Secure Video Watermarking Via Embedding Strength Modulation

Gwenaël Doërr and Jean-Luc Dugelay

Eurécom Institute
Department of Multimedia Communications
2229, route des Crêtes BP 193
06904 Sophia-Antipolis Cédex, FRANCE
{doerr,dugelay}@eurecom.fr
http://www.eurecom.fr/∼image

Abstract. Straightforward adaptations of results for still images watermarking have led to non-secure video watermarking algorithms. The very specific nature of digital video has indeed to be considered so that robustness and security issues are handled efficiently. As a result, a novel video watermarking scheme is presented in this paper: security is achieved by using a smooth time-dependent strength and payload is encoded in the phase difference between several signals transmitted along non-interfering communication channels. Moreover, temporal synchronization can be done in a blind manner on the detector side. The proposed scheme is finally proven to be secure against traditional intra-video collusion attacks and robust against MPEG compression.

1 Introduction

Digital watermarking has been introduced in the 90's as a complementary technology to protect digital multimedia data along its lifetime. Protecting digital data is necessary since it can be copied rapidly, perfectly, at a large scale and without any limitation on the number of copies. Consequently, encryption is usually enforced to render the data useless for people not having the correct decryption key. Nevertheless, encrypted digital data has to be decrypted sooner or later to be finally presented to a human observer/listener. In others terms, encryption protects digital data along its transport but this protection falls during content presentation. As a result, digital watermarking comes as a second line of defense to fill this *analog gap.* It basically embeds a secret invisible and robust watermark, which should be closely tied to the data so that it survives Digital/Analog conversion. This hidden signal encodes a message related to the targeted application: rights associated with the data for copyright protection, client signature for traitor tracing, data signature for authentication. There exists a complex trade-off between several conflicting parameters (*visibility*, *payload*, *robustness*, *security*) and a compromise has to be found which is often tied to the targeted application. The fresh watermarker is redirected towards existing books [1, 2] for further insight regarding those issues.

If digital watermarking has been mostly devoted to still images at the beginning, watermarking other types of multimedia data is now being investigated and digital video is one of those *new objects* of interest. Many applications can indeed benefit from digital watermarking in the context of video [3]. Cinema studios own very high valued video films. However, disseminating them is highly hazardous since released videos are then likely to be freely exchanged on popular peer-to-peer networks, leading thus to a drastic loss of royalties for the majors. Large amounts of money are at stakes and security mechanisms have to be introduced to safeguard the rights of the copyright owners. Digital watermarking has consequently been evocated to enforce copy and playback control [4] in the Digital Versatile Disk (DVD). The upcoming introduction of the digital cinema format also raises some concerns, in particular regarding camcorder capture of the screen [5, 6]. As a result, it has been proposed to embed a watermark during show time to identify the cinema and the presentation date and time to be able to trace back the source of the leak in the distribution network. In the same fashion, digital watermarking can be inserted in Pay-Per-View (PPV) and Video-On-Demand (VOD) frameworks [7]. Thus, when an illegal copy is found, the customer who has broken his/her license agreement can be identified and sanctioned. Digital watermarking can also be exploited for broadcast monitoring [8] i.e. to check that video items are effectively broadcasted during their associated booked air time.

To date, video watermarking is mainly considered as an extension of still image watermarking. Some algorithms address the specificities of a compression standard [9, 10] or embed a watermark in a three dimensional domain [11, 12]. Nevertheless, watermarking digital video content is still regarded most of the time as watermarking a sequence of images. Unfortunately, this straightforward adaptation has led to weak algorithms in terms of security [13, 14] i.e. resistance of the watermark against hostile intelligence. Depending on the specifications of the targeted application, such a weakness can be critical. A novel watermarking scheme is consequently presented in this article to address this issue. In Section 2 an original embedding strategy is proposed. It basically consists in encoding the payload in the phase difference between several signals transmitted along non-interfering communication channels. A self-synchronized detection procedure is then described in Section 3. The performances of the system are then evaluated in terms of security (intra-video collusion) and robustness (MPEG compression). Finally, conclusions are drawn and tracks for future work given in Section 5.

2 Watermark Embedding

Hartung and Girod [15] have described one of the pioneer video watermarking systems based on the Spread Spectrum theory [16]. In few words, a pseudo-random watermark encoding the payload is scaled by an embedding strength and added to the video signal. This approach is still used in recent video watermarking schemes. Either a different watermark is embedded in each video frame [17], or the same watermark is embedded in each video frame [8]. Unfortunately, both strategies have been shown to be weak against intra-video collusion attacks [13, 14]. As a result, a novel approach based on embedding strength modulation instead of watermark modulation is proposed in Subsection 2.1 so that the inserted watermark is immune against traditional collusion attacks. An application based on sinusoidal modulation is then presented in Subsection 2.2 and a discussion is conducted to show how multibit payload can be obtained.

