Countermeasures for Collusion Attacks Exploiting Host Signal Redundancy

Gwenaël Doërr and Jean-Luc Dugelay

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

Abstract. Multimedia digital data is highly redundant: successive video frames are very similar in a movie clip, most songs contain some repetitive patterns, etc. This property can consequently be exploited to successively replace each part of the signal with a similar one taken from another location in the same signal or with a combination of similar parts. Such an approach is all the more pertinent when video content is considered since such signals exhibit both temporal and spatial self-similarities. To counter such attacking strategies, it is necessary to ensure that embedded watermarks are coherent with the redundancy of the host content. To this end, both motion-compensated watermarking and self-similarities inheritance will be surveyed.

1 Introduction

Digital watermarking was initially introduced in the early 90's as a complementary protection technology [1] since encryption alone is not enough. Indeed, sooner or later, encrypted multimedia content is decrypted to be eventually presented to human beings. At this very moment, multimedia content is left unprotected and can be perfectly duplicated, manipulated and redistributed at a large scale. Thus, a second line of defense has to be added to address this issue. This is the main purpose of digital watermarking which basically consists in hiding some information into digital content in an imperceptible manner. Up to now, research has mainly investigated how to improve the trade-off between three conflicting parameters: imperceptibility, robustness and capacity. Perceptual models have been exploited to make watermarks less perceptible, benchmarks have been released to evaluate robustness, channel models have been studied to obtain a theoretical bound for the embedding capacity.

A lot of attention has focused on security applications such as Intellectual Property (IP) protection and Digital Rights Managements (DRM) systems. Digital watermarking was even thought of as a possible solution to combat illegal copying which was a forthcoming issue in the mid-90's. However the few attempts to launch watermarking-based copy-control mechanisms [2, 3] have resulted in partial failures, which have significantly lowered the initial enthusiasm related to this technology. These setbacks were in part due to the claim that embedded watermarks would survive in a highly hostile environment even if very few works addressed this issue. Indeed, if the survival of the watermark against common signal processing primitives - filtering, lossy compression, global desynchronization - has been carefully surveyed, almost no work has considered that an attacker may try to learn some knowledge about the watermarking system to defeat it. Nevertheless, in applications such as copy control or fingerprinting, digital watermarking is usually seen as a disturbing technology. Therefore, it is likely to be submitted to strong hostile attacks when it is released to the public.

Security evaluation is now a growing concern and collusion attacks have often been mentioned as a possible mean to do it [4, 5]. Collusion consists in collecting several watermarked documents and combining them to obtain unwatermarked content. Such attacks are all the more relevant in video since each individual frame can be regarded as a single watermarked document. In Section 2, a specific kind of collusion attack is reviewed. When similar contents carry uncorrelated watermarks, colluders can average them so that watermark samples sum to zero. In this perspective, an attacker can exploit both the temporal and spatial redundancy of the video signal to design efficient attacks. Next, signal coherent watermarking is introduced in Section 3 to circumvent the previously exhibited threats. The goal is basically to make the embedded watermark have the same redundancy as the host signal. To this end, motion-compensated watermarking and self-similarities inheritance will be studied. Finally, conclusions are drawn in Section 4 and tracks for future work are given.

2 Combine Similar Contents Carrying Uncorrelated Watermarks

Previous works have stressed the fact that using a redundant watermarking structure is likely to induce some information leakages [6–8]. Considering multiple watermarked contents, a hostile attacker is able to gain some knowledge about the embedded watermark signal and exploit it to confuse the detector. Nevertheless, completely independent watermarks are not the solution either. If an attacker can collect similar contents carrying uncorrelated watermarks, averaging them will usually sum the watermark samples to zero. Since video material is highly redundant, such a strategy can lead to powerful attacks. In Subsection 2.1, the correlation between successive frames is exploited to estimate the background in each frame using the neighbor ones. Furthermore, spatial self-similarities will also be considered in Subsection 2.2 to elaborate efficient Block Replacement Attacks (BRA).

