Secure Background Watermarking Based on Video Mosaicing

Gwenaël Doërr and Jean-Luc Dugelay

Eurécom Institute
Multimedia Communications Department
2229 route des Crêtes, BP 193
06904 Sophia-Antipolis Cédex, FRANCE

ABSTRACT

Digital watermarking was introduced during the last decade as a complementary technology to protect digital multimedia data. Watermarking digital video material has already been studied, but it is still usually regarded as watermarking a sequence of still images. However, it is well-known that such straightforward frame-by-frame approaches result in low performance in terms of security. In particular, basic intra-video collusion attacks can easily defeat basic embedding strategies. In this paper, an extension of the simple temporal frame averaging attack will be presented, which basically considers frame registration to enlarge the averaging temporal window size. With this attack in mind, video processing, especially video mosaicing, will be considered to produce a temporally coherent watermark. In other words, an embedding strategy will be proposed which ensures that all the projections of a given 3D point in a movie set carry the same watermark sample along a video scene. Finally, there will be a discussion regarding the impact of this novel embedding strategy on different relevant parameters in digital watermarking e.g. capacity, visibility, robustness and security.

Keywords: Video watermarking, intra-video collusion, frame registration, video mosaicing

1. 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 easily and perfectly duplicated and rapidly redistributed on a large scale. Today, even non-hacker users can exchange copyrighted material via Peer-to-Peer networks and multimedia content providers have requested security mechanisms (copyright protection, data authentication, traitor tracing) 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, encrypted data has to be decrypted sooner or later to eventually be presented to a human user i.e. the encryption protection falls within media presentation. As a result, digital watermarking[1] was introduced in the 90's as a second line of defense to fill this *analog gap*.

Digital watermarking basically consists in 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 exists a complex trade-off between several conflicting parameters (*visibility*, *payload*, *robustness* and *security*) and a compromise has to be found which is often related to the targeted application. If digital watermarking has been mostly devoted to still images at the beginning, watermarking other types of multimedia data is now being explored and digital video is one of these *new objects* of interest.[2] Cinema studios own very high valued movies and are reluctant to disseminate them on a risky environment. Large amounts of money are at stakes and security mechanisms have to be introduced to safeguard the rights of copyright owners. Thus, digital watermarking is worth being introduced in many applications: verification watermarks for broadcast monitoring, fingerprinting watermarks in Pay-Per-View (PPV) and Video-on-Demand

Send correspondence to Professor Jean-Luc Dugelay: E-mail: dugelay@eurecom.fr, Switchboard: +33 4.93.00.26.41, Fax: +33 4.93.00.26.27

(VoD) frameworks, copy control watermarks in the Digital Versatile Disk (DVD), identification watermarks to manage theater screen capture.

To date, video watermarking mainly extends results previously obtained for still images. As a result, frame-by-frame approaches are commonly used. Unfortunately, such straightforward adaptations have led to weak algorithms in terms of security, in particular against intra-video collusion as reminded in Section 2. An extension of the simple temporal frame averaging attack is then presented in Section 3. It basically considers video frame registration to allow frame averaging with large temporal window size, even in dynamic scenes. To counter this attack, a new video watermarking architecture is introduced in Section 4. The goal is to introduce a watermark in the background of the scene which is coherent with the camera motion. With this end in view, video mosaicing is exploited so that all the projections of a given 3D point in a movie set carry the same watermark sample all along the video scene. The impact of this approach on traditional parameters in watermarking is discussed in Section 5 and directions for future work are finally proposed in Section 6.

2. FRAME-BY-FRAME WATERMARKING AND COLLUSION ISSUES

Some video watermarking algorithms exploit the specificities of a compression standard or embed a watermark in a three dimensional transform. However, video watermarking mostly extends results previously obtained for still images and, today, watermarking digital video content is regarded most of the time as watermarking a sequence of still images. Without any loss of generality, such a frame-by-frame approach will be illustrated below with a simple additive spread spectrum watermark:

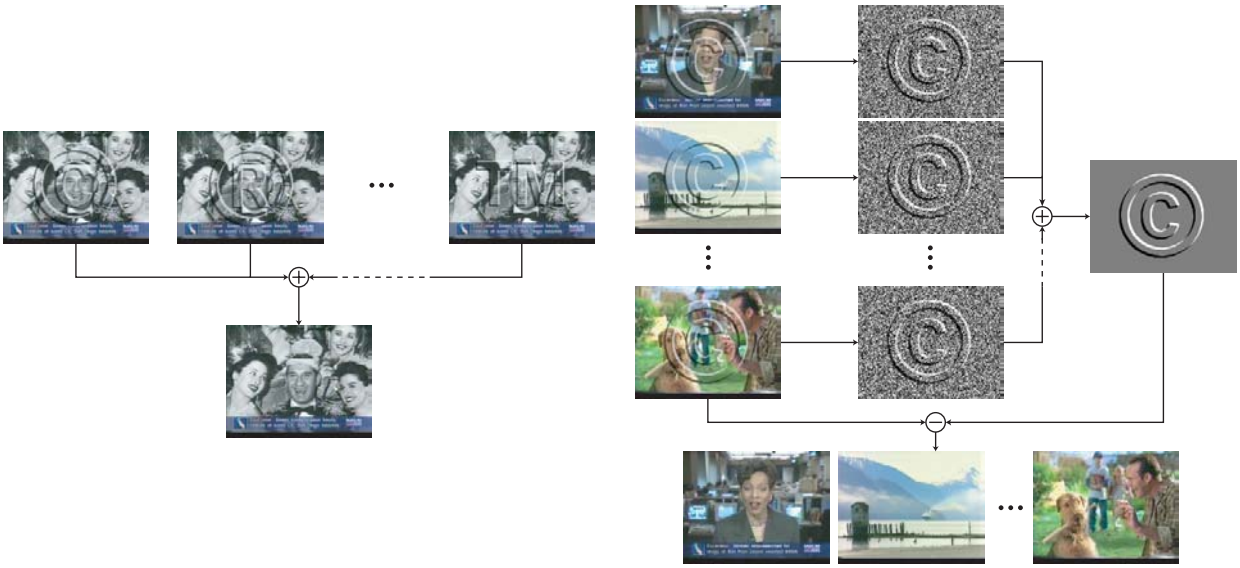
$$\check{\mathbf{F}}_t = \mathbf{F}_t + \alpha \mathbf{W}_t(K), \quad \mathbf{W}_t(K) \sim \mathcal{N}(0,1). \tag{1}$$