2.1 Time-Dependent Embedding Strength

To date, video watermarking has mostly inherited from the results obtained for still images and many algorithms rely on the insertion of a spread-spectrum watermark in the luminance channel in a frame by frame fashion. Such approaches can be basically summarized with the following equation:

$$\check{F}_k = F_k + \alpha\, W_k \quad W_k \sim \mathcal{N}(0,1) \tag{1}$$

where F_k is the k^{th} video frame, $\check{F}_k$ its watermark version and α the embedding strength. The pseudo-random watermark W_k has a normal distribution with zero mean and unit variance and has been pseudo-randomly generated with a secret key K used as a seed. Perceptual shaping can be subsequently introduced to improve the invisibility of the watermark by making for example the embedding strength α dependent of the local content of the frame [18]. On the receiver side, a simple correlation-based detector permits to assert the presence or absence of the watermark.

Depending on the evolution of the embedded watermark W_k in time, two well-known systems can be obtained, each one having its strengths and weaknesses in terms of security. When a different watermark is inserted in each video frame [15, 17], averaging successive video frames spreads the watermark signal amongst neighbor frames, which makes the detector fail to detect the underlying hidden signal. On the other hand, if the same watermark is embedded in each video frame, it can be finely estimated and a simple remodulation removes the watermark signal [19]. Both approaches can be regarded as specific cases of a more general framework, where the embedder switches between P orthogonal watermarks [14]. The detector should then be slightly modified to obtain robust performances against traditional intra-video collusion attacks. Nevertheless, such a scheme is potentially weak against an attack, which combines watermark estimation remodulation and vector quantization.

The security issue is consequently not entirely solved and further investigations have to be conducted to securely embed a watermark in a video. Previous approaches basically rely on a modulation of the watermark W_k to achieve security. However, to the best knowledge of the authors, no study has been conducted which considers the temporal modulation of the embedding strength to achieve security as described by the following equation.

$$\check{F}_k = F_k + \alpha_k\, W \quad W \sim \mathcal{N}(0,1) \tag{2}$$

On the receiver side, each video frame is correlated with the fixed watermark W and the detector checks if the received temporal signal matches the expected α_k.

For security issues, the modulation law should respect some constraints. On one hand, α_k should be zero mean to be immune against the watermark estimation / remodulation attack. On the other hand, α_k should vary smoothly in time to be immune against temporal frame averaging. This approach is further discussed by considering a sinusoidal modulation law in the remaining of the article.

2.2 Achieving Payload with Multiple Sinusoids

For invisibility reasons, the watermarking process should introduce the same distortion in all the video frames, which will not be the case if a single sinusoid is embedded as given by Equation 2. A basic idea consists then in using several watermarks W_i to carry the same sinusoidal signal modulo a phase difference ϕ_i as written below:

$$\check{F}_k = F_k + \alpha \sum_{i=0}^{P-1} \sin\left(\frac{2\pi k}{T\tau} + \phi_i + \phi_r\right) W_i \tag{3}$$

where τ is the frame rate e.g. 25 frames/sec, T the sinusoid period in seconds, ϕ_r a random phase shared by all the W_i's and ϕ_i a phase specific to each W_i. The P watermarks are also orthonormalized with the Gram-Schmidt algorithm [20] to prevent cross-talk on the detector side i.e. $W_i \odot W_j = \delta_i^j$ if $\odot$ denotes the linear correlation and δ the Kronecker delta. The W_i's can be regarded as spatial carrier signals carrying the same temporal signal modulo a phase difference. In other terms, the same signal is transmitted along several non-interfering communication channels with a phase difference between them. The Mean Square Error (MSE) between a video frame of dimension $W \times H$ and its watermarked version is given by:

$$\begin{aligned}
MSE_k &= \frac{1}{WH} \sum_{x=1}^{W} \sum_{y=1}^{H} \Big[\check{F}_k(x,y) - F_k(x,y)\Big]^2 \\
&= \frac{\alpha^2}{WH} \sum_{x=1}^{W} \sum_{y=1}^{H} \Big[\sum_{i=0}^{P-1} \sin(2\pi k\rho + \phi_i + \phi_r)\, W_i(x,y)\Big]^2 \\
&= \alpha^2 \sum_{i=0}^{P-1} \sin^2(2\pi k\rho + \phi_i + \phi_r) \\
&= \frac{\alpha^2}{2} \Big[P - A(k) \sum_{i=0}^{P-1} \cos(2\phi_i) + B(k) \sum_{i=0}^{P-1} \sin(2\phi_i)\Big]
\end{aligned} \tag{4}$$

where $1/\rho = T\tau$ is the sinusoid period in number of frames, $A(k) = \cos(4\pi k\rho + 2\phi_r)$ and $B(k) = \sin(4\pi k\rho + 2\phi_r)$. In order to make the distortion independent of the temporal index k, it is necessary to chose the ϕ_i's so that both sums in Equation 4 are equal to zero and the P^{th} roots of unity in $\mathbb{C}$ are good candidates. As a result, the several ϕ_i's can be defined as follows:

$$\forall i \in [1, P-1] \qquad 2\phi_i = \frac{i2\pi}{P} \pmod{2\pi}$$

$$\forall i \in [1, P-1] \qquad \phi_i = \frac{i\pi}{P} \quad \text{or} \quad \phi_i = \left(\frac{i}{P} + 1\right)\pi \pmod{2\pi} \tag{5}$$

