and Method I was denoted as Method II. However, we would like to particularly emphasize that the proposed VFDW can be readily applied to other video watermarking algorithms. On the other hand, Lee's Wiener filter [5] was used to perform denoising-based blind watermark extraction. The threshold T used to determine the existence of a watermark was selected as 0.11 if the desired false positive probability was approximately 10^{-7} [1].

A. Estimation of Watermark's Sign Bits

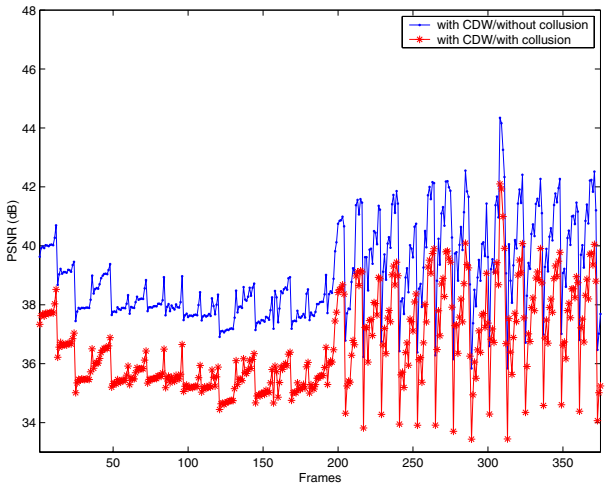
All the watermarks that were extracted from visually similar frames with respect to the three video sequences were averaged to obtain the estimated watermark. By comparing the estimated watermark and the original one, the bit error rates (BERs) regarding the watermark's sign bits is summarized from an attacker's perspective as follows: (i) if VFDW is not used, $1\% \leq$BERs$\leq 13\%$ is yielded; (ii) when VFDW is used, BERs are all raised to about 50% (i.e., resemble random guess). This experiment confirms that the VFDW can efficiently confuse the watermark estimation conducted by attackers.

B. VFDW Resistance to Collusion Attack

The collusion attack was applied to Method I and Method II, respectively. The impacts of the collusion attack and the VFDW will be examined with respect to the two scenarios: (s1) the quality of a colluded video; and (s2) watermark detection after performing collusion. Some results are depicts in Figs. 2 and 3, respectively. In summary, as long as a frame hash is involved in constructing a watermark, even a collusion attack is applied owners still can extract the watermarks and fidelities of colluded videos cannot be improved.



(a) Method I



(b) Method II

Fig. 2. Quality of a colluded video: (a) the PSNR values of the colluded frames (top) are higher than those of the stego frames; (b) when VFDW is applied, the PSNR values of the colluded frames (bottom) become lower than those of the stego frames. This experiment reveals that a collusion attack will fail to improve the fidelity of a colluded video when VFDW is involved.

C. VFDW Resistance to Copy Attack

The copy attack was applied to Method I and Method II to compare their capability of resistance. One of the videos was first watermarked, and then the watermark was estimated and copied to the other unwatermarked videos to form counterfeit stego videos. The PSNR values of the attacked video frames were in the range of $36 \sim 55$dB (no masking was used).

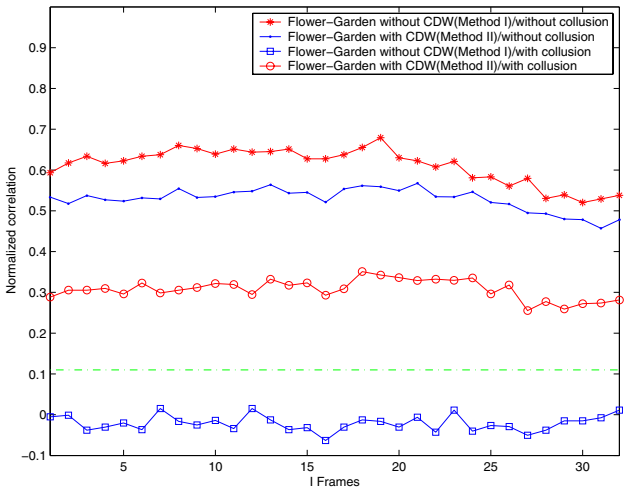


Fig. 3. Watermark detection under collusion (the dashdot line indicates the threshold $T = 0.11$). By comparing the detection curves obtained using different $|\mathcal{C}|$'s, it can be found that performance degradation of the proposed anti-disclosure watermarking method is lower bounded by the denoising-based removal (i.e., $|\mathcal{C}| = 1$). If VFDW is not used, collusion with $|\mathcal{C}| > 1$ provides more effective removal. These results are exactly consistent with Proposition 2.