For each incoming original video frame $\mathbf{F}_t$, a watermark $\mathbf{W}_t(K)$ is generated using a secret key K and added with an embedding strength α to obtain the associated watermarked video frame $\check{\mathbf{F}}_t$. The watermark signal has a normal distribution with zero mean and unit variance. Perceptual shaping can subsequently be introduced to improve the invisibility of the watermark e.g. by making the embedding strength α dependent of the local variance of the video frame.[3] On the detector side, the presence or absence of a watermark is checked thanks to a correlation score, which can be computed using several successive video frames. When a frame-by-frame approach is enforced, two major embedding strategies are usually observed even if more complex ones can be explored.[4] Typically, either a *different watermark* is inserted in each video frame, or the *same watermark* is embedded in all the video frames. In other words, most of the current watermarking systems can be described thanks to Equation (1) with one of the following strategy:

1. *SS Strategy*: $\forall (t,t')\ \ t \neq t' \Rightarrow \mathbf{W}_t \neq \mathbf{W}_{t'}$ i.e. uncorrelated watermarks,[5]
2. *SS-1 Strategy*: $\forall t\ \ \mathbf{W}_t = \mathbf{W}$ i.e. fixed watermark.[6]

The main asset of such video watermarking systems is simplicity, which is tightly related with real-time processing. However it is also well-known that they result in poor performance in terms of security. This can be a major drawback depending on the targeted application.

In the watermarking context, security is defined as *the inability by unauthorized users to have access to the raw watermarking channel*.[7] In other words, it can be seen as the resistance of the hidden watermark against hostile intelligence. If applications such as broadcast monitoring, video authentication or data hiding do not have strong security requirements, this issue has to be addressed as soon as embedded watermarks are likely to be submitted to advanced hostile attacks e.g. fingerprinting or copy control watermarks. In particular, resistance to collusion attacks has to be considered. The goal of collusion attacks is to produce unwatermarked content by *combining* several watermarked contents. This can be regarded as eavesdropping the watermarking channel to identify some hidden properties and exploiting this knowledge to damage information transmitted on this secret communication channel. Collusion is crucial in digital video watermarking since each video frame can be seen as one watermarked document. As a result, examining a single watermarked video is enough to stir out the watermark signal from the whole video, which explains the term *intra-video* collusion. Previous work[8] has demonstrated that straightforward frame-by-frame systems can easily be defeated with simple collusion attacks. For example, Figure 1 depicts two attacks which can be applied to beat down a video watermarking system enforcing either SS or SS-1 strategy.



(a) Temporal Frame Averaging (TFA): Similar video frames carrying uncorrelated watermarks are averaged to produce unwatermarked content.

(b) Watermark Estimation Remodulation (WER): Several watermark estimations obtained from different video frames are combined to refine the estimation and allow watermark removal.

Figure 1. Visual illustration of traditional intra-video collusion attacks.

1. *Temporal Frame Averaging (TFA)*: Averaging uncorrelated watermarks generally converges toward zero. Furthermore, neighboring video frames are highly similar and can be averaged without introducing much visual distortion. Thus, temporal frame averaging has the following impact with video frames watermarked using the SS strategy:

$$\dot{\mathbf{F}}_t = \frac{1}{2w+1}\sum_{\delta=-w}^{w}\check{\mathbf{F}}_{t+\delta} = \frac{1}{2w+1}\Big[\sum_{\delta=-w}^{w}\mathbf{F}_{t+\delta} + \alpha\sum_{\delta=-w}^{w}\mathbf{W}_{t+\delta}\Big] \approx \frac{1}{2w+1}\sum_{\delta=-w}^{w}\mathbf{F}_{t+\delta} \approx \mathbf{F}_t, \tag{2}$$

where $\dot{\mathbf{F}}_t$ is the resulting attacked video frame and w the temporal window half-size. The larger the temporal window size is, the more attenuated is the watermark signal but the more distorted are the attacked video frames. In practice, this attack is particularly relevant in static scene when the SS strategy is enforced.