The problem is underconstrained i.e. for each watermark W_i, there are two alternatives to choose the associated phase ϕ_i. This ambiguity will be exploited in the remaining of the article to encode the payload. Depending on the binary value of the i^{th} bit b_i of the payload, a phase can be associated to the i^{th} watermark according to the following equation:

$$\phi_i = \left(\frac{i}{P} + b_i\right)\pi \qquad b_i \in \{0, 1\} \tag{6}$$

On the detector side, it will be necessary to estimate the ϕ_i's in a blind manner to obtain back the payload. In other terms, a temporal reference is required and the first sinusoid will be dedicated to that purpose. As a result, b_0 is set to 0 and it is then necessary to use $P+1$ watermarks to transmit a P bits payload. An example of the resulting mixture of sinusoids is shown in Figure 1.

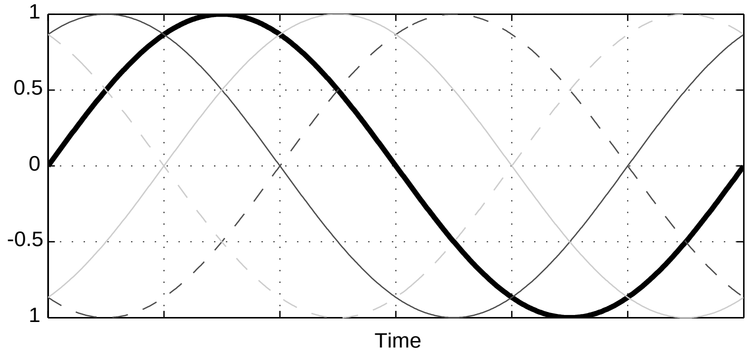

Fig. 1. Example of a mixture of sinusoids for a two bits payload. The bold line will be used for synchronization during detection. Dark (resp. light) gray lines suggest the two possible positions (plain and dash line) of the sinusoid associated with the first (resp. second) bit. Plain lines indicate the positions to encode the payload 01.

Once the ϕ_i's have been set as defined in Equation 6, the watermark embedding process introduces the same distortion in each video frame. The MSE is indeed equal to $\alpha^2 P/2$ according to Equation 4. This relation expresses the traditional trade-off in digital watermarking between payload, distortion and embedding strength, which is often related with robustness. For example, if the MSE is fixed to 9 and the payload to 16 bits, then the embedding strength is determined by the previous relation and should be around 1.

3 Watermark Retrieval

Once a content owner has embedded a secret digital watermark in a video, this later one can be transmitted over a digital network and delivered to customers.

If the content owner finds one day an illegal copy, he/she can check the presence of an underlying watermark and extract the payload to determine the rights associated with this video. This quite novel embedding strategy calls for a new detection procedure. The framework of the detector can be divided in three major modules. In a first step, the several temporally sinusoidal signals transmitted along non-interfering communication channels are extracted (Subsection 3.1). Next, the period of the reference sinusoid is estimated (Subsection 3.2). Finally, the phase differences between the different sinusoids and the reference one are estimated to evaluate the bit carried by each sinusoid (Subsection 3.3).

3.1 Parallel Signal Extraction

As previously pointed out, the embedding process can be regarded as the transmission of the same temporally sinusoidal signal, modulo some phase differences ϕ_i's, along several non-interfering communications channels, whose carrier signals can be assimilated to the watermarks W_i's. The first task of the detector is consequently to separate those several hidden temporal signals. This can be easily done by performing parallel linear correlations between the incoming video frames $\check{F}_k$ and the set of watermarks W_i as follows:

$$\begin{aligned}\beta_i(k) &= \check{F}_k \odot W_i \\ &= F_k \odot W_i + \alpha \sum_{j=0}^{P-1} \sin(2\pi k\rho + \phi_j + \phi_r)\, W_j \odot W_i \\ &= F_k \odot W_i + \alpha\, \sin(2\pi k\rho + \phi_i + \phi_r) \\ &\approx \alpha\, \sin(2\pi k\rho + \phi_i + \phi_r) \qquad (7)\end{aligned}$$

where $\beta_i(.)$ is the extracted temporal signal associated with the i^{th} watermark W_i. It should be noted that the orthonormalization of the set of watermarks has played a key role to simplify the previous equation. Moreover, it has been assumed that all the video frames have no correlation with each one of the watermarks ($F_k \odot W_i \approx 0$). Since this hypothesis is not necessarily true in practice, a preprocessing step [21] can be introduced before the embedding process which removes any interfering correlation from the original video frames. The obtained temporal signals β_i's are then normalized to have zero mean and a variance equal to 0.5 i.e. the average energy of a sinusoidal signal of amplitude 1 over a period. This normalization allows to compensate some alterations of the transmitted signals and the resulting normalized signals $\bar{\beta}_i(.)$'s should be almost equal to $\sin(2\pi k\rho + \phi_i + \phi_r)$. Thus, this first module outputs P normalized sinusoidal signals $\bar{\beta}_i(.)$ which only differ by some phase differences ϕ_i's. The next two modules will consequently be devoted to the estimation of those phase differences in order to extract the hidden payload.

