Security Pitfalls of Frame-by-Frame Approaches to Video Watermarking

Gwenaël Doërr and Jean-Luc Dugelay *Senior Member, IEEE*

***Abstract*— Watermarking digital video material is usually considered as watermarking a sequence of still images. However, such a frame-by-frame approach is very risky since straightforward embedding strategies can result in poor performance in terms of security i.e. against hostile attacks. As examples, two very common video watermarking systems will be presented as well as the associated intra-video collusion attacks which defeat them. Then, both watermark modulation and embedding strength modulation will be surveyed to design alternative embedding strategies which exhibit superior performance against such attacks. Nevertheless, it will also be shown that an expert attacker can still construct an effective watermark removal attack. Finally, there will be a discussion to assert whether or not security against intra-video collusion can be achieved with such blind frame-by-frame embedding strategies.**

***Index Terms*— Video watermarking, security, intra-video collusion attacks, watermark estimation**

I. Introduction

THE last century saw the enormous growth of the digital world: old analog audio tapes were substituted by digital disks, personal computers with internet connections took homes by storms and Digital Versatile Disk (DVD) players invaded living rooms. Unfortunately, this has also raised many concerns regarding copyright protection since digital data can be perfectly duplicated and rapidly redistributed on a large scale. Today, even non-technical users can exchange copyrighted material via Peer-to-Peer networks and multimedia content providers have requested security mechanisms before releasing their highly valued property. Many Digital Right Management (DRM) frameworks rely on end-to-end encryption to make digital data completely unusable without the proper decryption key. However, this protection falls when encrypted data is decrypted to eventually be presented to a human user. Digital watermarking [1], [2] was consequently introduced in the 90's as a second line of defense to fill this *analog hole*.

Digital watermarking basically consists of embedding a key dependent secret signal into digital data in a robust and invisible way. Moreover, this underlying signal is closely tied to the host data so that it survives digital to analog conversion. There is a complex trade-off between three parameters: *data payload*, *fidelity* and *robustness*. Data payload is the number of bits encoded by the hidden watermark. Fidelity is related to the distortion, which the watermark embedding process is bound to introduce: the inserted watermark should remain imperceptible to a human user. Finally, the robustness of a watermarking scheme can be seen as the ability of the detector to extract the watermark from some altered watermarked data. These parameters are in conflict and a compromise must be found depending on the targeted application.

Embedding a watermark in video content can be useful in many applications [3] to provide for example services such as copy control for DVD or traitor tracing in Video-on-Demand frameworks. Today, video watermarking basically extends results obtained for still images. Thus, two common frame-by-frame approaches are presented in Section II. Such straightforward adaptations have however led to non-secure algorithms and two specific attacks are introduced to illustrate this point in the next section. Sections IV and V explore then two strategies to improve performance against collusion attacks: watermark modulation and embedding strength modulation. Nevertheless, it is also shown that an expert attacker is still able to defeat these new strategies. Finally, lessons to be learned are gathered in the last section and the need for *informed watermarking* is discussed.

II. Frame-by-frame Video Watermarking

Some video watermarking algorithms exploit the specificities of a compression standard. Others embed a watermark in a three dimensional transform. However, watermarking digital video is mostly considered today as watermarking a sequence of still images [3]. Once this approach is enforced, two major embedding strategies are used: either a *different watermark* is inserted in each video frame, or the *same watermark* is embedded in all the video frames. For sake of simplicity, both strategies are illustrated with an additive watermark based on the Spread Spectrum (SS) theory in the next subsections.

A. Uncorrelated Watermarks Embedding

In the pioneering spread spectrum based video watermarking technique [4], video was considered as a one-dimensional signal. From a frame-by-frame perspective, this can be seen as a system which *always embeds a different watermark* as depicted in Figure 1. In such a *SS system*, the embedder inserts a pseudo-random watermark in each video frame:

$$\check{\mathbf{F}}_t = \mathbf{F}_t + \alpha \mathbf{W}_t(K), \quad \mathbf{W}_t(K) \sim \mathcal{N}(0,1) \tag{1}$$

where $\mathbf{F}_t$ is the luminance of the t^{th} video frame, $\check{\mathbf{F}}_t$ the luminance of the t^{th} watermarked frame, α the embedding strength and K a secret key. The inserted watermark $\mathbf{W}_t(K)$

Manuscript version: February 20, 2004

M. Doërr and M. Dugelay are with the Eurécom Institute, 2229 route des Crêtes, BP 193, 06904 Sophia-Antipolis, France. Email: {doerr, dugelay}@eurecom.fr

has a normal distribution with zero mean and unit variance and is different at every instant t. Using $K + t$ as a seed for the pseudo-random generator is a simple way to obtain this property. Perceptual shaping can be introduced to improve the invisibility of the watermark even if a global embedding strength has been used in practice. From a subjective point of view, always changing the embedded watermark introduces an annoying flicker artifact [5].

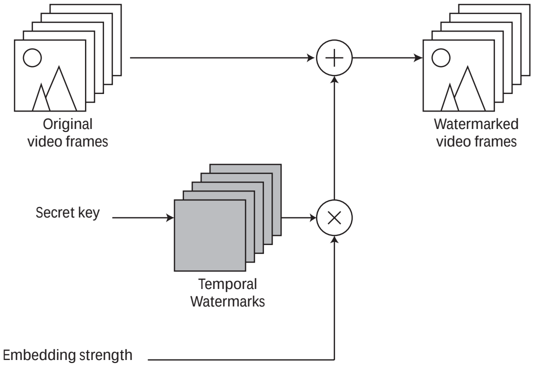


Fig. 1. SS system: A different watermark is embedded in each video frame.

The detector computes then the following correlation score:

$$\rho\big(\{\check{\mathbf{F}}_t\}\big) = \frac{1}{T}\sum_{t=1}^{T}\check{\mathbf{F}}_t \cdot \mathbf{W}_t = \alpha + \frac{1}{T}\sum_{t=1}^{T}\mathbf{F}_t \cdot \mathbf{W}_t \approx \alpha \qquad (2)$$

where T is the number of considered video frames and $\cdot$ denotes the linear correlation operation. This score should be equal to α if a watermark is present in the video, while it should be almost equal to zero if no watermark has been inserted. Moreover, host interference can be cancelled in a preprocessing step [6] to enhance the detection statistics. As a result, the computed score is compared to a threshold τ_{detect} to assert the presence or absence of the watermark. The value $\alpha/2$ has been chosen in practice to obtain equal false positive and false negative probabilities[1].

B. Redundant Watermark Embedding

The SS system is highly sensitive to temporal desynchronization. A simple frame drop or insertion succeeds in confusing the detector. The alternative *SS-1 system* depicted in Figure 2 has consequently been introduced. It basically *always embeds the same watermark* [7]. In other terms, the embedder redundantly inserts the same pseudo-random watermark in each video frame:

$$\check{\mathbf{F}}_t = \mathbf{F}_t + \alpha\mathbf{W}(K), \quad \mathbf{W}(K) \sim \mathcal{N}(0,1) \qquad (3)$$

where $\mathbf{W}(K)$ is a key-dependent reference watermark. From a subjective perspective, this embedding strategy produces an annoying persistent pattern [5] when the camera moves.