2. *Watermark Estimation Remodulation (WER)*: Digital watermarks are generally located in high frequencies. A rough estimation $\tilde{\mathbf{W}}_t$ of the watermark $\mathbf{W}_t$ embedded in the video frame $\mathbf{F}_t$ can consequently be obtained thanks to denoising techniques, or more simply by computing the difference between the watermarked frame $\check{\mathbf{F}}_t$ and its low-pass filtered version.[9] If the same watermark pattern $\mathbf{W}$ has been redundantly embedded using the SS-1 strategy, several individual watermark estimates can be combined, e.g. averaged, to further refine the estimation. Next, a simple remodulation allows stirring out the watermark signal as follows:

$$\dot{\mathbf{F}}_t = \check{\mathbf{F}}_t - \alpha\tilde{\mathbf{W}} = \mathbf{F}_t + \alpha(\mathbf{W} - \tilde{\mathbf{W}}) \approx \mathbf{F}_t, \tag{3}$$

where $\tilde{\mathbf{W}}$ is the refined watermark estimate. The more individual watermark estimates are combined, the finer is the watermark estimation. Furthermore, the more different are the considered video frames, the more efficient is the refinement process. As a result, this attack is pertinent in dynamic scene when the SS-1 strategy is enforced.

3. TEMPORAL FRAME AVERAGING AFTER REGISTRATION

A major shortcoming of temporal frame averaging is that it is limited by the content of the considered video. When the scene consists of dynamic content, e.g. fast moving object and/or camera motion, video frames cannot

be averaged without strongly degrading the video quality. If neighboring frames are highly correlated, they need to be registered to permit efficient averaging.[10] Each video frame is indeed a projection of a 3D movie set and different video frames from a shot can be seen as different 2D projections of the same scene. Thus, frame registration can be exploited to bring all these projections onto the same reference frame so that all the projections of a given 3D point overlap. As a result, temporal averaging can be done with a large temporal window without introducing much visual distortion. A detailed description of Temporal Frame Averaging after Registration (TFAR) is given below and the whole process is depicted in Figure 2.

The goal is to estimate a given video frame $\mathbf{F}_t$ from its neighboring ones $\mathbf{F}_{t+\delta}$ thanks to frame registration. However these frames may contain objects which cannot be used to reconstruct the target video frame. As a result, a binary mask $\mathbf{M}_t$ has to be built for each frame to distinguish useful areas in the frame (e.g. the background) from useless ones (e.g. moving objects). This mask is somewhat similar to the Video Object Plane (VOP) in the MPEG-4 video coding standard.[11] Once this mask has been defined, the background $\mathbf{B}_t$ and the moving objects $\mathbf{O}_t$ can be retrieved using the following equations:

$$\mathbf{O}_t = \mathbf{F}_t \otimes \mathbf{M}_t \quad \text{and} \quad \mathbf{B}_t = \mathbf{F}_t \otimes \bar{\mathbf{M}}_t, \tag{4}$$

where $\otimes$ is the pixel-wise multiplication operator and $\bar{\cdot}$ is the binary negation operator. No specific work has been done to design an object-based segmentation in this paper and an existing algorithm based on semi-automatic initial segmentation of the first video frame, followed by an automatic tracking of the selected objects[12] has been reused.

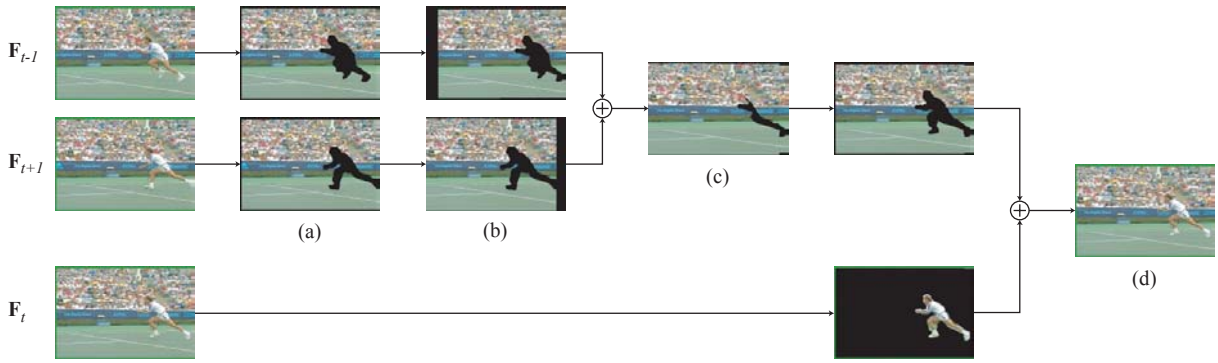


Figure 2. Temporal Frame Averaging after Registration (TFAR): Once the video objects have been removed (a), neighbor frames are registered (b) and combined to estimate the background of the current frame (c). Next, the missing video objects are inserted back (d). In this illustration, the temporal window half-size w is equal to 1.

Once several observations $\mathbf{B}_{t'}$ of the movie set have been obtained from neighboring frames, they can be exploited to estimate the background $\tilde{\mathbf{B}}_t$ of the current frame. To this end, it is necessary to find a registration function which *pertinently* associates to each pixel position (x_t, y_t) in the current frame $\mathbf{F}_t$ a position $(x_{t'}, y_{t'})$ in a neighboring frame $\mathbf{F}_{t'}$ i.e. which minimizes for example the mean square error between the target background $\mathbf{B}_t$ and the registered one $\mathbf{B}_{t'}^{(t)}$. In other words, the goal is to define a model which describes the apparent displacement generated by the camera motion. Physically, camera motion is a combination of traveling displacements (horizontal, vertical, forward and backward translations), rotations (pan, roll and tilt) and zooming effects (forward and backward). As the background of the scene is often far from the camera, pan and tilt rotations can be assimilated, for small rotations, to translations in terms of 2D apparent motion. Thus, the zoom, roll and traveling displacements can be represented, under some assumptions, by a first order polynomial motion model[13] as follows:

$$\begin{cases} x_{t'} = t_x + z(x_t - x_o) - z\theta(y_t - y_o) \\ y_{t'} = t_y + z(y_t - y_o) + z\theta(x_t - x_o) \end{cases}, \tag{5}$$

where z is the zoom factor, θ the 2D rotation angle, (t_x, t_y) the 2D translational vector and (x_o, y_o) the coordinates of the camera optical center. Obviously, this model is quite simple and may not be accurate when the camera displacement or the scene structure is very complicated. More complex motion representations can be introduced

such as the affine model,[13] the projection model[14] or the trifocal motion model.[15] Nevertheless, the model described in Equation (5) has been used in this paper for simplicity reasons.