3.2 Self-Synchronization

The estimation of the several phase differences ϕ_i's will rely on the unbiased cross-correlations between the reference signal $\bar{\beta}_0$ and the other ones. Since those correlations output phase differences in number of frames, it is necessary to estimate the period $1/\rho$ of the sinusoidal signals to get back the phase differences in radians. Even if the period T used during embedding is known on the detector side, it cannot be used directly since the video may have experienced temporal attacks e.g. small increase / decrease of the video speed. The period of the reference signal $\bar{\beta}_0$ should consequently be estimated and the autocorrelation $\bar{\beta}_0 \otimes \bar{\beta}_0$ of this signal is computed, with $\otimes$ the unbiased cross-correlation operator defined as follows:

$$f \otimes g\,(\delta) = \frac{1}{N - |\delta|} \sum_{n=\max(0,-\delta)}^{\min(N,N-\delta)} f(n)\,g(n+\delta) \qquad \delta \in \mathbb{Z} \tag{8}$$

where δ is a varying lag used in the cross-correlation and N the shared length of the signals f and g. Since the reference signal $\bar{\beta}_0$ is expected to be almost sinusoidal, basic trigonometric addition formulas insure that $\bar{\beta}_0 \otimes \bar{\beta}_0\,(\delta) \approx \cos(2\pi\delta\rho)/2$ i.e. the autocorrelation is sinusoidal and has the same period as the reference signal. The estimated period $1/\tilde{\rho}$ is then twice the average distance between two extrema. This estimation is performed on the autocorrelation instead of the extracted reference signal $\bar{\beta}_0$ because it is often far less noisy which facilitates extrema detection.

Such a procedure will always output an estimation of the period $1/\rho$ even if $\bar{\beta}_0$ is not sinusoidal or even periodic e.g. when a video is not watermarked. A matching criterion has consequently to be defined to determine if the extracted reference signal is effectively a sinusoid with a period $1/\tilde{\rho}$ or not. For example, one can compute the cross-correlation between $\bar{\beta}_0$ and a generated sinusoid of period $1/\tilde{\rho}$ with a lag δ varying between 0 and $\lceil 1/\tilde{\rho} \rceil$. The resulting signal is expected to be a period of a cosinusoid oscillating between -0.5 and 0.5. As a result, the maximum value M of this cross-correlation can be compared to a threshold ζ_{match} to assert whether the estimated sinusoidal signal matches the extracted reference one or not. If M is lower than ζ_{match}, the detector assumes that the estimated period $1/\tilde{\rho}$ does not match the periodicity of the extracted reference signal $\bar{\beta}_0$ and reports that no watermark has been detected. Otherwise, the detection procedure continues with the estimated period $1/\tilde{\rho}$. At this point, it should be noted that the detector can estimate the temporal distortions that the video has been subjected to, if it has access to the period $1/\rho$ used during the embedding process.

3.3 Payload Extraction

Once the period of the underlying sinusoidal signals has been estimated, the detector can then performed its final task, which is to estimate the phase differences ϕ_i between the different extracted $\bar{\beta}_i$'s and the reference $\bar{\beta}_o$ to estimate the payload bits b_i. The process will again rely on the unbiased cross-correlation operator. Since the extracted temporal signals are expected to be almost sinusoidal, trigonometric addition formulas insure that $\bar{\beta}_i \otimes \bar{\beta}_0(\delta) \approx \cos(2\pi\delta\rho - \phi_i)/2$.

As a result, the phase difference between the temporal signal carried by W_i and the one carried by W_0 can be estimated according to the following equation:

$$\tilde{\phi_i} = 2\pi\tilde{\rho}\ \arg \max_{\delta \in [0, \lceil 1/\tilde{\rho} \rceil]} \left(\bar{\beta}_i \otimes \bar{\beta}_0(\delta) \right) \tag{9}$$

This estimated $\tilde{\phi_i}$ is then compared with the only two possible phase differences for this specific communication channel given by Equation 6. The detector finally concludes that the bit, associated with the phase difference which is the nearest from the estimated one, has been embedded. This can be mathematically written as follows:

$$\tilde{b_i} = \arg \max_{b \in \{0,1\}} \left(\left| \tilde{\phi_i} - \left(\frac{i}{P} + b \right) \pi \right| \right) \tag{10}$$

As soon as the detector has asserted that the period $1/\rho$ has been correctly estimated, a sequence of bits is extracted in a blind manner, whatever the phase differences ϕ_i are. Whether the estimated $\tilde{\phi_i}$'s are near the phase differences given by Equation 6 or not, this will have no influence at all on the detector and the output result. Reliability measures $\mathcal{R}_i$ should consequently be introduced to indicate how confident the detector is for each estimated bit $\tilde{b_i}$. For example, the absolute difference $\varDelta_i$ between the estimated phase difference $\tilde{\phi_i}$ and the expected one $\left(\frac{i}{P} + \tilde{b_i} \right) \pi$ can be considered. When this difference $\varDelta_i$ is around 0, the estimated phase difference is really close to the expected one and the associated reliability should be very high. On the other hand, when $\varDelta_i$ is around $\pi/2$, the estimated phase difference is almost in the middle of the expected one and the bit $\tilde{b_i}$ should be regarded as unreliable. Several functions can be used to obtain such reliability measures e.g. a triangular function ($\mathcal{R}_i = 1 - 2\varDelta_i/\pi$) or a Hanning function ($\mathcal{R}_i = 0.5 + 0.5 \cos 2\varDelta_i$). Those reliability measures are then averaged to obtain a global reliability score $\mathcal{R}$, which is then then compared to a threshold $\zeta_{reliable}$ to determine if a message has been effectively embedded or not.