On the detector side, the correlation score defined in (2) is computed. Now that the same watermark is embedded in each video frame, the linearity of the operator $\cdot$ can be exploited to reduce the number of computations [7] required for detection:

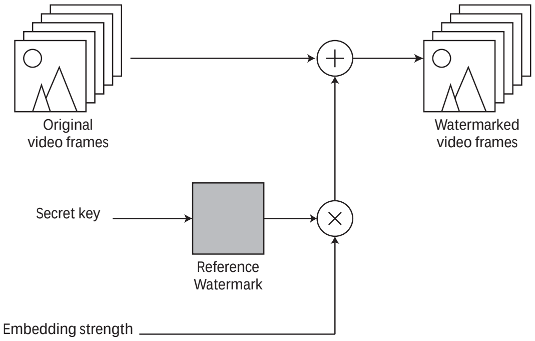


Fig. 2. SS-1 system: The same reference watermark is redundantly embedded in each video frame.

$$\rho\big(\{\check{\mathbf{F}}_t\}\big) = \frac{1}{T}\sum_{t=1}^{T}\check{\mathbf{F}}_t \cdot \mathbf{W} = \left(\frac{1}{T}\sum_{t=1}^{T}\check{\mathbf{F}}_t\right) \cdot \mathbf{W} \qquad (4)$$

This means that averaging several correlations between different video frames and the same watermark is equivalent to computing a single correlation between the average of the video frames and this watermark. Here again, the correlation score should be equal to α if a watermark is present in the video, while it should be almost equal to zero if no watermark has been inserted. As a result, the computed score is compared to a threshold τ_{detect}, which is set equal to $\alpha/2$ in practice to obtain equal false positive and false negative probabilities.

III. Weaknesses against Collusion Attacks

Previous works have mainly focused on robustness i.e. resilience against non-malicious attacks. For example, for applications such as broadcast monitoring, video authentication or data hiding, the watermark has to undergo some signal processing e.g. noise addition, filtering, lossy compression. However, for fingerprinting or copy-control applications, the embedded watermark has also to survive in a hostile environment with malicious users. In this context, security issues have to be addressed. Security has been defined as *the inability by unauthorized users to have access to the raw watermarking channel* [8] and is usually neglected during watermarking evaluation. The remainder of this section consequently introduces collusion attacks, which can be used to evaluate security.

A. Inter-videos Collusion

Collusion can be seen as eavesdropping the watermarking channel to identify some hidden properties and exploiting this knowledge to damage information transmitted on this secret communication channel. In practice, several watermarked documents are combined with a linear or non-linear operator [9] to obtain unwatermarked content. For example, a group of malicious customers can gather several versions of the same movie containing different watermarks and average them to wash out the underlying watermarks. A known countermeasure consists in designing the set of distributed watermarks so that a coalition, gathering few customers in comparison with the total

[1] Adding some noise to the watermarked video introduces an interfering term in (2), which has zero mean and a variance proportional to $1/\sqrt{T}$. In other words, modifying T enables to adjust the false positive and false negative probabilities.

number of users, cannot remove the entire watermark [10]. Furthermore, the remaining watermark signal should identify at least one of the colluders without ever framing any innocent customer. Collusion can also occurs when several movies carry the same watermark. In this case, the attacker roughly estimates the embedded watermark from each movie and combines them to refine the watermark estimate, which is later remodulated to stir out the watermark signal. A simple counterattack consists then in making the hidden watermark dependent on the host signal so that watermark estimation is not possible.

B. *Intra-video Collusion*

The previous attacks require several watermarked videos to produce unwatermarked video content. In contrast, intra-video collusion attacks aim at removing an underlying watermark using only a single watermarked video. Since frame-by-frame watermarking is commonly used, an attacker can indeed view each single video frame as a watermarked content to be exploited for collusion. As a result, unless such attacks are carefully considered, video watermarking schemes are doomed to be broken once released to a large hostile audience [11]. To support this idea, the remainder of this subsection presents two basic intra-video collusion attacks which succeed in removing the watermarks inserted by both SS and SS-1 systems.

1) Temporal Frame Averaging (TFA): Since neighboring video frames are highly similar, temporal low-pass filtering can be performed without introducing much visual distortion:

$$\dot{\mathbf{F}}_t = \mathrm{L}_w(\mathcal{F}_t), \quad \mathcal{F}_t = \{\mathbf{F}_u, -w/2 \leq t - u < w/2\} \tag{5}$$

where w is the size of the temporal window, L_w is a temporal low-pass filter and $\dot{\mathbf{F}}_t$ is the resulting t^{th} attacked video frame. In experiments, a simple 3-frames temporal averaging filter has been used. Assuming that a watermarked video $\{\check{\mathbf{F}}_t\}$ is temporally averaged, the following correlation score is obtained on the detector side:

$$\rho\big(\{\dot{\mathbf{F}}_t\}\big) \approx \frac{\alpha}{wT} \sum_{t=1}^{T} \Bigg(\sum_{u \in [-\frac{w}{2}, \frac{w}{2}[} \mathbf{W}_{t+u} \cdot \mathbf{W}_t \Bigg) \tag{6}$$

If the same watermark has been redundantly embedded (SS-1 system), all the correlation terms $\mathbf{W}_{t+u} \cdot \mathbf{W}_t$ are equal to 1 and the correlation score is equal to α. In other words the TFA attack fails. Alternatively, if uncorrelated watermarks have been inserted in successive video frames (SS system), the term corresponding to the index $u = 0$ in the second summation is the only one not to be null and the correlation score is reduced to α/w. As a result, for w greater than 2, the correlation score drops below the detection threshold τ_{detect} and the attack is a success. Averaging many video frames is likely to result in poor quality video in dynamic scenes. This attack is consequently more relevant in static scenes even if it can be adapted to cope with dynamic ones thanks to frame registration [12].

2) Watermark Estimation Remodulation (WER): Computing the difference $\Delta_{\text{o}}(\check{\mathbf{F}}) = \check{\mathbf{F}} - \mathbf{F}$ is the optimal approach to estimate the watermark embedded in a given video frame. However, the attacker does not have access to the original digital content and has to blindly estimate in practice the hidden watermark. Digital watermarks are usually located in high frequencies. A rough estimation of the watermark can consequently be obtained with denoising techniques, or more simply by computing the difference between the watermarked frame and its low-pass filtered version [13]:

$$\Delta(\check{\mathbf{F}}) = \check{\mathbf{F}} - \mathrm{L}(\check{\mathbf{F}}) \tag{7}$$

where $\mathrm{L}(.)$ is a low-pass filter e.g. a simple 5×5 spatial averaging filter. Then, estimations obtained from different video frames are averaged [11]:

$$\tilde{\mathbf{W}}_T = \frac{1}{T} \sum_{t=1}^{T} \tilde{\mathbf{W}}_t = \frac{1}{T} \sum_{t=1}^{T} \Delta(\check{\mathbf{F}}_t) \tag{8}$$

where T is the number of considered video frames for collusion. In practice, the estimator defined in (7) produces badly estimated samples around discontinuities (edges or textured areas). An additional thresholding operation is consequently performed to discard samples whose magnitude is greater than τ_{valid}. The threshold value has been set to 8 for experiments and the number of valid estimations for each watermark sample has been counted to allow pertinent normalization in (8). The resulting watermark $\tilde{\mathbf{W}}_T$ is then subtracted from each watermarked video frame with a remodulation strength β. This strength is chosen to introduce a distortion similar to the one due to the watermarking process in terms of Peak Signal to Noise Ratio (PSNR). The attacked video frames are thus given by:

$$\dot{\mathbf{F}}_t = \check{\mathbf{F}}_t - \alpha \tilde{\mathbf{W}}_{T_n} = \check{\mathbf{F}}_t - \alpha \frac{\tilde{\mathbf{W}}_T}{\sqrt{\tilde{\mathbf{W}}_T \cdot \tilde{\mathbf{W}}_T}} \tag{9}$$

Assuming that the attacker has access to the estimator $\Delta_{\text{o}}(.)$, when a watermarked video is submitted to the WER attack, the detector obtains the following correlation score:

$$\rho\big(\{\dot{\mathbf{F}}_t\}\big) \approx \alpha \left[1 - \frac{1}{T^2 \sqrt{\tilde{\mathbf{W}}_T \cdot \tilde{\mathbf{W}}_T}} \sum_{t=1}^{T} \sum_{u=1}^{T} \mathbf{W}_u \cdot \mathbf{W}_t \right] \tag{10}$$

If the watermarks embedded in different video frames are uncorrelated (SS system), the correlation term $\mathbf{W}_u \cdot \mathbf{W}_t$ is equal to δ_u^t where δ is the Kronecker delta and the correlation score after attack is equal to $\alpha(1 - 1/\sqrt{T})$ which is almost equal to α for large T. As a result, the attack does not succeed in removing an embedded watermark if a strategy which *always embeds a different watermark* is enforced. On the other hand, if the same watermark has been redundantly embedded in all the video frames (SS-1 system), each correlation term is equal to 1 and the correlation score drops to zero. This result has to be contrasted since the attacker has not access to $\Delta_{\text{o}}(.)$. However, combining several individual estimates as in (8) refines the final one and the attack proves to be a success in practice [11]. In fact, the more the video frames are different, the more each individual watermark estimate refines the final one i.e. the attack is more relevant in dynamic scenes.