The computed registered backgrounds $\mathbf{B}_{t+\delta}^{(t)}$, obtained from the video frames in the temporal window, are then combined to obtain an estimation $\tilde{\mathbf{B}}_t$ of the background in the current frame. For each pixel position p in the frame, the value of the background is estimated:

$$\tilde{\mathbf{B}}_t(p) = \begin{cases} \sum_{\delta \in [-w,w]^*} \mathbf{B}_{t+\delta}^{(t)}(p) \;/\; \sum_{\delta \in [-w,w]^*} \bar{\mathbf{M}}_{t+\delta}^{(t)}(p) & \text{if the denominator is not equal to 0} \\ 0 & \text{otherwise} \end{cases}, \quad (6)$$

where $\bar{\mathbf{M}}_{t'}^{(t)}$ the registered binary mask and w the temporal window half-size. In other words, the registered backgrounds are averaged using the proper normalization factor. A binary mask $\mathbf{R}_t$ is also built to indicate, for each pixel position, whether a background value has been effectively estimated ($\mathbf{R}_t(p) = 1$) or not ($\mathbf{R}_t(p) = 0$). The whole reconstruction process can then written as follows:

$$\dot{\mathbf{F}}_t = \underbrace{\tilde{\mathbf{B}}_t \otimes \bar{\mathbf{M}}_t}_{\text{Background}} + \underbrace{\mathbf{F}_t \otimes \mathbf{M}_t}_{\text{Objects}} + \underbrace{\mathbf{F}_t \otimes (\bar{\mathbf{M}}_t \;\&\; \bar{\mathbf{R}}_t)}_{\text{Missing pixels}}, \quad (7)$$

where & is the binary AND operator. The first term is associated with the current estimated background: pixel values have to be discarded if the related positions do not belong to the current background binary mask $\bar{\mathbf{M}}_t$. The second term indicates that moving video objects $\mathbf{O}_t$ from the original frame are inserted back. The last term in (7) points out that, at this point, some background pixels may have not been estimated. In this case, the pixel values from the original video frame $\mathbf{F}_t$ are retrieved. It should be noted that this attack does not affect the moving video objects $\mathbf{O}_t$. As a result, if such objects occupy most of the video scene, the attack is not likely to trap the detector. However, the background is usually the main part in many video shots and the attack is still pertinent. In particular, it is likely to defeat a system enforcing either SS, or SS-1 watermark embedding strategy.[10]

4. COHERENT BACKGROUND WATERMARKING

Skeptical people might argue that this attack is too intensive in terms of computation to be realistic. However, video mosaics or *sprite panoramas* are expected to be exploited for efficient background compression in the upcoming video standard MPEG-4.[11] Such video coding algorithms will have a similar impact on a potentially embedded watermark. As a result, this issue has to be addressed and, in particular, the reasons explaining the weakness of the previous embedding strategies have to be isolated. The results reminded in Section 2 basically recommend to watermark correlated video frames with the same watermark on one hand, and uncorrelated video frames with uncorrelated watermarks on the other one. These rules have subsequently been extended to give the following well-known fundamental embedding principle. *Watermarks embedded in distinct frames should be as correlated as the host video frames*, as written below:

$$\forall (t, t') \quad \rho(\mathbf{W}_t, \mathbf{W}_{t'}) \approx \rho(\mathbf{F}_t, \mathbf{F}_{t'}), \quad (8)$$

where $\rho(.)$ is a given correlation score, e.g. the correlation coefficient. Alternative approaches have been proposed to meet this specification e.g. the embedded watermark can be made frame-dependent,[16] a frame-dependent binary string can be exploited to generate a watermark pattern which degrades gracefully with an increased number of bit errors,[17,18] the watermark can be embedded in some frame-dependent positions.[19] However, is it enough to achieve security? The correlation score between an image and a shifted version of it may be very low. Nonetheless, the embedded watermarks should not be completely uncorrelated. In fact, the watermark embedded in the shifted image should also be a shifted version of the watermark embedded in the reference image. An additional mechanism has consequently to be introduced in watermarking embedding frameworks to ensure such a behavior.

4.1. Watermark Embedding Exploiting Video Mosaicing

For a given scene, backgrounds of video frames can be considered as several 2D projections of the same 3D set. The weakness of SS and SS-1 embedding strategies against Temporal Frame Averaging after Registration is due to the fact that camera motion is not considered at all. They are completely *blind*. As a result, a given 3D point which is projected in different locations in different video frames is associated with uncorrelated watermark samples. Thus, averaging registered video frames succeeds in confusing the watermark detector. The goal is consequently to inform the embedder about camera motion and to find an embedding strategy which forces each 3D point to carry the same watermark sample whenever it is visible in the video scene. In other terms, the basic idea is to simulate an utopian world where the movie set would already be watermarked.