4 Performances

Watermarked videos experience various non hostile video processings when they are transmitted on a network: noise addition, frame filtering, chrominance resampling (4:4:4, 4:2:2, 4:2:0), lossy compression, transcoding, changes of spatio-temporal resolution (NTSC, PAL, SECAM), etc. Such processings can even be performed by content providers. Moreover, high-valued watermarked videos are also likely to be submitted to strong hostile attacks. Basically, several watermarked contents can be colluded to produce unprotected content [13, 14]. Collusion traditionally occurs when a clique of malicious customers gathers together to produce unwatermarked content. That is *inter-videos* collusion i.e. several watermarked video are required to produce unprotected content. Additionally, successive frames of a watermarked video can be regarded as several watermarked images. Thus, a single malicious user can collude several watermarked frames to produce an unprotected video. That is *intra-video* collusion i.e. a watermarked video alone permits to stir out the watermark signal from the video stream. This section will consequently be devoted to the evaluation of the performances of the presented video watermarking algorithm. In particular, Subsections 4.1 and 4.2 will focus on the security of the embedded watermark against two basic intra-video collusion attacks to demonstrate the superiority of the presented algorithm in comparison with previous ones [8, 15, 17]. Subsequently, the algorithm is also checked to be robust against moderate lossy compression with the popular MPEG standard in Subsection 4.3.

4.1 Temporal Frame Averaging

Digital watermarks are generally localized mostly in high frequencies since the Human Visual System (HVS) is less sensible to noise addition. As a result, one of the earliest proposed attacks to remove hidden watermarks is to apply a low-pass filter to the protected data [22]. Spatial filtering has been investigated extensively and most watermarking algorithms for still images are robust against it today. In the context of video, since neighbor video frames are highly similar, temporal low-pass filtering can be used to obtain an estimate of the original video frames i.e. without the underlying watermark. This can be written:

$$\dot{F}_k = \mathcal{L}_w(E_k), \quad E_k = \{F_{k+d}, -w \leq d \leq w\} \tag{11}$$

where w is half the size of the temporal window, $\mathcal{L}_w$ is the used temporal low-pass filter and $\dot{F}_k$ is the resulting k^{th} attacked video frame. In practice, a simple temporal averaging filter is often used even if non-linear operations can be performed [23]. Previous works [13, 14] have shown that such an attack succeeds in trapping video watermarking systems which always embed a different watermark in each video frame [15, 17]. This result has to be contrasted with the content of the video scene. Indeed, averaging several successive frames may result in a video of poor quality if fast moving objects are present in the scene or if there is a camera global motion. As a result, this attack is particularly relevant in static shots even if it can be adapted to cope with dynamic ones thanks to frame registration [24].

Now, if a video watermarked with the previously presented scheme is attacked, the following video frames are obtained:

$$\begin{aligned}
\dot{F}_k &= \frac{1}{2w+1}\sum_{d=-w}^{w} \check{F}_{k+d} \\
&= \frac{1}{2w+1}\sum_{d=-w}^{w} F_{k+d} + \frac{\alpha}{2w+1}\sum_{i=0}^{P-1} W_i \sum_{d=-w}^{w} \sin\Big(2\pi(k+d)\rho + \phi_i + \phi_r\Big) \\
&= \underline{F_k} + \alpha\gamma \sum_{i=0}^{P-1} \sin\Big(2\pi k\rho + \phi_i + \phi_r\Big) W_i
\end{aligned} \tag{12}$$

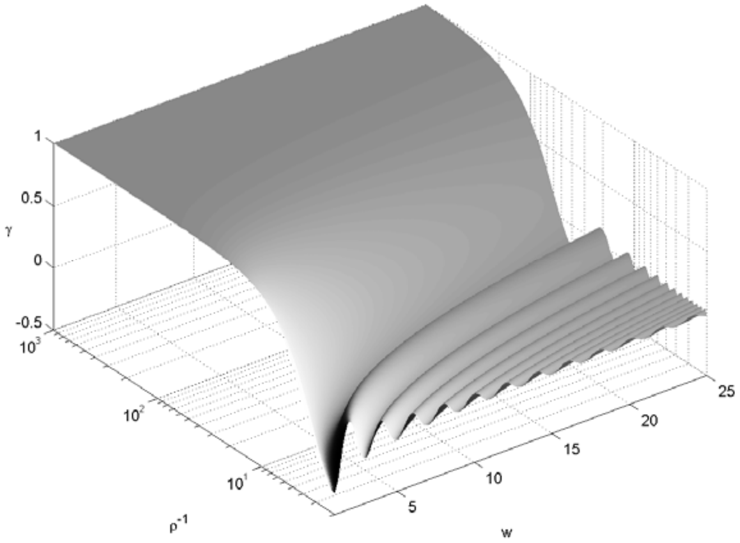

Fig. 2. Variations of the signed attenuation factor γ due to temporal frame averaging with the period of the sinusoid $1/\rho$ and w which is half the size of the temporal window.