IV. Switching between Alternative Watermarks

Section III highlighted two important facts. First, uncorrelated watermarks can be washed out with temporal frame averaging. Second, a redundant watermark can be estimated and later removed via remodulation. Watermark modulation is explored in the remainder of this section: for each video frame, the watermark is picked out from a finite pool of reference watermark patterns. The superiority of this strategy in terms of security is demonstrated both theoretically and experimentally. Its limitations against an expert attacker are also outlined.

A. SS-N System

Periodic watermark schedules have been investigated for temporal synchronization [14]. However, from a security point of view, repeating the same sequence of watermarks allows an attacker to group frames carrying the same watermark before performing a WER attack. Thus, for each video frame, the watermark should rather be randomly chosen from a finite set of N watermarks $\{\mathbf{W}_i\}$ as depicted in Figure 3. Both previous systems are specific cases of this novel architecture: $N = 1$ for SS-1 system and $N = \infty$ for SS system. Watermarks are orthonormalized to prevent cross-talk on the detector side. The embedding process can then be rewritten:

$$\check{\mathbf{F}}_t = \mathbf{F}_t + \alpha \mathbf{W}_{\Phi(t)}, \quad \mathrm{P}\big(\Phi(t) = i\big) = p_i \tag{11}$$

where the p_i's are the emission probabilities of the system. From a subjective point of view, changing the watermark pattern still introduces a flicker artifact.

On the detector side, a new correlation score[2] is computed:

$$\rho\big(\{\check{\mathbf{F}}_t\}\big) = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N} |\check{\mathbf{F}}_t \cdot \mathbf{W}_i| \tag{12}$$

For each video frame, N linear correlations are computed and their absolute values are summed before being temporally averaged. This detection process does not require synchronization. However, the complexity[3] of the detector is increased by a factor N and the linearity of the operator $\cdot$ cannot be exploited as in (4) because of the absolute values. Immediately after embedding, the detector obtains:

$$\begin{aligned} \rho\big(\{\check{\mathbf{F}}_t\}\big) &= \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N} \Big|\mathbf{F}_t \cdot \mathbf{W}_i + \alpha \mathbf{W}_{\Phi(t)} \cdot \mathbf{W}_i\Big| \\ &\approx \frac{\alpha}{T}\sum_{t=1}^{T}\sum_{i=1}^{N} \delta^{\,i}_{\Phi(t)} \approx \alpha \end{aligned} \tag{13}$$

Host interference is cancelled in a preprocessing step [6] to improve detection statistics. The correlation score is then equal to α if a watermark is present in the video and to zero otherwise. This score is consequently compared to a threshold τ_{detect}, which is set equal to $\alpha/2$ in practice, in order to assert the presence or absence of a hidden watermark.

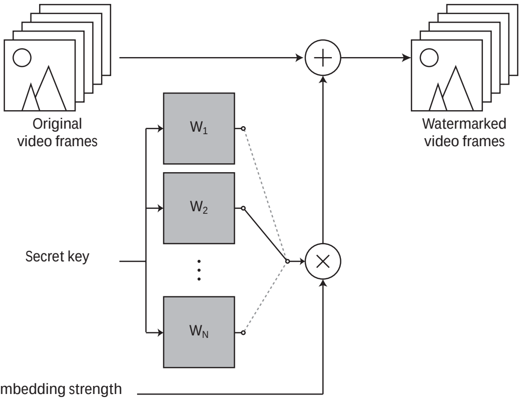


Fig. 3. SS-N system: the embedder inserts a watermark randomly chosen from a collection of N reference watermarks.

B. Enhanced Security

If a watermarked video is temporally averaged with a large window size w i.e. a strong attack without any concern for video quality, the attacked video frames are then given by:

$$\dot{\mathbf{F}}_t = \frac{1}{w}\sum_{u \in [-\frac{w}{2}, \frac{w}{2}[} \mathbf{F}_{t+u} + \alpha \sum_{i=1}^{N} p_i \mathbf{W}_i \tag{14}$$

Thus, the detector obtains the following correlation score:

$$\rho\big(\{\dot{\mathbf{F}}_t\}\big) = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N} |\dot{\mathbf{F}}_t \cdot \mathbf{W}_i| \approx \frac{\alpha}{T}\sum_{t=1}^{T}\sum_{i=1}^{N} p_i \approx \alpha \tag{15}$$

TFA spreads the energy of a watermark embedded in a video frame over its neighboring frames. In the SS system, when the detector checks for the presence of the watermark that should be embedded in each video frame, it misses most of the watermark signal. On the other hand, the SS-N detector checks the presence of *all the watermarks* of the set $\{\mathbf{W}_i\}$ in each video frame and thus retrieves all the parts of each watermark. As a result, the TFA attack fails.

Assuming that the attacker has access to the estimator $\Delta_{\text{o}}(.)$, if a watermarked video is submitted to the WER attack, the final watermark estimate is equal to $\frac{1}{T}\sum_{t=1}^{T} \mathbf{W}_{\Phi(t)}$. After remodulation, the following video frames are produced:

$$\dot{\mathbf{F}}_t = \mathbf{F}_t + \alpha\left[\Big(1 - \frac{p_t}{\nu}\Big)\mathbf{W}_{\Phi(t)} - \sum_{i \neq \Phi(t)} \frac{p_i}{\nu}\mathbf{W}_i\right] \tag{16}$$

where $\nu = \sqrt{\tilde{\mathbf{W}}_T \cdot \tilde{\mathbf{W}}_T}$. Subsequently, the detector obtains the following correlation score:

$$\begin{aligned} \rho\big(\{\dot{\mathbf{F}}_t\}\big) &\approx \frac{\alpha}{T}\sum_{t=1}^{T}\sum_{i=1}^{N} \left|\Big(1 - \frac{p_{\Phi(t)}}{\nu}\Big)\delta^{\,i}_{\Phi(t)} - \sum_{j \neq \Phi(t)} \frac{p_j}{\nu}\delta^{\,j}_{\Phi(t)}\right| \\ &\approx \alpha \sum_{i=1}^{N} p_i \left[\Big(1 - \frac{p_i}{\nu}\Big) + \sum_{j \neq i} \frac{p_j}{\nu}\right] \end{aligned} \tag{17}$$

If all the p_i are equal to $1/N$, the norm ν is equal to $1/\sqrt{N}$ and (17) becomes:

$$\rho\big(\{\dot{\mathbf{F}}_t\}\big) = \alpha\left[1 + (N-2)\frac{\sqrt{N}}{N}\right] \tag{18}$$

[2] Changing the detector has of course an impact on the detection statistics. In particular, the variance is increased by a factor $\sqrt{N}$ in comparison with SS and SS-1 systems i.e. more frames need to be accumulated to have the same false positive and false negative probabilities.