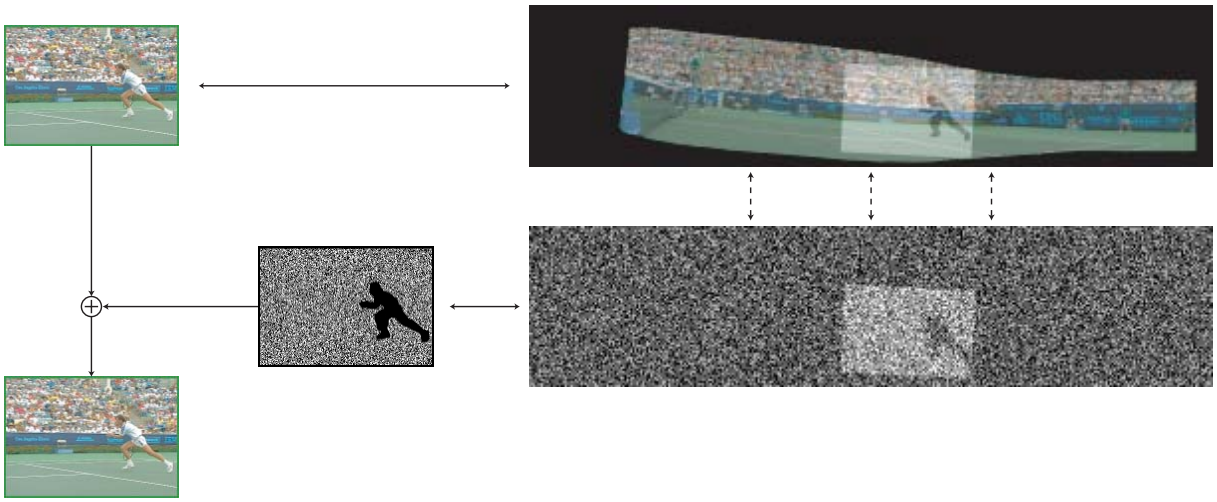


Figure 3. Embedding procedure for camera motion coherent watermarking (SS-Reg): The part of the watermark pattern which is associated with the current video frame is retrieved and registered back. Next, it is embedded in the background portion of the video frame.

Video mosaicing[20] consists in aligning all the frames of a video sequence to a fixed coordinate system. The resulting mosaic image provides a snapshot view of the sequence i.e. an estimation of the background of the scene if the moving objects have been removed. A straightforward and naive approach would consist in watermarking this video mosaic and to use it as the background of the video frames. However, such a process is likely to introduce some visible artifacts due to double interpolation and an alternative but somewhat equivalent approach is proposed in this article as depicted in Figure 3. First of all, warping parameters are computed for each video frames. Hence, each frame $\mathbf{F}_t$ is associated with a set of warping parameters (θ, z, (x_o, y_o) and (t_x, t_y) for the motion model defined in Equation (5)) i.e. the frame background $\mathbf{B}_t$ is associated with a portion $\mathbf{B}_t^{(m)}$ of the video mosaic. Next, a key-dependent watermark $\mathbf{W}$ is generated which has the same the dimensions as the mosaic representation of the video shot. Now, a portion $\mathbf{W}_t$ of the watermark can be associated to each video frame using the same warping parameters as for the mosaic. Finally, this watermark portion only has to be registered back to obtain the watermark pattern $\mathbf{W}_t^{(t)}$ to be embedded in the video frame. The overall embedding process can consequently be written as follows:

$$\check{\mathbf{F}}_t = \mathbf{F}_t + \alpha \bar{\mathbf{M}}_t \otimes \mathbf{W}_t^{(t)}, \quad \mathbf{W} \sim \mathcal{N}(0,1). \tag{9}$$

Again, perceptual shaping can be introduced to make the embedded watermark less noticeable. This novel embedding strategy will be referred as the *SS-Reg* strategy in the remainder of this paper. It should be noted that moving video objects $\mathbf{O}_t$ are left unprotected, which is due to the pixel-wise multiplication by the binary mask $\bar{\mathbf{M}}_t$. This operation can be removed and the watermark pattern $\mathbf{W}_t^{(t)}$ spread over the whole video frame. However, this would contradict the underlying philosophy of this strategy: *a 3D point carries the same watermark sample all along the video scene*. As a result, alternative approaches have to be inserted to protect these objects

if needed. Previous works have watermarked MPEG-4 video objects according to their main directions,[21] their animation parameters[22] or their texture.[23]

4.2. Watermark Detection

On the detector side, the procedure is very similar to the embedding one. In a first step, warping parameters are computed for each frames of the video scene to be checked and the watermark $\mathbf{W}$ is generated using the shared secret key. Next, the detector only checks if the portion $\mathbf{W}_t$ associated with each incoming frame $\tilde{\mathbf{F}}_t$ has been effectively embedded in the background. This can be done using the following correlation score:

$$\rho(\tilde{\mathbf{F}}_t, \mathbf{W}) = \frac{\tilde{\mathbf{F}}_t \cdot \mathbf{W}_t^{(t)}}{m_t} \approx \frac{\epsilon\alpha}{m_t}\left(\bar{\mathbf{M}}_t \otimes \mathbf{W}_t^{(t)}\right) \cdot \mathbf{W}_t^{(t)} = \epsilon\alpha, \tag{10}$$

where $\cdot$ denotes the linear correlation, ϵ equals 0 or 1 depending whether the video is watermarked or not and m_t the percentage of pixels contained in the background of frame $\tilde{\mathbf{F}}_t$. A preprocessing step[24] can be added to remove host interference in Equation (10) and thus improve the detection statistics. The proposed correlation score should then be equal to α if a watermark is present in the video frame, 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 in order to assert the presence or absence of the watermark. The value given to this detection threshold determines the false positive and false negative probabilities and the value $\alpha/2$ can be chosen for equal false positive and false negative probabilities. In practice, successive video frames are exploited to establish if a watermark is embedded in a video sequence and different correlation scores are accumulated as follows:

$$P_w(\tilde{\mathbf{F}}_t, \mathbf{W}) = \frac{1}{2w+1} \sum_{\delta=-w}^{w} \rho(\tilde{\mathbf{F}}_{t+\delta}, \mathbf{W}) \approx \epsilon\alpha. \tag{11}$$