where $\underline{F_k}$ is the k^{th} original video frame after temporal frame averaging and $\gamma = \text{sinc}\big((2w+1)\pi\rho\big)/\text{sinc}(\pi\rho)$. Regarding Equation 3, one can notice that temporal frame averaging has basically scaled the watermark signal by a factor γ. Since the absolute value of this scaling factor is always lower than 1, it can be regarded as a signed attenuation factor whose variations are depicted in Figure 2. For a given sinusoid period $1/\rho$, this attenuation factor decreases before oscillating around zero as the temporal window size increases. On the other hand, for a given temporal window size, γ decreases before oscillating around zero as the period $1/\rho$ of the sinusoid decreases. Such a behavior could have been predicted, since the values of the sinusoid are almost equal inside the temporal window when the period of the sinusoid is large. The period $1/\rho$ can consequently be chosen in such a way that the attenuation factor γ is always higher than a given value γ_{lim} as long as the temporal window size is below a given value w_{max}. For example, a content provider can estimate that the watermark is not required to survive if more than 11 successive frames are averaged ($w_{max} = 5$) since the resulting video has very poor visual quality. As a result, if the period $1/\rho$ is chosen to be greater than 30, then the attenuation factor γ is guaranteed to be always greater than 0.8. In other terms, the parameters w_{max} and γ_{lim} give a lower bound for the period of the sinusoid. Thus, temporal frame averaging only results in a relatively small attenuation of the hidden temporal signal and experiments have shown that the detector can counterbalance it thanks to the normalization.

4.2 Watermark Estimation Remodulation

When all the video frames carry the same watermark, the attacker can estimate the embedded watermark in each video frame and obtain a refined estimation of the watermark by combining (e.g. taking the average) those different estimations [13]. The ideal watermark estimator consists in computing the difference between a watermarked video frame and the associated original one. However, in practice, an attacker has not access to the original video frames and the watermark estimation process should be done in a blind manner. Previous work [19] has been done to estimate a watermark inserted in an image. As previously mentioned, a digital watermark is generally localized in high frequencies and a reasonable estimation can be obtained by computing the difference between a watermarked video frame and its low-pass filtered version[1]. As a result, the estimated watermark is given by:

$$\tilde{W_T} = \frac{1}{T} \sum_{n=1}^{T} \left[\check{F}_{\psi(n)} - \mathcal{L}(\check{F}_{\psi(n)}) \right] \tag{13}$$

where T is the number of combined watermark estimations and $\mathcal{L}(.)$ a spatial low-pass filter. The mapping function $\psi(.)$ indicates that the frames providing an estimation of the watermark are not necessarily adjacent. Once the embedded watermark has been estimated, it is subtracted from each watermarked video frame $\check{F}_t$ with a remodulation strength β. In practice, an attacker sets this remodulation strength so that the visual distortion introduced by the attack is equal to the one introduced by the watermark embedding process. As a result, the following attacked video frames are obtained:

$$\dot{F}_k = \check{F}_k - \sqrt{\frac{MSE_{embed}}{\tilde{W}_T \odot \tilde{W}_T}}\, \tilde{W}_T \tag{14}$$

Previous work [13, 14] has shown that this attack is particularly efficient against video watermarking schemes which always embed the same watermark [8]. Furthermore, the more the video frames are different, the more each individual watermark estimate refines the final one. In other terms, this attack is more efficient in dynamic scenes.

In the context of the watermarking scheme presented in Section 2, such an attack is doomed to fail since there is more than a single watermark to be estimated. Assuming that the attacker has access to the perfect watermark estimator $\check{F}_k - F_k$, the resulting estimated watermark will be a linear combination of the several W_i's as written below:

$$\tilde{W_T} = \sum_{i=1}^{P-1} \left(\frac{\alpha}{T} \sum_{n=1}^{T} \sin\left(2\pi\psi(n)\rho + \phi_i + \phi_r\right) \right) W_i = \sum_{i=1}^{P-1} \lambda_i(T) W_i \tag{15}$$

[1] In practice, some samples are badly estimated e.g. around the edges and in textured regions. An additional thresholding operation can consequently be performed to remove those non-pertinent samples.