[3] Complexity can be reduced by using non full frame watermark patterns. In other terms, each frame is partitioned in N non-overlapping areas and each watermark pattern is spread over one of these areas. As a result, each$\check{\mathbf{F}}_t \cdot \mathbf{W}_i$ has N times fewer terms. However, this also alters detection statistics i.e. robustness performance.

In other terms, for N greater or equal to 2, the correlation score is greater or equal to τ_{detect} and the attack fails. Here, using several watermarks has interfered with the watermark estimation process. Thus, the attacker can only remove a small fraction $\sqrt{N}/N$ of the embedded watermark in each video frame. On the other hand, a small part of all the other watermarks from the set $\{\mathbf{W}_i\}$ is also removed. Then, summing the *absolute values* of the linear correlations succeeds in compensating the loss of correlation with the originally embedded watermark. Absolute values play a key role in fact. If they are removed from (12), the algorithm is still immune to TFA but the WER attack causes then the correlation score to drop to zero. Equation (18) also reminds that the WER attack is a success for $N = 1$ (SS-1 system).

C. Experimental Results

Five videos (704×576, 25 frames per second, 375 frames) are used for experiments. Their content is summarized in Table I. They are watermarked with the three watermarking schemes presented, with a global embedding strength equal to 3. The PSNR is consequently around 38 dB which ensures the watermark invisibility. Four different watermarks have been used for the SS-N system. The watermarked videos are then submitted to TFA on one hand and to the WER attack on the other. Finally, the correlation score is computed for all the videos.

TABLE I
DESCRIPTION OF THE VIDEOS USED FOR EXPERIMENTS

Video shot	*Short description*
Ping-Pong	Moving players, camera zoom/static/pan
Ski	Fast moving skier tracked by the camera
Susie	Girl on phone close-up, lips/eye/head motion
Train	Many moving objects (train, ball, calendar), camera pan
Tree	Static landscape background, camera static/pan

Each watermarking scheme is represented in Figure 4 by a specific symbol: crosses for SS system, triangles for SS-1 system and circles for SS-N system. The figure has also been divided into four quadrants whose borders are defined by the detection threshold $\tau_{\text{detect}} = 1.5$. The crosses are located in the upper-left quadrant, which confirms that the SS system resists the WER attack while it is weak against TFA. In fact they are in the neighborhood of the line defined by $\mathbf{y} = w\mathbf{x}$ ($w = 3$ in the experiments) as can be predicted from theoretical results in Section III-B. On the other hand, the triangles are in the lower-right quadrant, which supports conjectures asserting that the SS-1 system is robust against TFA while the WER attack succeeds in stirring out the embedded watermark, even if this latter attack is more or less efficient depending on the video content of the shot. Finally, the circles are in the upper-right quadrant, meaning that the SS-N system effectively resists both TFA and WER attacks. The WER attack even increases the correlation score as asserted in (18).

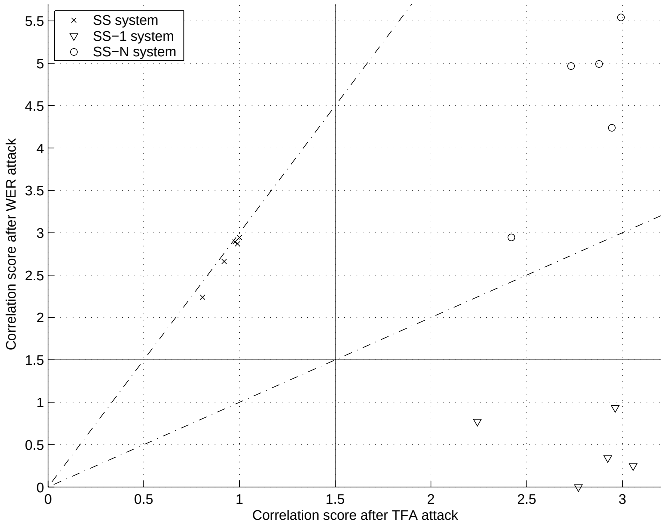


Fig. 4. Resilience of the three presented video watermarking systems (SS, SS-1 and SS-N) against TFA and WER intra-video collusion attacks.

D. Watermark Estimations Clusters Remodulation (WECR)

Attackers are likely to modify and adjust their approach according to this novel watermarking strategy. The security of the SS-N system basically relies on the assumption that attackers are unable to build sets of frames carrying the same watermark. Otherwise, a simple WER attack performed on each subset succeeds in estimating the pool of secret watermarks. A successful brute force attack can be theoretically designed [15] but its computational complexity may prevent its use in practice. Individual watermark estimates $\{\tilde{\mathbf{W}}_t\}$ obtained from different video frames can be seen as vectors in a very high dimensional space. Since these vectors should approximate the embedded watermarks $\{\mathbf{W}_i\}$, the problem comes down to vector quantization. In other words, the goal is to define N clusters $\mathcal{C}_i$ whose centroids $\mathbf{C}_i$ are good estimates of the secret watermarks.

1) Attack Description: The k-means algorithm is a simple way to perform vector quantization. In a first step, the individual watermark estimates $\{\tilde{\mathbf{W}}_t\}$ are distributed amongst different clusters $\{\mathcal{C}_i\}$, so that each vector is assigned to the cluster associated with its nearest centroid $\mathbf{C}_i$ according to the distance below:

$$\mathrm{d}(\tilde{\mathbf{W}}_t, \mathbf{C}_i)^2 = \frac{1}{P}\left[\sum_{x\in\mathcal{V}}\left(\tilde{\mathbf{W}}_t(x)-\mathbf{C}_i(x)\right)^2+\sum_{x\notin\mathcal{V}}\mathbf{C}_i^2(x)\right] \quad (19)$$

where P is the frame dimension and $\mathcal{V}$ the set of valid samples i.e. whose magnitude is lower than τ_{valid}. The first term in (19) measures how close the observation $\tilde{\mathbf{W}}_t$ is from the centroid $\mathbf{C}_i$ considering only the valid samples. The other term is a penalty term which favors observations having more valid samples. In a second step, the centroids are updated using only valid samples and the algorithm iterates until convergence.

To avoid random initialization, a splitting strategy [16] has been introduced. The basic idea is to start with a single cluster and to increment iteratively the number of clusters. Once the k-means algorithm has run until convergence, the log-likelihood

L_i of each cluster is computed:

$$L_i = -\frac{|\mathcal{C}_i|}{2}\left[1+\log\left(\frac{2\pi}{|\mathcal{C}_i|}\sum_{\tilde{\mathbf{W}}_t\in\mathcal{C}_i}\mathrm{d}(\tilde{\mathbf{W}}_t,\mathbf{C}_i)^2\right)\right] \tag{20}$$

where $|\mathcal{C}_i|$ is the number of vectors contained in the cluster $\mathcal{C}_i$. The worst cluster, the one with the lowest log-likelihood, is then identified and its associated centroid $\mathbf{C}_{\text{worst}}$ is split in $\mathbf{C}_{\text{worst}}\pm\epsilon\mathbf{D}$ where ϵ is a very small value and $\mathbf{D}$ is a direction to be set. This direction can be fixed, random or even better the direction of principal variation in the cluster. After each split, the k-means algorithm is run until convergence. This splitting strategy is stopped when the last split has not significantly reduced the average of the distances between each watermark estimate $\tilde{\mathbf{W}}_t$ and its nearest centroid.