It should be noted that, when the temporal window covers the whole video sequence, the detection procedure is equivalent to build the video mosaic of the scene and to compute the linear correlation with the watermark pattern $\mathbf{W}$. Considering many frames is commonly used[6] to enhance detection statistics. Indeed, some video processing, such as linear filtering, noise addition or lossy compression, are likely to introduce an interfering term in Equation (10). As a result, the correlation score is equal to $\epsilon\alpha + n$, where n can be considered as normally distributed with zero mean and variance σ. This has a direct impact on the false positive and false negative probabilities. Accumulating successive scores as in Equation (11) allows to reduce the effect of the interfering term n since it divides its variance by a factor $\sqrt{2w+1}$.

5. DISCUSSION

The novelty of the proposed embedding strategy lies in the fact that camera motion is compensated before embedding the watermark. To the best knowledge of the authors, such an approach has not been proposed in the literature yet. The most similar approach could be the SLIDE algorithm[19]: small watermark patches are embedded at some image dependent anchor locations. One can expect that these anchor points remain the same from a 3D point of view all along a video sequence and thus be coherent with camera motion. However, tracking of anchor point has not been explicitly addressed in that paper. The remainder of the article will consequently examine different outcomes, regarding some important properties in digital watermarking, due to this novel watermarking strategy.

5.1. Enhanced Security

In Section 4, the very first motivation for considering motion compensation before embedding was to enhance performances in terms of security. As a result, resilience against temporal frame averaging after registration has to be verified to demonstrate the superiority of the new embedding strategy. To this end, the video sequence *Stefan* has been watermarked using the three introduced embedding strategies: SS, SS-1 and SS-Reg. The embedding strength has been set to 3 so that the embedding process introduces a distortion around 38 dB in

terms of Peak Signal to Noise Ratio (PSNR). Furthermore, for a fair comparison between all three schemes, the watermark has been embedded in the background area as follows:

$$\check{\mathbf{F}}_t = \mathbf{F}_t + \alpha \bar{\mathbf{M}}_t \otimes \mathbf{W}_t, \tag{12}$$

where $\mathbf{W}_t = \mathbf{W}(K+t)$ for SS strategy, $\mathbf{W}_t = \mathbf{W}(K)$ for SS-1 strategy and $\mathbf{W}_t = \mathbf{W}_t^{(t)}$ for SS-Reg strategy. On the detector side, the presence or absence of the watermark is checked only in the background of the video frames using the following correlation score:

$$\rho(\tilde{\mathbf{F}}_t, \mathbf{W}_t) = \frac{\tilde{\mathbf{F}}_t \cdot \mathbf{W}_t}{m_t}. \tag{13}$$

The final score is obtained by averaging the correlation scores obtained for five successive video frames according to Equation (11). Once the detection score has been computed for all three watermarked videos, temporal frame averaging after registration is performed for each video using a window half-size $w = 1$ and the detection score computed once again. The ratio between the detection score after and before attack is then calculated and the results are depicted in Figure 4. A bold horizontal line has been drawn to illustrate the fact that the detector

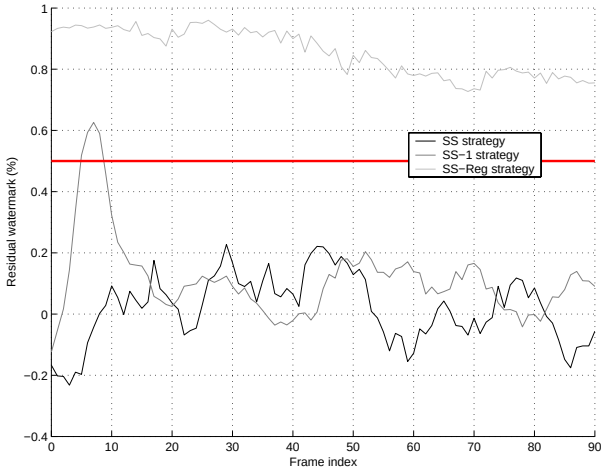


Figure 4. Resilience against Temporal Frame Averaging after Registration: Correlation scores before and after TFAR have been computed for each watermarking strategy. Their ratio is then calculated to evaluate the impact of the attack.

needs to retrieve at least 50% of the embedded signal to detect the watermark. If the plot corresponding to an embedding strategy is above this line, the watermark is detected in the video. Otherwise, the detector reports that no watermark has been found. It is clear that the SS-Reg strategy outperforms both SS and SS-1 embedding schemes. It should be noted that there is a peak around the 8^{th} video frame when SS-1 strategy is enforced. At this moment, there is indeed almost no camera motion and the SS-1 strategy is optimal. Performances with SS-Reg appear to degrade as the frame index is increasing. After examination, it seems to be related with variations of the zoom factor z. This issue will be further explored in future work to improve the efficiency of the proposed SS-Reg embedding strategy.