If the frames providing the individual estimates are chosen randomly, then the coefficients $\lambda_i(T)$ are drawn from a truncated Gaussian distribution with zero mean and a variance equal to $\alpha^2/2T$. In other terms, the more video frames are considered, the more the λ_i's are close to zero. Since the attacker has not access to the perfect watermark estimator, each watermark estimation is noisy and accumulating several watermark estimations decreases the power of the watermark signal. Thus, combining several individual watermark estimates hampers the final estimation of the embedded watermark, which is in complete contradiction with the paradigm behind the original attack and experiments have shown that the embedded watermark is completely immune to the watermark estimation remodulation attack. Nevertheless parameters need to be chosen cautiously to prevent new security breaches. First the period $1/\rho$ of the sinusoid should remain secret or pseudo-secret. Otherwise the attacker would be able to separate the video frames in distinct sets of frames carrying the same watermark. Then he/she would only have to perform a simple watermark estimation remodulation attack on each set to remove the watermark. Moreover, the period $1/\rho$ should not be an integer. Otherwise, the attacker may be able to perform an attack based on watermark estimations clustering [14]. In the best case, $1/\rho$ should be chosen irrational ($\mathbb{R} - \mathbb{Q}$) so that a given mixture of sinusoidal coefficients is never used twice.

4.3 MPEG Compression

One hour of a video coded at 25 frames per second, with a frame size of 704×576 and pixels coded with 3 bytes, requires around 100 Gbytes for storage. In practical video storage and distribution systems, video sequences are consequently stored and transmitted in a compressed format. As a result, the behavior of the presented watermarking scheme against lossy compression, and in particular against MPEG-2 compression, has been investigated. Video sequences of 375 frames of size 704×576 at 25 frames per second have been watermarked with the embedding algorithm presented in Section 2 before being compressed with a freely available MPEG-2 encoder at 6 Mbits/s with a GOP of 12 (IBBPBBPBBPBBI) with default parameters. Figure 3 depicts one of the extracted sinusoids before and after the lossy compression in a given video sequence. First of all, it should be noted that the hidden sinusoid has been globally attenuated, which is a well-known behavior of spread-spectrum watermarks facing low-pass filtering or DCT coefficients quantization. However, this attenuation is stronger in dynamic scenes (frames 0-80 and 331-374) than in static shots (frames 80-330). This can be explained by the fact that more bits are allocated to motion vectors in dynamic scenes, thus reducing the number of bits allocated to details i.e. the hidden watermark located in high frequencies. Furthermore, it should be noted that the attenuation factor seems to be dependent of the MPEG frame type (I, P or B). In fact, the zoomed area reveals that the watermark signal is more attenuated in B frames than in P or I ones, which could have been expected since B frames are predicted from P frames, which are themselves predicted from I frames. In few words, MPEG compression basically more or less attenuates the hidden signal depending of the content of the scene with a factor which is dependent of the MPEG frame type. As a result, the *shape* of the hidden signal is slightly altered. Nevertheless, the detector estimates the several phase differences from the cross-correlated signal which is far much smoother and experiments have shown that the detector still succeeds in extracting the payload with a good confidence score (95%). Of course, the more periods of the sinusoid are considered for detection, the smoother are the cross-correlated signals and the more accurate is the detection.

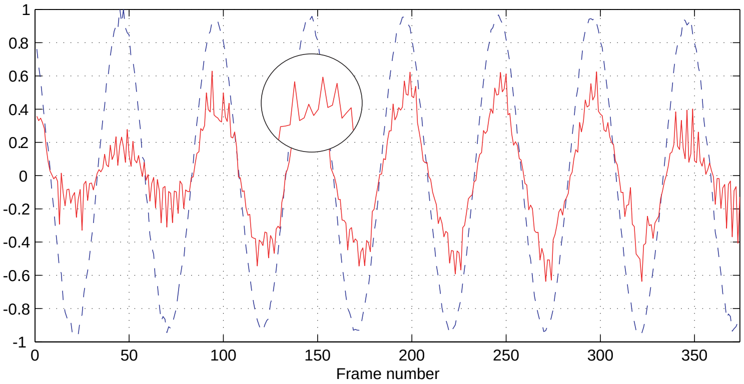

Fig. 3. Visualization of a given embedded sinusoid before (dashed line) and after (plain line) MPEG compression with the video sequence *pingpong*.

5 Conclusions and Perspectives

Digital video watermarking has mostly relied for the moment on direct adaptations of results for still image i.e. a video clip is regarded as a sequence of still images, which are watermarked with either an *embed always the same watermark* strategy, or an *embed always a different watermark* one. Unfortunately, such extensions have opened security breaches, in particular against collusion attacks. An innovative embedding strategy based on embedding strength modulation has consequently been presented in this paper to achieve security: the detector only considers a finite set of pseudo-random watermarks while the attacker views the video frames as carrying each one a different watermark. Furthermore, a moderate payload has been hidden in the video by introducing some phase differences between several non-interfering communication channels carrying the same temporal signal. Finally, this approach has been proven to be secure against traditional intra-video collusion attacks. It has also been shown to be robust against moderate MPEG compression and future work will evaluate the robustness of the algorithm against other non-hostile attacks.

Security against collusion has been notably enhanced with the proposed approach in comparison with previous systems. However, it can still be broken by an attacker with a higher level of expertise. First, the attacker can try to estimate the scene background, e.g. via video mosaicing [24], and then to generate a video similar to the original one from this estimated scene. Alternatively, it can be noticed that, for a given secret key, the inserted watermarks always lie in the same low dimensional subspace $\mathcal{W}$ generated by the W_i's. As a result, an attacker can estimate $\mathcal{W}$, e.g. by computing the PCA of the several individual watermark estimates obtained from each frame, and remove the part of the frame projected on this subspace. As a result, *informed coding* should be investigated to make the watermarks spread all over the watermarking space and not only a low dimensional subspace of it.