At this point, M centroids have been obtained which are assumed to estimate the embedded watermark patterns. Thus, they can be remodulated to alter the watermark signal:

$$\dot{\mathbf{F}}_t = \check{\mathbf{F}}_t - \alpha\frac{\mathbf{C}_{\tilde{\phi}(t)}}{\sqrt{\mathbf{C}_{\tilde{\phi}(t)}\cdot\mathbf{C}_{\tilde{\phi}(t)}}} \tag{21}$$

where $\tilde{\phi}(t) = \arg\max_i \check{\mathbf{F}}_t\cdot\mathbf{C}_i$. If the attacker knows how many watermarks have been used during embedding, an additional merging step [17] can be introduced to have exactly the same number N of centroids. The basic idea consists in successively merging the two most similar centroids, according to a given metric such as the correlation coefficient for example:

$$\mathbf{C}_{i\cup j} = \frac{|\mathcal{C}_i|\mathbf{C}_i + |\mathcal{C}_j|\mathbf{C}_j}{|\mathcal{C}_i| + |\mathcal{C}_j|} \tag{22}$$

2) Attack Performance: The videos presented in Table I have been watermarked with the SS-N system using 4 alternative watermarks and an embedding strength α equal to 3. Next, the watermarked videos have been submitted to the WECR attack with and without an additional merging step. The detection score has been computed before and after the attack and the results have been gathered in Table II. The value in brackets indicates the detection score when a merging step is introduced. It is clear that the efficiency of the attack depends on the content of the video. The more dynamic the video content, the more different the individual watermark estimates and the more effective the watermark estimation refinement process. Furthermore, if the video contains long static shots, it can interfere with the splitting strategy and results in *bad* centroids i.e. which gathers video frames not carrying the same watermark pattern $\mathbf{W}_i$. Adding a merging step may then alter the efficiency of the attack (*ping-pong* video). In real life, an attacker would not use successive frames from a video, but would rather extract some key frames of the watermarked video. As an example, a TV news video with commercial breaks has been watermarked with the SS-N system and 325 key frames have been extracted to perform the WECR attack. In this case, almost 90% of the watermark signal has been properly estimated, which succeeds in lowering the correlation score from 2.91 to 0.52 (0.45) i.e. a score below the detection threshold.

TABLE II
IMPACT OF THE WECR ATTACK ON THE DETECTION SCORE OF THE SS-N SYSTEM

Video shot	*Before WECR attack*	*After WECR attack*
Ping-Pong	2.92	1.73 (3.24)
Ski	2.82	0.46 (0.45)
Susie	3.00	0.30 (0.27)
Train	2.89	0.70 (0.54)
Tree	2.37	1.63 (1.02)

V. EMBEDDING STRENGTH MODULATION

The SS-N system exploits watermark modulation to obtain superior performance against intra-video collusion attacks. However, an expert attacker can still remove the embedded watermark with an attack based on vector quantization. A new geometrical interpretation is consequently introduced in this section to obtain a novel perspective and thus a better understanding of the weaknesses of the previous watermarking schemes. From these observations, embedding strength modulation is explored to achieve security. Limitations of such an approach against hostile intelligence are also evaluated.

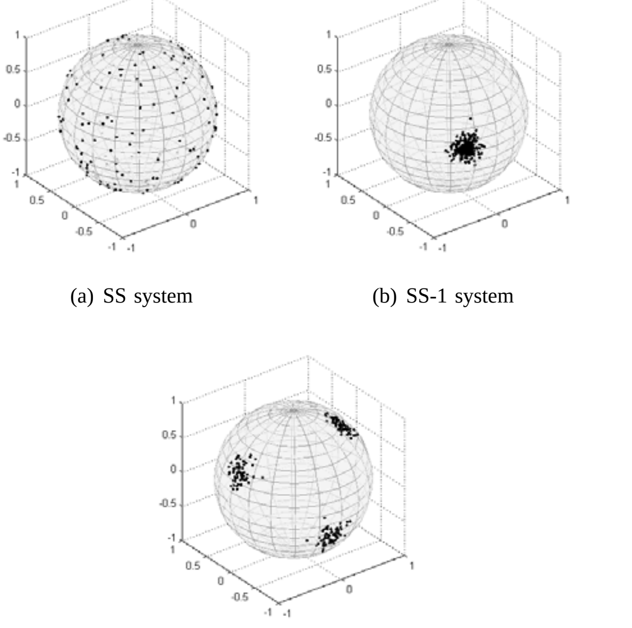
A. A Novel Perspective

The three video watermarking systems presented all embed a normally distributed watermark $\mathbf{W}_t$ with zero mean and unit variance in each frame $\mathbf{F}_t$ with a fixed embedding strength α:

$$\check{\mathbf{F}}_t = \mathbf{F}_t + \alpha\mathbf{W}_t(K), \quad \mathbf{W}_t(K)\sim\mathcal{N}(0,1) \tag{23}$$

The embedded watermark can be seen as a low-power pseudo-random image of P pixels which is scaled and added to a video frame. Alternatively, it can be considered as a disturbing random vector drawn from a P dimensional space which is added to a host vector. In this case, the norm of the first vector has to be far lower than the norm of the latter to fulfill the invisibility constraint. Since watermarks are zero mean, they are in fact drawn from a $(P-1)$ dimensional subspace. Furthermore, they are bound to lie on the unit sphere associated with the distance $\mathrm{d}(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})\cdot(\mathbf{x}-\mathbf{y})}$ as they have unit variance. Now, even if the presented watermarking schemes share a common framework, they enforce a different embedding strategy. This has a direct impact on how the different watermarks are distributed over the unit sphere as illustrated in Figure 5.

This geometric approach sheds a new light on the link between embedding strategies and security issues. When embedded watermarks are uniformly distributed over the unit sphere (SS system), averaging successive watermarks results then in a very small vector in the middle of the unit sphere i.e. there is very little residual watermark energy. Alternatively, when watermarks are gathered in a single narrow area (SS-1 system), or even several areas (SS-N system), the watermarks can be distributed amongst well-identified clusters. As a conclusion, successive watermarks define a trajectory over the unit sphere and this watermark trajectory should have some properties to resist intra-video collusion attacks. First it should be continuous so that averaging successive watermarks results



(a) SS system

(b) SS-1 system

(c) SS-N system (N=3)

Fig. 5. Distribution of the embedded watermarks over the unit sphere depending on the enforced watermarking strategy in a 3-dimensional watermarking subspace.

in a watermark near the surface of the unit sphere. Second, the trajectory should not have accumulation points to prevent weaknesses against WECR attacks.

B. SS-α System

The SS-N system relies on watermark modulation to achieve security. However, an alternative strategy exploiting the embedding strength can also be explored. The basic idea consists in using a time dependent embedding strength $\beta(t)$:

$$\check{\mathbf{F}}_t = \mathbf{F}_t + \alpha\beta(t)\mathbf{W}(K), \quad \mathbf{W}(K) \sim \mathcal{N}(0,1) \tag{24}$$

Here the embedding strength is modulated for security reasons and not to improve watermark invisibility as usual. With this end in view, the modulation function $\beta(t)$ has to respect the three following constraints:

(i) It should vary smoothly in time to be immune to TFA attacks,

(ii) It should be zero mean to resist a potential WER attacks,

(iii) It should have a large number of values after discrete sampling to avoid WECR attacks.