5.2. Video Capacity

The presented motion compensated video watermarking scheme has a zero bit capacity. It only gives an answer to the question: *is there a portion of the watermark* $\boldsymbol{W}$ *in each video frame*? However, it should be possible to modify the embedding strategy so that some payload can be hidden in a video scene. In comparison with still images, a video sequence provides a larger number of digital samples which can be exploited to carry some hidden information. A common mistake consists then in asserting that a greater payload can be embedded. Such a claim is true if there is no security requirement. For example, digital watermarking can be used for data hiding i.e. to embed some additional useful information in an invisible way. However, if the targeted application

includes strong security specifications (copy control, fingerprinting), advanced hostile attacks such as temporal frame averaging after registration are likely to occur and have to be addressed. As a result, the embedding strategy has to ensure that a given 3D point of the movie set always carries the same watermark sample in a video sequence. It is somewhat related with statistical invisibility introduced in previous work.[19] The proposed SS-Reg embedding strategy gives then some intuitive insight on how many bits can be *securely* embedded in a video sequence. Looking at Figure 3, the embedding procedure can be regarded as inserting a watermark in the mosaic representation of the video shot and subsequently exploiting this watermarked mosaic to replace the background in each video frame. In other words, the capacity is related with the dimensions of the mosaic i.e. with camera motion. If the camera is static, the mosaic image has the same dimensions as a video frame and a moderate payload can be embedded. On the other hand, as soon as the camera moves, new areas are revealed and they can be used to hide a larger payload.

5.3. Watermark Visibility

Evaluating the impact of distorting a signal as perceived by a human user is a great challenge. The amount and perceptibility of distortions, such as those introduced by lossy compression or digital watermarking, are indeed tightly related to the actual signal content. This has motivated the modeling of the human perception system to design efficient metrics. For example, when considering an image, it is now admitted that a low-frequency watermark is more visible than a high-frequency one or that a watermark is more noticeable in a flat area than in a texture one. The knowledge of such a behavior can then be exploited to perform efficient perceptual shaping. In the context of video, the Video Quality Experts Group (VQEG)[25] was formed in 1997 to devise objective methods for predicting video image quality. In 1999, they stated that no objective measurement system at test was able to replace subjective testing and that no objective model outperforms the others in all case. This explains while the Peak Signal to Noise Ratio is still the most often used metric today to evaluate the visibility of a video watermark. However, from a subjective point of view, previous works[26,27] have isolated two kinds of impairments which appear in moving video, when the embedding strength is increased, but not in still frames:

1. *Temporal flicker*: Embedding uncorrelated watermarks in successive video frames (SS strategy) usually results in annoying twinkle or flicker artifacts similar to the existing ones in video compression,
2. *Stationary pattern*: Embedding the same watermark pattern in all the video frames (SS-1 strategy) is visually disturbing since it gives the feeling that the scene has been filmed with a dirty camera when it pans across the movie set.

With the proposed motion compensated embedding strategy, different watermarks are still embedded in successive video frames. However, these differences are coherent with the camera motion and the user is no longer annoyed by flickering. In fact, the user has the feeling that the noise was already present in the filmed movie set and find it more *natural*. On the other hand, the proposed embedding strategy introduces a new kind of artifacts. All the embedded watermarks $\mathbf{W}_t^{(t)}$ originate from the same watermark pattern $\mathbf{W}$. Nevertheless, they have been obtained using different warping parameters, and in particular different zoom factor z. As a result, the embedded watermarks have not the same frequency content: if the camera zooms in, the watermark slides towards low frequencies and thus becomes more visible. This issue will be addressed in future work.

6. CONCLUSION AND PERPECTIVES

Watermarking video is still regarded today most of the time as watermarking a sequence of still images. This has resulted in weak algorithms in terms of security. In particular, intra-video collusion attacks can be designed to trap non secure watermarking schemes. In this paper, an attack exploiting frame registration has been presented to demonstrate how algorithms based on either SS or SS-1 strategy can be defeated. The SS-Reg strategy has then been introduced to watermark digital video. The main point is to compensate camera motion before inserting the watermark so that a given 3D point of the movie set carries the same watermark sample all along a video scene. A practical implementation of this embedding strategy using video mosaicing has been proposed. This approach has been proven to resist temporal frame averaging after registration. Furthermore, it has also been noted that such a motion coherent watermark is likely to improve the invisibility of the watermark.

This mosaicing-based approach is computationally intensive, which may prevent real-time processing. As a result, future work will explore whether this camera motion compensated watermark can be introduced using an alternative approach. From a more general point of view, this embedding strategy ensures temporal coherency: each point is tracked in time so that it carries the same watermark sample. On the other hand, recent attacks[28, 29] have exploited weaknesses due to spatially non-coherent watermark i.e. similar parts in an image do not carry similar watermarks. It has been shown that such watermarks can be removed by replacing each part of the protected signal with a combination of similar parts from the same signal. Thus, future work will also explore how to obtain spatially coherent watermarks, so that the combination of both approaches gives a really secure watermark.

ACKNOWLEDGMENTS

The authors want to thank Henri Nicolas[13, 30] from IRISA, Rennes, France for providing useful video processing support (video objects segmentation and warping parameters of the video *stefan*) and enlightening discussions on video mosaicing.