2.1 Temporal Frame Averaging after Registration

One of the pioneering algorithm for video watermarking basically considers video content as a mono-dimensional signal and simply adds a pseudo-random sequence as a watermark [9]. From a frame-by-frame point of view, such a strategy can be

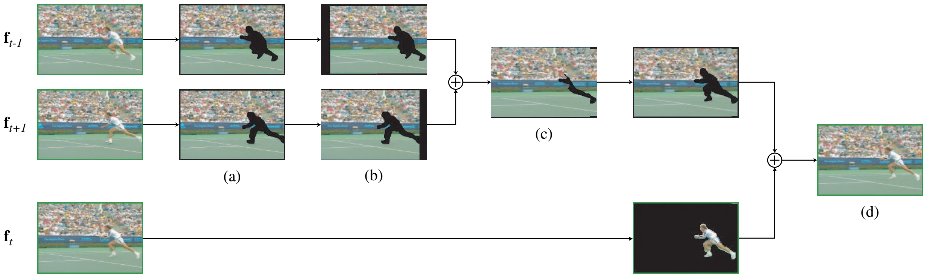

Fig. 1. Temporal Frame Averaging after Registration (TFAR): Once the video objects have been removed (a), neighbor frames are registered (b) and combined to estimate the background of the current frame (c). Next, the missing video objects are inserted back (d).

seen as always embedding a different watermark[1]. The drawback of this approach is that temporal frame averaging usually succeeds in confusing the watermark detector [4]. In static scenes, video frames are highly similar and can be averaged without introducing strong visible artifacts. On the other hand, since successive watermarks are uncorrelated, temporal averaging significantly decreases the power of the embedded watermark $\mathbf{w}_t$ in the frame $\mathbf{f}_t$. Nevertheless, in practice, video material usually contains dynamic components such as fast moving objects and/or camera motion. Therefore, this simple attack needs to be improved to ensure that the quality of the video is not destroyed.

Each video frame is a projection of a single 3D movie set and different video frames from a shot can be seen as different 2D projections of the same scene. As a result, even if some dynamic components are present, successive frames are still highly correlated. However, they need to be aligned to enable efficient averaging [11, 12]. The goal is to register the video frames, so that all the projections of a given 3D point overlap, to enable large temporal averaging without introducing much visual distortion. In other words, Temporal Frame Averaging after Registration (TFAR) aims at estimating a given video frame $\mathbf{f}_t$ from its neighboring ones $\mathbf{f}_{t+\delta}$ thanks to frame registration as depicted in Figure 1. Moving objects are difficult to predict from one frame to the other. This is the reason why segmentation is used to separate two alternative Video Object Planes (VOP) [13]: the background $\mathbf{b}_t$ on one side and the moving objects $\mathbf{o}_t$ on the other side. Video objects are then ignored for the rest of the attack, which simply comes down then to estimate the current background $\mathbf{b}_t$ from the neighbor ones.

[1] Frame-by-frame watermarking is a commonly used strategy in video [10]. The following notation will be used in the remainder of this article: $\check{\mathbf{f}}_t = \mathbf{f}_t + \alpha \mathbf{w}_t$, where $\mathbf{f}_t$ is the original video frame at instant t, $\check{\mathbf{f}}_t$ its watermarked version, α the embedding strength and $\mathbf{w}_t$ the embedded watermark which is normally distributed with zero mean and unit variance.

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 [14] as follows:

$$\begin{cases} x_{t'} = t_x + z(x_t - x_o) - z\theta(y_t - y_o) \\ y_{t'} = t_y + z(y_t - y_o) + z\theta(x_t - x_o) \end{cases} \tag{1}$$

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 simple model may be inaccurate when the camera displacement or the scene structure is very complicated. In this case, more complex motion representations can be introduced [14–16].

The registered backgrounds $\mathbf{b}_{t+\delta}^{(t)}$, obtained from the video frames in the considered temporal window, are averaged to obtain an estimation $\tilde{\mathbf{b}}_t$ of the background in the current frame. The moving objects $\mathbf{o}_t$ are then inserted back to obtain the attacked video frame $\tilde{\mathbf{f}}_t$. It should be noted that this attack does not affect the moving 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, since the background is usually the main part in a video shot, the attack remains pertinent. From a coding perspective, TFAR can be seen as encoding the background with an advanced forward-backward predictive coder e.g. B-frames in MPEG. Alternatively, it can also be considered as temporal averaging along the motion axis. Whatever, since most watermarking algorithms do not consider the evolution of the structure of the scene during embedding, this attack has been shown to confuse several watermark detectors [12]. The only exception is when the same watermark pattern $\mathbf{w}$ is embedded in all the video frames in a static scene. In this case, TFAR has no impact. Skeptical people might argue that such attacks are too computationally intensive 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 and such video coding algorithms will have a similar impact on embedded watermarks [17].