References

1. Katzenbeisser, S., Petitcolas, F.: Information Hiding: Techniques for Steganography and Digital Watermarking. Artech House (1999)
2. Cox, I., Miller, M., Bloom, J.: Digital Watermarking. Morgan Kaufmann Publishers (2001)
3. Doërr, G., Dugelay, J.-L.: A guide tour of video watermarking. Signal Processing: Image Communication **18**(4) (2003) 263–282
4. Bloom, J., Cox, I., Kalker, T., Linnartz, J.-P., Miller, M., Traw, C.: Copy protection for DVD video. Proceedings of the IEEE **87**(7) (1999) 1267–1276
5. Haitsma, J., Kalker, T.: A watermarking scheme for digital cinema. In: Proceedings of the IEEE International Conference on Image Processing. Volume 2. (2001) 487–489
6. Bloom, J.: Security and rights management in digital cinema. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. Volume IV. (2003) 712–715
7. Griwodz, C., Merkel, O., Dittmann, J., Steinmetz, R.: Protecting VoD the easier way. In: Proceedings of the ACM Multimedia Conference. (1998) 21–28
8. Kalker, T., Depovere, G., Haitsma, J., Maes, M.: A video watermarking system for broadcast monitoring. In: Proceedings of SPIE 3657, Security and Watermarking of Multimedia Contents. (1999) 103–112
9. Jordan, F., Kutter, M., Ebrahimi, T.: Proposal of a watermarking technique for hiding/retrieving data in compressed and decompressed video. In: JTC1/SC29/WG11 MPEG97/M2281, ISO/IEC (1997)
10. Langelaar, G., Lagendijk, R., Biemond, J.: Real-time labelling of MPEG-2 compressed video. Journal of Visual Communication and Image Representation **9**(4) (1998) 256–270
11. Deguillaume, F., Csurka, G., Ó Ruanaidh, J., Pun, T.: Robust 3D DFT video watermarking. In: Proceedings of SPIE 3657, Security and Watermarking of Multimedia Contents. (1999) 113–124
12. Swanson, M., Zhu, B., Tewfik, A.: Multiresolution scene-based video watermarking using perceptual models. IEEE Journal on Selected Areas in Communications **16**(4) (1998) 540–550
13. Su, K., Kundur, D., Hatzinakos, D.: A novel approach to collusion resistant video watermarking. In: Proceedings of SPIE 4675, Security and Watermarking of Multimedia Contents IV. (2002) 491–502

14. Doërr, G., Dugelay, J.-L.: Switching between orthogonal watermarks for enhanced security against collusion in video. Technical Report RR-03-080, Eurécom Institute (2003)
15. Hartung, F., Girod, B.: Watermarking of uncompressed and compressed video. Signal Processing **66**(3) (1998) 283–301
16. Pickholtz, R., Schilling, D., Millstein, L.: Theory of spread spectrum communications - a tutorial. IEEE Transactions on Communications **30**(5) (1982) 855–884
17. Mobasseri, B.: Exploring CDMA for watermarking of digital video. In: Proceedings of SPIE 3657, Security and Watermarking of Multimedia Contents. (1999) 96–102
18. Voloshynovskiy, S., Herrigel, A., Baumgärtner, N., Pun, T.: A stochastic approach to content adaptive digital image watermarking. In: Proceedings of the Third International Workshop on Information Hiding (LNCS 1768). (1999) 211–236
19. Voloshynovskiy, C., Pereira, S., Herrigel, A., Baumgärtner, N., Pun, T.: Generalized watermarking attack based on watermark estimation and perceptual remodulation. In: Proceedings of SPIE 3971, Security and Watermarking of Multimedia Contents II. (2000) 358–370
20. Cohen, H.: A Course in Computational Algebraic Number Theory. Springer-Verlag (1993)
21. Cox, I., Miller, M.: Preprocessing media to facilitate later insertion of a watermark. In: Proceedings of the International Conference on Digital Signal Processing. Volume 1. (2002) 67–70
22. Kutter, M., Petitcolas, F.: A fair benchmark for image watermarking systems. In: Proceedings of SPIE 3657, Security and Watermarking of Multimedia Contents. (1999) 226–239
23. Zhao, H., Wu, M., Wang, Z., Liu, K.: Non-linear collusion attacks on independent fingerprints for multimedia. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. Volume V. (2003) 664–667
24. Doërr, G., Dugelay, J.-L.: New intra-video collusion attack using mosaicing. In: Proceedings of the IEEE International Conference on Multimedia and Expo. Volume II. (2003) 505–508
25. Kutter, M.: Watermarking resisting to translation, rotation and scaling. In: Proceedings of SPIE 3528, Multimedia Systems and Applications. (1998) 423–431
26. Herrigel, A., Voloshynovskiy, S., Rytsar, Y.: The watermark template attack. In: Proceedings of SPIE 4314, Security and Watermarking of Multimedia Contents III. (2001) 394–405