Keeping these specifications in mind, a set $\{\mathbf{W}_i\}$ of N orthonormal watermark patterns is built. The embedding procedure of the SS-α system is then defined as follows:

$$\check{\mathbf{F}}_t = \mathbf{F}_t + \alpha \sum_{i=1}^{N} \beta_i(t)\mathbf{W}_i = \mathbf{F}_t + \alpha\mathbf{W}_t \tag{25}$$

The modulation functions $\beta_i(t)$ have to be chosen in accordance with the precited specifications to achieve security. The SS-N system can indeed be seen as a specific case of this new system where the modulation functions are equal to $\beta_i(t) = \delta^i_{\Phi(t)}$. However, such modulation functions only give N possible combinations of watermarks and this system can be defeated by a WECR attack. An additional constraint is introduced so that embedded watermarks $\mathbf{W}_t$ all lie on the unit sphere. In other terms, the modulation functions should verify:

$$\forall t \quad \sum_{i=1}^{N} \beta_i^2(t) = \mathbf{W}_t \cdot \mathbf{W}_t = 1 \tag{26}$$

As a result, the embedding process introduces a Mean Square Error (MSE) equal to α^2 and an embedding strength α equal to 3 induces a distortion of about 38 dB. The detector computes the energy[4] contained in the subspace spanned by the watermark patterns $\mathbf{W}_i$:

$$\begin{aligned} \rho(\{\check{\mathbf{F}}_t\}) &= \sqrt{\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N}(\check{\mathbf{F}}_t \cdot \mathbf{W}_i)^2} \\ &\approx \alpha \sqrt{\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N}\beta_i^2(t)} = \alpha \end{aligned} \tag{27}$$

Host interference is cancelled in a preprocessing step [6] to enhance detection statistics. The detection score should be equal to α if a watermark is present in the video, while it should be almost equal to zero if no watermark has been inserted. It is consequently compared to a threshold τ_{detect}, which is set equal to $\alpha/2$ in practice, to assert the presence or absence of a hidden watermark.

1) Sinusoidal Modulation: A sinusoidal embedding strength [18] can be used to have a practical implementation of this strategy:

$$\beta_i(t) = \sqrt{\frac{2}{N}}\sin(\Omega t + \phi_i) \tag{28}$$

where Ω is a shared radial frequency and ϕ_i are phases to be set appropriately. From a communication perspective, this system can be considered as transmitting the same low-power temporal signal $\sin(\Omega t)$ along several non-interfering channels $\mathbf{W}_i$ with some phase differences ϕ_i. The square norm of the embedded watermarks $\mathbf{W}_t$ is then given by:

$$\begin{aligned} \mathbf{W}_t \cdot \mathbf{W}_t = 1 - &\frac{\cos(2\Omega t)}{N}\sum_{i=1}^{N}\cos(2\phi_i) \\ + &\frac{\sin(2\Omega t)}{N}\sum_{i=1}^{N}\sin(2\phi_i) \end{aligned} \tag{29}$$

The phase differences ϕ_i should be chosen so that both sums are equal to zero to fulfill (26). The N^{th} roots of unity in $\mathbb{C}$ can be taken into account and $2\phi_i = i2\pi/N$ modulo 2π. An ambiguity regarding the value of ϕ_i still remains, leaving room for embedding a moderate payload:

$$\phi_1 = 0, \quad \phi_i = \left(\frac{i}{N} + b_i\right)\pi \mod 2\pi \tag{30}$$

[4] As for the SS-N system, changing the detector has an impact on the detection statistics. Here again, the variance is increased and more frames need to be accumulated to obtain similar false positive and false negative probabilities than for SS or SS-1 systems

where $b_i \in \{0,1\}$ is a bit of payload. Since the detector will only be able to estimate phase differences, the phase ϕ_1 is set 0 to allow payload retrieval. The whole embedding process is depicted in Figure 6. On its side, the detector correlates each incoming video frame $\check{\mathbf{F}}_t$ with all the watermark patterns $\mathbf{W}_i$ to obtain an estimate $\tilde{\beta}_i(t) = \check{\mathbf{F}}_t \cdot \mathbf{W}_i$ of the temporal signal transmitted along each communication channel. Next, the detection score given in (27) is computed to assert whether an underlying watermark is present in the video or not. If a watermark is detected, the payload bits are extracted by estimating the phase differences ϕ_i. This can be easily done by computing the unbiased cross-correlation between the reference signal $\tilde{\beta}_1(t)$ and the other $\tilde{\beta}_i(t)$ whose phase difference encodes a payload bit b_i according to (30).

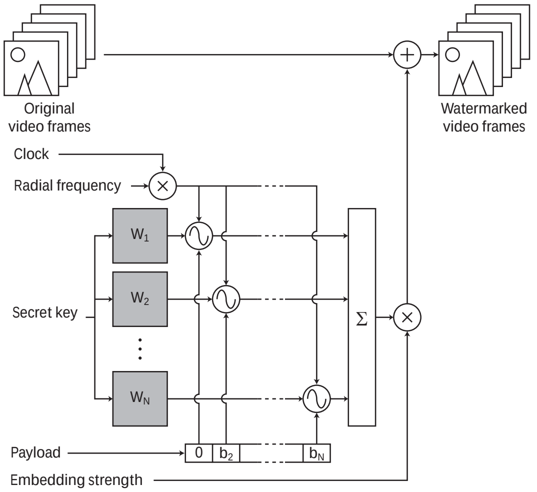


Fig. 6. SS-α system with sinusoidal modulation: the embedder inserts a linear combination of N reference watermark patterns, whose mixing coefficients are temporally sinusoidal.

2) Security Constraints: Even if this novel system has been designed to resist intra-video collusion attacks, some parameters need to be carefully chosen. First, the radial frequency Ω should remain secret or pseudo-secret to prevent an attacker from separating the watermarked video frames into distinct sets of frames carrying almost the same watermark signal[5]. Otherwise, a WER attack can then be successfully applied to each set. Now, if an attacker performs a WER attack on the whole video using the optimal watermark estimator $\Delta_{\text{o}}(.)$, the following watermark estimate is obtained:

$$\tilde{\mathbf{W}}_T = \sum_{i=1}^{N} \Big(\frac{\alpha}{T}\sum_{t=1}^{T}\beta_i(t)\Big)\mathbf{W}_i = \sum_{i=1}^{N}\lambda_i(T)\mathbf{W}_i \qquad (31)$$

The more video frames are considered, the closer the coefficients $\lambda_i(T)$ are to zero. Since the attacker does not have access to the optimal watermark estimator in practice, each watermark estimation is noisy and accumulating several watermark estimations decreases the power of the watermark signal. That is to say that combining several individual watermark estimates hampers the final watermark estimation, which is in complete contradiction with the paradigm behind the original attack. The same property can be demonstrated with non-adjacent video frames. The radial frequency Ω should also be set so that a given mixture of sinusoidal coefficients $\beta_i(t)$ is never used twice. It should consequently be selected from $\mathbb{R} - \pi\mathbb{Q}$ so that any WECR attack is then doomed to fail.

Alternatively, an attacker can perform a TFA attack and obtain the following attacked video frames:

$$\dot{\mathbf{F}}_t = \frac{1}{w}\sum_{u\in[-\frac{w}{2},\frac{w}{2}[}\mathbf{F}_{t+u} + \alpha\gamma_w\mathbf{W}_t, \quad \gamma_w = \frac{\text{sinc}(\frac{w\Omega}{2})}{\text{sinc}(\frac{\Omega}{2})} \qquad (32)$$

Regarding (25), TFA has basically scaled the embedded watermark signal by a signed attenuation factor γ_w. The larger the temporal window size w, the lower the attenuation factor. Similarly, the higher the radial frequency Ω, the closer the attenuation factor to zero. As a result, the radial frequency Ω should be chosen in such a way that the attenuation factor remains higher than a threshold value γ_{lim} as long as the temporal window size is lower than a given value w_{max}. If a larger window size is used, the content provider considers that the video has lost its commercial value due to the loss of visual quality. In other words, the parameters γ_{lim} and w_{max} give a higher bound for the radial frequency Ω so that TFA only results in a small attenuation of the hidden signal.