The normalized correlations obtained by employing the copy attack to Method I fell within the interval $[0.487 \quad 0.650]$ (all were sufficiently larger than $T = 0.11$), which indicated the presence of watermarks. However, when the VFDW was introduced, these correlations decreased significantly to the interval $[-0.056 \quad 0.060]$, which indicated the absence of watermarks. The experimental results are consistent with the analytic result that the proposed VFDW is able to deter the detection of copied watermarks.

V. Concluding Remarks

Robustness is still a major issue that determines whether a watermarking scheme could play a major role in a digital right management (DRM) system. Inherently, a video watermarking method is prone to the collusion and copy attacks, which are recognized to be able to defeat watermarking systems without needing much cost. In this paper, the video frame-dependent watermark (VFDW), which is a mixture of a frame hash and a hidden message, has been explored to withstand watermark-estimation attacks (WEAs). Notably, 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 for achieving complete watermark removal. The characteristics of the VFDW have been analyzed to justify its resistance to WEAs. Overall, the experimental results have confirmed our mathematical analyses about WEAs and VFDW. To our knowledge, we are the first to employ the content-dependent video watermark in resisting the collusion and copy attacks, simultaneously.

The proposed media hash at its current status is sensitive to geometric distortions and could potentially affects the resistance of the VFDW to them. In our future work, we will continue to study this challenging problem, i.e., geometrical invariance of the media hash.

Appendix: Confidence of Watermark Sign Estimation under the Collusion Attack

Here, we will justify our confidence in collusive estimation of a watermark's sign using binomial probability distribution. Suppose each $sgn(W_f^e(i))$ $f \in \mathcal{C}$ ($\mathcal{C}$ is a collusion set) is regarded as a trial, and that the trials are independent. Each trial will result in one of two possible outcomes: $+1$ and -1. Each outcome will occur with equal probability, 0.5. Now, our confidence in the occurrence of $sgn(W_f^e(i))$ is formulated as the probability P_s of $sgn(W_f^e(i))$'s, which are 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 $sgn(W_f^e(i))$ observed in Δ. As a consequence, P_s can be expressed as

$$P_s(\beta > \frac{|\mathcal{C}|}{2}) = \sum_{n=\frac{|\mathcal{C}|}{2}+1}^{|\mathcal{C}|} \binom{|\mathcal{C}|}{n} 0.5^{|\mathcal{C}|}. \qquad (8)$$

Looking at the table of binomial probabilities, we can find that P_s will increase rapidly (usually it will be larger than 0.8) as long as n is slightly larger than $\frac{|\mathcal{C}|}{2}$. The larger P_s is, the more confident we are. Hence, we have sufficient confidence to rely on the collusion attack to determine the sign of a hidden watermark.

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

References

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

[2] 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.

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

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

[5] 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.

[6] 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.

[7] C. S. Lu, J. R. Chen, H. Y. Mark Liao, and K. C. Fan, "Real-Time MPEG2 Video Watermarking in the VLC Domain", *Proc. Int. Conf. on Pattern Recognition*, Vol. II, Canada, 2002.

[8] 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.

[9] C. S. Lu and C. Y. Hsu, "Content-dependent Anti-Disclosure Image Watermark", *Proc. Int. Workshop on Dogotal Watermarking*, LNCS 2939, Seoul, Korea, 2003.

[10] C. S. Lu, C. Y. Hsu, S. W. Sun, and P. C. Chang, "Robust Mesh-based Hashing for Copy Detection and Tracing of Images," submitted to *IEEE Int. Conf. on Multimedia and Expo*, Taipei, Taiwan, 2004.

[11] K. Su, D. Kundur, D. Hatzinakos, "Statistical Invisibility for Collusion-resistant Digital Video Watermarking," to appear in *IEEE Trans. on Multimedia*.

[12] 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.

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