2.2 Block Replacement Attack

If similarities can be easily exhibited in successive video frames as noticed in the previous subsection, less obvious ones are also present at a lower resolution level: the block level. Such self-similarities have already been exploited to obtain efficient image compression tools [18]. The signal to be processed is first partitioned into a set of blocks $\mathbf{b}_T$ of size S_T. Those blocks can overlap or not. The asset of using overlapping blocks is that it prevents strong blocking artifacts on the border of the blocks by averaging the overlapping areas. The Block Replacement Attack (BRA) processes then each one of these blocks sequentially. For each block, a search window is defined. It can be chosen in the vicinity of the block $\mathbf{b}_T$ or randomly to prevent system designers to systematically invert the attack. This search window is partitioned to obtain a codebook $\mathcal{Q}$ of blocks $\mathbf{b}_{Q_i}$ of size S_Q. Once again, these blocks can overlap or not. Next a candidate block for replacement $\mathbf{b}_R$ is computed using the blocks present in the codebook. Of course, the larger the codebook $\mathcal{Q}$ is, the more choices there are to compute a replacement block which is *similar* enough to the input block $\mathbf{b}_T$ so that it can be substituted without introducing strong visual artifacts. On the other hand, the larger the codebook $\mathcal{Q}$ is, the higher the computational complexity is and a trade-off has to be found. The Mean Square Error (MSE) can be used to evaluate how similar are two blocks. The lower the MSE is, the more similar are the two blocks. Thus, the original block $\mathbf{b}_T$ is substituted by the replacement block $\mathbf{b}_R$ associated with the lowest MSE.

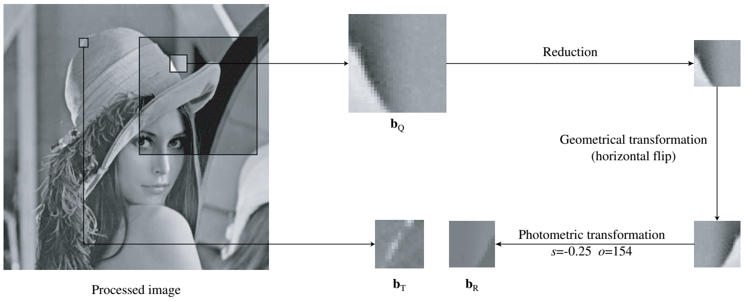

Fig. 2. Block Replacement Attack (BRA) implementation using a fractal coding strategy: each block is replaced by the one in the search window which is the most similar modulo a geometrical and photometric transformation.

There are many ways of computing the replacement block $\mathbf{b}_R$. One of the first proposed implementation was based on fractal coding [19] and is illustrated in Figure 2. The codebook is first artificially enlarged by also considering geometrically transformed versions of the blocks within the search window. For complexity reasons, a small number of transformations are considered e.g. downsampling by a factor 2 and 8 isometries (identity, 4 flips, 3 rotations). Next, the candidate replacement blocks are computed with a simple affine photometric compensation. In other words, each block $\mathbf{b}_{Q_i}$ of the codebook is transformed in $s\mathbf{b}_{Q_i} + o\mathbf{1}$, where $\mathbf{1}$ is a block containing only ones, so that the MSE with the target block $\mathbf{b}_T$ is minimized. This is a simple least squares problem and the scale s and offset o can be determined as follows:

$$s = \frac{(\mathbf{b}_T - \mathrm{m}_T\mathbf{1}) \cdot (\mathbf{b}_{Q_i} - \mathrm{m}_{Q_i}\mathbf{1})}{|\mathbf{b}_{Q_i} - \mathrm{m}_{Q_i}\mathbf{1}|^2} \tag{2}$$

$$o = \mathrm{m}_T - s.\mathrm{m}_{Q_i} \tag{3}$$

where m_T (resp. m_{Q_i}) is the mean value of block $\mathbf{b}_T$ (resp. $\mathbf{b}_{Q_i}$), $\cdot$ the linear correlation and $|\mathbf{b}|$ the norm defined as $\sqrt{\mathbf{b} \cdot \mathbf{b}}$. At this point, the transformed blocks $s\mathbf{b}_{Q_i} + o\mathbf{1}$ are sorted in ascending order according to their similarity with the target block $\mathbf{b}_T$ and the most similar one is retained for replacement. In the same fashion, an alternative approach consists in building iteratively sets of similar blocks and randomly shuffling their positions [20, 21] until all the blocks have been replaced.

The main drawback of this implementation is that it is not possible to modify the strength of the attack. Furthermore, the computation of the replacement block is not properly managed: either it is too close from the target block $\mathbf{b}_T$ and the watermark is reintroduced, or it is too distant and strong visual artifacts appear. Optimally, one would like to ensure that the distortion $\Delta = \mathrm{MSE}(\mathbf{b}_R, \mathbf{b}_T)$ remains within two bounds τ_{low} and τ_{high}. To this end, several blocks $\mathbf{b}_{Q_i}$ can be combined to compute the replacement block instead of a single one i.e. $\mathbf{b}_R = \sum_{i=1}^{N} \lambda_i \mathbf{b}_{Q_i}$ where the λ_i's are mixing parameter chosen in such a way that Δ lies within the specified interval. This combination can take into account a fixed number of blocks [22] or also adapt the number of considered blocks for combination according to the nature of the block to be reconstructed [23]. Intuitively, approximating flat blocks requires to combine fewer blocks than for highly textured ones.