C. Watermarking Subspace Estimation Draining (WSED)

Embedded watermarks $\mathbf{W}_t$ are always a linear combination of a small number of reference watermark patterns $\mathbf{W}_i$ as written in (25). In other terms, embedded watermarks are restricted to a low dimensional watermarking subspace which can be estimated[6] using space dimension reduction techniques [19]. Having a collection of T individual watermark estimates of size P and knowing that the embedded watermarks are contained in a N-dimensional subspace ($N \ll P$), the attacker wants to find N vectors $\mathbf{E}_i$ which span the same subspace as the one generated by the secret patterns $\mathbf{W}_i$:

$$\mathcal{W} = \text{span}(\mathbf{W}_i) = \text{span}(\mathbf{E}_i) = \mathcal{E} \qquad (33)$$

With this end in view, Principal Component Analysis (PCA) can be performed since it is an optimal dimension reduction technique. Let $\tilde{\mathbf{W}}$ be a $P \times T$ matrix whose columns are the individual watermark estimates $\tilde{\mathbf{W}}_t$. The goal is to find a $P\times N$ matrix $\mathbf{E}$ and a $N \times T$ matrix $\mathbf{V}$ which minimize the norm $\|\tilde{\mathbf{W}} - \mathbf{E}\mathbf{V}\|$. Each column of the matrix $\mathbf{V}$ can be viewed as the coordinates of the associated watermark estimate in the matrix $\tilde{\mathbf{W}}$ in the principal subspace spanned by the vectors defined by the columns of matrix $\mathbf{E}$.

[5] In fact, Ω can be estimated with a very simple temporal spectral estimation. This is a major security flaw for any watermarking system based on periodic watermark schedule. However this system is only presented for illustrative purpose. In the general case, this attack does not defeat the SS-α system.

[6] It should be noted that this estimation of the watermarking subspace can be exploited to enhance the previously described WECR attack. The watermark estimates $\tilde{\mathbf{W}}_t$ are projected onto the estimated subspace $\mathcal{E}$ prior to vector quantization. Once the coordinates of the clusters have been identified in the watermarking subspace, the centroids $\mathbf{C}_i$ can then be easily retrieved.

1) Attack Description: As standard methods for PCA require too much memory for high dimensional data, an approach based on the Expectation-Maximization (EM) algorithm is exploited [20]. The PCA procedure is then reduced to an iterative algorithm using two steps:

$$\text{E-step:} \quad \mathbf{V} = (\mathbf{E}^{\mathrm{T}}\mathbf{E})^{-1}\mathbf{E}^{\mathrm{T}}\tilde{\mathbf{W}} \tag{34a}$$

$$\text{M-step:} \quad \mathbf{E} = \tilde{\mathbf{W}}\mathbf{V}^{\mathrm{T}}(\mathbf{V}\mathbf{V}^{\mathrm{T}})^{-1} \tag{34b}$$

where $.^{\mathrm{T}}$ denotes the transposition operator. A major asset of this approach is that it can be performed *online* using only a single watermark estimate at a time, which significantly reduces storage requirements. Moreover, the EM framework supports missing data, i.e. non pertinent estimated samples. During the E-step, for each incomplete watermark estimate $\tilde{\mathbf{W}}_t$, the coordinates $\mathbf{V}_t$ in the current estimated subspace are computed using only valid samples and missing information is completed so that the distance to the current principal subspace is minimized. The completed watermark estimate $\tilde{\mathbf{W}}_t^*$ is then used for the M-step. After the PCA iterations, a N-dimensional subspace $\mathcal{E}$ has been estimated which is assumed to be close to the watermarking subspace $\mathcal{W}$. Thus it is drained from any energy:

$$\begin{aligned}\dot{\mathbf{F}}_t &= \check{\mathbf{F}}_t - \sum_{i=1}^{N}(\check{\mathbf{F}}_t \cdot \mathbf{E}_i)\mathbf{E}_i \\ &\approx \mathbf{F}_t + \sum_{i=1}^{N}\alpha_i(t)\Big(\mathbf{W}_i - \sum_{j=1}^{N}(\mathbf{W}_i \cdot \mathbf{E}_j)\mathbf{E}_j\Big)\end{aligned} \tag{35}$$

where $\{\mathbf{E}_i\}$ is an orthonormalized basis of the subspace $\mathcal{E}$ e.g. the eigenvectors of matrix $\mathbf{E}$. If the watermarking subspace $\mathcal{W}$ has been finely estimated, the terms $\mathbf{W}_i - \sum_{j=1}^{N}(\mathbf{W}_i \cdot \mathbf{E}_j)\mathbf{E}_j$ are null and the embedded watermark is removed.

2) Attack Performance: A TV news video with commercial breaks has been watermarked with the sinusoidal implementation of the SS-α system. An 8 bit payload has been hidden using $N = 9$ watermark patterns $\mathbf{W}_i$ and the embedding strength α has been set equal to 3. Previous experiments have shown that intra-video collusion attacks are more efficient when the several individual watermark estimates originate from video frames with uncorrelated contents. As a result, key frames of the watermarked video have been extracted and used to estimate the watermarking subspace $\mathcal{W}$. Eventually, all the frames of the watermarked video were drained of any energy contained in the estimated subspace $\mathcal{E}$. This WSED attack has reduced the detection score given in (27) from 2.96 to 0.53. In other terms, there is no longer enough watermark energy and the attack is a success. This result however has to be contrasted. First, for a given dimension N, the more watermarked video frames $\tilde{\mathbf{W}}_t$ are considered, the finer the estimated watermarking subspace and the more efficient the attack. Second, with a given number T of watermarked video frames, the greater the dimension of the watermarking subspace $\mathcal{W}$, the harder it is to estimate.

VI. Conclusion

Robustness is usually considered as a key-property for watermarking systems. However, it is only a first requirement when the technology is to be deployed in a hostile environment. In this case, malicious users will surely design some advanced attacks to defeat the system. The security issue has consequently to be addressed. When robustness ensures the survival of the watermark after blind attacks, security ensures its survival even if it is submitted to hostile *intelligent* attacks. Ideal systems are utopian in security and the goal is only to always make the task more difficult for an attacker. Thus, in this paper, a basic frame-by-frame embedding strategy has been improved step by step so that more sophisticated attacks are needed to defeat the system. All the proposed systems can be defeated as reminded in Table III but the attacks are also more and more complex. Now that security pitfalls have been identified, the introduced geometrical perspective gives some intuitive insight regarding which trajectory successive watermarks should follows. It should be continuous, without any accumulation point and should go all over the whole watermark space.

TABLE III
WATERMARK EMBEDDING STRATEGIES ASSOCIATED WITH THEIR DEDICATED INTRA-VIDEO COLLUSION ATTACK

Embedding strategy	*Collusion attack*
SS system	Temporal frame averaging
SS-1 system	Watermark estimation remodulation
SS-N system	Watermark estimations clusters remodulation
SS-α system	Watermarking subspace estimation draining

The previous theoretical statement does not give any clue on how such trajectories can be built in practice. All the watermarking systems presented can be labeled as *blind* as they do not in any way consider the data to be watermarked. Considering the host data may have a significant impact on performance and possible tracks for future work are given below:

(i) *Anchor-based watermarks:* Security is somewhat related to statistical invisibility [21]. In such an approach, two watermarks should be as similar as the associated host video frames. An implementation of this idea consists in embedding small watermark patches at some anchor locations of the video frames. These anchor points should be pseudo-secret, and also host signal dependent.

(ii) *Image signature:* Another approach to obtain such coherent watermarks exploits key-dependent image signatures [22], [23]. The goal is to obtain binary strings related with the host content i.e. image signatures should be as correlated as the associated images. They can then be used to generate a watermark pattern which degrades gracefully with an increased number of bit errors.