REFERENCES

1. I. Cox, M. Miller, and J. Bloom, *Digital Watermarking*, Morgan Kaufmann Publishers, 2001.
2. G. Doërr and J.-L. Dugelay, "A guide tour of video watermarking," *Signal Processing: Image Communication* **18**, pp. 263–282, April 2003.
3. S. Voloshynovskiy, A. Herrigel, N. Baumgärtner, and T. Pun, "A stochastic approach to content adaptive digital image watermarking," in *Proceedings of the Third International Workshop on Information Hiding*, *Lecture Notes on Computer Science* **1768**, pp. 211–236, September 1999.
4. E. Lin and E. Delp, "Temporal synchronization in video watermarking - further studies," in *Security and Watermarking of Multimedia Contents V*, *Proceedings of SPIE* **5020**, pp. 493–504, January 2003.
5. F. Hartung and B. Girod, "Watermarking of uncompressed and compressed video," *Signal Processing* **66**, pp. 283–301, May 1998.
6. T. Kalker, G. Depovere, J. Haitsma, and M. Maes, "A video watermarking system for broadcast monitoring," in *Security and Watermarking of Multimedia Contents*, *Proceedings of SPIE* **3657**, pp. 103–112, January 1999.
7. T. Kalker, "Considerations on watermarking security," in *Proceedings of the IEEE Workshop on Multimedia Signal Processing*, pp. 201–206, October 2001.
8. M. Holliman, N. Memon, and M. Yeung, "On the need for image dependent keys for watermarking," in *Proceedings of the IEEE Symposium on Content Security and Data Hiding in Digital Media*, May 1999.
9. 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*, *Proceedings of SPIE* **3971**, pp. 358–370, January 2000.
10. G. Doërr and J.-L. Dugelay, "New intra-video collusion attack using mosaicing," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, **II**, pp. 505–508, July 2003.
11. R. Koenen, "MPEG-4 overview," in *JTC1/SC29/WG11 N4668*, ISO/IEC, March 2002.
12. A. Smolic, M. Lorei, and T. Sikora, "Adaptive kalman-filtering for prediction and global motion parameter tracking of segments in video," in *Proceedings of the Picture Coding Symposium*, March 1996.
13. H. Nicolas and C. Labit, "Motion and illumination variation estimation using a hierarchy of models: Application to image sequence coding," *Journal of Visual Communication and Image Representation* **6**, pp. 303–316, December 1995.
14. R. Szeliski and H.-Y. Shum, "Creating full view panoramic image mosaics and environment maps," in *Proceedings of the International Conference on Computer Graphics and Interactive Techniques*, pp. 251–258, April 1997.
15. Z. Sun and M. Tekalp, "Trifocal motion modeling for object-based video compression and manipulation," *IEEE Journal on Circuits and Systems for Video Technology* **8**, pp. 667–685, May 1998.

16. M. Holliman, W. Macy, and M. Yeung, "Robust frame-dependent video watermarking," in *Security and Watermarking of Multimedia Contents II*, *Proceedings of SPIE* **3971**, pp. 186–197, January 2000.
17. J. Fridrich and M. Goljan, "Robust hash functions for digital watermarking," in *Proceedings of the International Conference on Information Technology: Coding and Computing*, pp. 178–183, March 2000.
18. 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*, *Proceedings of SPIE* **4675**, pp. 548–554, January 2002.
19. K. Su, D. Kundur, and D. Hatzinakos, "A novel approach to collusion resistant video watermarking," in *Security and Watermarking of Multimedia Contents IV*, *Proceedings of SPIE* **4675**, pp. 491–502, January 2002.
20. M. Irani, P. Anandan, J. Bergen, R. Kumar, and S. Hsu, "Mosaic representations of video sequences and their applications," *Signal Processing: Image Communication* **8**, pp. 327–351, May 1996.
21. P. Bas and B. Macq, "A new video-object watermarking scheme robust to object manipulation," in *Proceedings of the IEEE International Conference on Image Processing*, **II**, pp. 526–529, October 2001.
22. F. Hartung, P. Eisert, and B. Girod, "Digital watermarking of MPEG-4 facial animation parameters," *Computers & Graphics* **22**, pp. 425–435, July 1998.
23. E. Garcia and J.-L. Dugelay, "Texture-based watermarking of 3D video objects," *IEEE Transactions on Circuits and Systems for Video Technology* **13**, pp. 853–866, August 2003.
24. I. Cox and M. Miller, "Preprocessing media to facilitate later insertion of a watermark," in *Proceedings of the International Conference on Digital Signal Processing*, **1**, pp. 67–70, July 2002.
25. Visual Quality Expert Group (VQEG), "http://www.vqeg.org."
26. W. Macy and M. Holliman, "Quality evaluation of watermarked video," in *Security and Watermarking of Multimedia Contents II*, *Proceedings of SPIE* **3971**, pp. 486–500, January 2000.
27. S. Winkler, E. Gelasca, and T. Ebrahimi, "Towards perceptual metrics for video watermark evaluation," in *Applications of Digital Image Processing*, *Proceedings of SPIE* **5203**, pp. 371–378, August 2003.
28. C. Rey, G. Doërr, J.-L. Dugelay, and G. Csurka, "Toward generic image dewatermarking?," in *Proceedings of the IEEE International Conference on Image Processing*, **III**, pp. 633–636, September 2002.
29. D. Kirovski and F. Petitcolas, "Blind pattern matching attack on watermarking systems," *IEEE Transactions on Signal Processing* **51**, pp. 1045–1053, April 2003.
30. H. Nicolas, "New methods for dynamic mosaicing," *IEEE Transactions on Image Processing* **10**, pp. 1239–1251, August 2001.