However, the computational load induced by computing optimal mixing parameters for each candidate replacement block has motivated the design of an alternative implementation which is described in Table 1. First, for each block $\mathbf{b}_T$, the codebook $\mathcal{Q}$ is built and photometric compensation is performed. Next, Principal Component Analysis (PCA) is performed considering the different blocks $\mathbf{b}_{Q_i}$ in the codebook. This gives a centroid $\mathbf{c}$ defined as follows:

$$\mathbf{c} = \frac{1}{|\mathcal{Q}|} \sum_{\mathbf{b}_{Q_i} \in \mathcal{Q}} \mathbf{b}_{Q_i} \tag{4}$$

and a set of eigenblocks $\mathbf{e}_i$ associated with their eigenvalues ϵ_i. These eigenblocks are then sorted by descending eigenvalues i.e. there are more variations in direction e_1 than in any other one. Then, a candidate block for replacement $\mathbf{b}_R$ is computed using the N first eigenblocks so that the distortion Δ is minimized. In other words, the block $\mathbf{b}_T - \mathbf{c}$ is projected onto the subspace spanned by the N first eigenblocks and $\mathbf{b}_R$ can be written:

$$\mathbf{b}_R = \mathbf{c} + \sum_{i=1}^{N} \frac{(\mathbf{b}_T - \mathbf{c}) \cdot \mathbf{e}_i}{|\mathbf{e}_i|^2} \mathbf{e}_i \tag{5}$$

Table 1. BRA procedure using block projection on a PCA-defined subspace.

```
For each block $\mathbf{b}_T$ of the signal
$\boxed{1}$ Build the block codebook $\mathcal{Q}$
$\boxed{2}$ Perform photometric compensation
$\boxed{3}$ Performs the PCA of the blocks in $\mathcal{Q}$ to obtain a set of orthogonal eigenblocks $\mathbf{e}_i$
    associated with their eigenvalues $\epsilon_i$
    Set $N = 1$ and flag $= 0$
$\boxed{4}$ While (flag $= 0$) AND ($N \leq S_T$)
    (a) Build the optimal replacement block $\mathbf{b}_R$ using the eigenblocks $\mathbf{r}_i$ associated
        with the $N$ first eigenvalues
    (b) Compute $\Delta = \mathrm{MSE}(\mathbf{b}_R, \mathbf{b}_T)$
    (c) If $\tau_{\mathrm{low}} \leq \Delta \leq \tau_{\mathrm{high}}$, set flag $= 1$
    (d) Else increment $N$
$\boxed{5}$ Replace $\mathbf{b}_T$ by $\mathbf{b}_R$
```

Of course, the distortion Δ gracefully decreases as the number N of combined eigenblocks increases. Thus, an adaptive framework is introduced to identify which value N should have so that the distortion Δ falls within the range $[\tau_{\mathrm{low}}, \tau_{\mathrm{high}}]$. It should be noted that the underlying assumption is that most of the watermark energy will be concentrated in the last eigenblocks since the watermark can be seen as details. As a result, if a valid candidate block can be built without using the last eigenblocks, the watermark signal will not be reintroduced. In fact, BRA has been shown to defeat both Spread Spectrum (SS) and Quantization Index Modulation (QIM) watermarks [21, 23].

3 Signal Coherent Watermarking

On one hand, using a redundant watermarking structure is not secure since it can be estimated when several watermarked uncorrelated documents are colluded. On the other hand, uncorrelated watermarks can be removed by averaging similar watermarked documents. These observations intuitively lead to the intuitive embedding principle: *watermarks embedded in distinct contents should be as correlated as the host contents themselves.* Alternative approaches have been proposed to meet this specification e.g. the embedded watermark can be made frame-dependent [24], a frame-dependent binary string can be exploited to generate a watermark pattern which degrades gracefully with an increased number of bit errors [25, 26], the watermark can be embedded in some frame-dependent positions [4]. However, those methods are likely to be still defeated by the attacks presented in Section 2. Indeed, the watermark needs to be coherent with the redundancy of the host signal. First, camera motion should be carefully considered to resist to TFAR. Optimally, the embedding process should ensure that *the watermark moves with the camera.* Second, the embedded watermark should exhibit the same spatial self-similarities as the host video frames to make sure it is immune to BRA. If a pattern is repeated in a frame, it should always carry the same watermark.