(iii) *Informed coding:* Recently, dirty paper codes [24], [25] have been explored to make the embedded watermark dependent on the host signal. Basically, for a given payload, a constellation of possible watermarks is defined on the unit sphere and the nearest watermark from the host signal is embedded. As a result, the induced watermark trajectory varies as smoothly as the host content and links several points of the constellation.

(iv) *Registration-based watermarks:* From an MPEG-4 perspective, frames of a video scene are several 2D projections of the same 3D movie set. Frame registration can consequently be exploited to combine several redundant areas, for instance the background, and thus produce unwatermarked content. A straightforward counterattack is then to simulate an *ideal world*, so that each 3D point of the scene, and thus its 2D projections, always carry the same watermark sample [26].

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers and Professor Ton Kalker for their helpful comments which improved the quality of this paper.

REFERENCES

[1] S. Katzenbeisser and F. Petitcolas, *Information Hiding: Techniques for Steganography and Digital Watermarking*. Artech House, 1999.

[2] I. Cox, M. Miller, and J. Bloom, *Digital Watermarking*. Morgan Kaufmann Publishers, 2001.

[3] G. Doërr and J.-L. Dugelay, “A guide tour of video watermarking,” *Signal Processing: Image Communication*, vol. 18, no. 4, pp. 263–282, April 2003.

[4] F. Hartung and B. Girod, “Watermarking of uncompressed and compressed video,” *Signal Processing*, vol. 66, no. 3, pp. 283–301, May 1998.

[5] W. Macy and M. Holliman, “Quality evaluation of watermarked video,” in *Security and Watermarking of Multimedia Contents II*, ser. Proceedings of SPIE, vol. 3971, January 2000, pp. 486–500.

[6] I. Cox and M. Miller, “Preprocessing media to facilitate later insertion of a watermark,” in *Proceedings of the International Conference on Digital Signal Processing*, vol. 1, July 2002, pp. 67–70.

[7] T. Kalker, G. Depovere, J. Haitsma, and M. Maes, “A video watermarking system for broadcast monitoring,” in *Security and Watermarking of Multimedia Contents*, ser. Proceedings of SPIE, vol. 3657, January 1999, pp. 103–112.

[8] T. Kalker, “Considerations on watermarking security,” in *Proceedings of the IEEE Workshop on Multimedia Signal Processing*, October 2001, pp. 201–206.

[9] H. Zhao, M. Wu, Z. Wang, and R. Liu, “Non-linear collusion attacks on independent fingerprints for multimedia,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. V, April 2003, pp. 664–667.

[10] D. Boneh and J. Shaw, “Collusion secure fingerprinting for digital data,” *IEEE Transaction on Information Theory*, vol. 44, no. 5, pp. 1897–1905, September 1998.

[11] M. Holliman, W. Macy, and M. Yeung, “Robust frame-dependent video watermarking,” in *Security and Watermarking of Multimedia Contents II*, ser. Proceedings of SPIE, vol. 3971, January 2000, pp. 186–197.

[12] G. Doërr and J.-L. Dugelay, “New intra-video collusion attack using mosaicing,” in *Proceedings of the IEEE International Conference on Multimedia and Expo*, vol. II, July 2003, pp. 505–508.

[13] C. Voloshynovskiy, S. Pereira, A. Herrigel, N. Baumgärtner, and T. Pun, “Generalized watermarking attack based on watermark estimation and perceptual remodulation,” in *Security and Watermarking of Multimedia Contents II*, ser. Proceedings of SPIE, vol. 3971, January 2000, pp. 358–370.

[14] E. Lin and E. Delp, “Temporal synchronization in video watermarking,” in *Security and Watermarking of Multimedia Contents IV*, ser. Proceedings of SPIE, vol. 4675, January 2002, pp. 478–490.

[15] G. Doërr and J.-L. Dugelay, “Switching between orthogonal watermarks for enhanced security against collusion in video,” Eurécom Institute, Tech. Rep. RR-03-080, July 2003. [Online]. Available: http://www.eurecom.fr/∼doerr

[16] Y. Linde, A. Buzo, and R. Gray, “An algorithm for vector quantizer design,” *IEEE Transactions on Communications*, vol. 28, no. 1, pp. 84–95, January 1980.

[17] A. Sankar, “Experiments with a gaussian merging-splitting algorithm for hmm training for speech recognition,” in *Proceedings of DARPA Speech Recognition Workshop*, February 1998, pp. 99–104.

[18] G. Doërr and J.-L. Dugelay, “Secure video watermarking via embedding strength modulation,” in *Proceedings of the Second International Workshop on Digital Watermarking*, ser. Lecture Notes in Computer Science, To be published.

[19] ——, “Danger of low-dimensional watermarking subspaces,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, To be published.

[20] S. Roweis, “EM algorithms for PCA and SPCA,” *Neural Information Processing Systems*, vol. 10, pp. 626–632, 1998.

[21] K. Su, D. Kundur, and D. Hatzinakos, “A novel approach to collusion resistant video watermarking,” in *Security and Watermarking of Multimedia Contents IV*, ser. Proceedings of SPIE, vol. 4675, January 2002, pp. 491–502.

[22] J. Fridrich and M. Goljan, “Robust hash functions for digital watermarking,” in *Proceedings of the International Conference on Information Technology: Coding and Computing*, March 2000, pp. 178–183.

[23] D. Delannay and B. Macq, “Method for hiding synchronization marks in scale and rotation resilient watermarking schemes,” in *Security and Watermarking of Multimedia Contents IV*, ser. Proceedings of SPIE, vol. 4675, January 2002, pp. 548–554.

[24] B. Chen and G. Wornell, “Dither modulation: A new approach to digital watermarking and information embedding,” in *Security and Watermarking of Multimedia Contents*, ser. Proceedings of SPIE, vol. 3657, January 1999, pp. 342–353.

[25] M. Miller, G. Doërr, and I. Cox, “Applying informed coding and informed embedding to design a robust, high capacity watermark,” *IEEE Transactions on Image Processing*, To be published.

[26] G. Doërr and J.-L. Dugelay, “Secure background watermarking based on video mosaicing,” in *Security, Steganography and Watermarking of Multimedia Contents VI*, ser. Proceedings of SPIE, vol. 5306, To be published.

Gwenaël Doërr received in 2001 the telecommunications engineering degree from Institut National des Télécommunications (Télécom INT), Evry, France and the M.S. degree in computer science from Université de Nice Sophia-Antipolis (UNSA), Sophia-Antipolis, France. He was an intern at NEC Research Institute, Princeton, NJ from April to September 2001, winning the Louis Leprince Ringuet Award for his work on trellis dirty paper codes. He then started a Ph.D. thesis at the Eurécom Institute, Sophia-Antipolis, France on video watermarking. His research interests currently include security against collusion, image hashing, video mosaicing, hostile attacks and turbo codes.

Jean-Luc Dugelay (Ph.D. 92, IEEE M’94-SM’02) joined the Eurécom Institute (Sophia Antipolis, France) in 1992. He is currently a Professor in the Department of Multimedia Communications and is in charge of the Image and Video Group for Multimedia Communications and Applications. His research interests include security imaging (watermarking and biometrics), image/video coding, facial image analysis, face cloning and talking heads. He has published over 80 technical papers and holds three international patents. He has in particular contributed to the first book on digital watermarking [1] and has given several tutorials on this topic co-authored with F. Petitcolas (Microsoft Research, Cambridge, England). In addition to national French projects, his group is involved in the European Network of Excellence *E-Crypt*. He is also serving as a Consultant in digital watermarking for France Télécom R&D and STMicroelectronics. He is an Associate Editor for the IEEE Transactions on Multimedia, the IEEE Transactions on Image Processing, the EURASIP Journal on Applied Signal Processing and the Kluwer Multimedia Tools and Applications.