3.1 Motion Compensated Watermarking

For a given scene, backgrounds of video frames can be considered as several 2D projections of the same 3D movie set. The weakness of common embedding strategies against TFAR is due to the fact that camera motion is not considered at all. These watermarking systems are completely *blind* with respect to camera motion. 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. A remedy would be 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 words, the basic idea is to simulate a utopian world where the movie set would already be watermarked. In this perspective, video mosaicing can be considered to design such a motion compensated watermarking scheme.

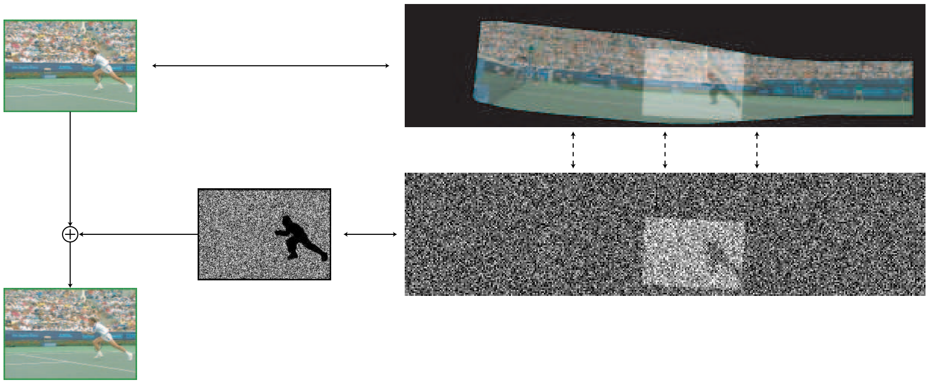

Fig. 3. Embedding procedure for camera motion coherent watermarking: The part of the watermark pattern which is associated with the current video frame is retrieved and registered back. Next, it is embedded in the background portion of the video frame.

Video mosaicing consists in aligning all the frames of a video sequence to a fixed coordinate system [27]. The resulting mosaic image provides a snapshot view of the video 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 embedding a digital watermark in the mosaic representation of the considered video scene. Next, the resulting watermarked mosaic would be used as the background of the video frames. However, such a process requires double interpolation for the background (frame → mosaic → frame) which is likely to alter the quality of the video. Therefore, an alternative but somewhat equivalent approach is depicted in Figure 3. First of all, warping parameters are computed for each video frame with respect to the considered motion model. For instance, if the motion model defined in Equation (1) is exploited, the warping parameters θ, z, (x_o, y_o) and (t_x, t_y) are computed for each video frame. Hence, each frame $\mathbf{f}_t$ is associated with a set of warping parameters i.e. the frame background $\mathbf{b}_t$ is associated with a portion $\mathbf{b}_{\mathrm{m}}^{(t)}$ of the video mosaic. Next, a key-dependent watermark $\mathbf{w}_{\mathrm{m}}$ is generated which has the same dimensions as the mosaic representation of the video shot. Now, using the same warping parameters as the ones used for building the mosaic, a portion $\mathbf{w}_{\mathrm{m}}^{(t)}$ of this watermark can be associated to each video frame $\mathbf{f}_t$. Finally, the resulting watermark portion only has to be registered back to obtain the watermark signal $\mathbf{w}_t$ to be embedded in the video frame. Similarly to TFAR, object segmentation can be performed to separate moving objects from the background. Next, the embedder only watermarks the background to follow the embedding philosophy: *a 3D point carries the same watermark sample all along the video scene*. In this case, alternative mechanisms have to be deployed to protect moving objects. Previous works have watermarked MPEG-4 video objects according to their main directions [28], their animation parameters [29] or their texture [30]. On the detector side, the procedure is very similar. In a first step, warping parameters are computed for each frames of the video scene to be verified and the watermark $\mathbf{w}_{\mathrm{m}}$ is generated using the shared secret key. Next, the detector only checks whether the portion $\mathbf{w}_t$ associated with each incoming frame $\tilde{\mathbf{f}}_t$ has been effectively embedded in the background or not using for instance a correlation score.

As expected, this novel embedding strategy has exhibited very good performances against TFAR [12]. Furthermore, this method also produces interesting results in terms of watermark imperceptibility. 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) [31] was formed in 1997 to devise objective methods for predicting video image quality. In 1999, they stated first, that no objective measurement system at test was able to replace subjective testing and second, that no objective model outperforms the others in all cases. This explains while the Peak Signal to Noise Ratio (PSNR) 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 [32, 33] have isolated two kinds of impairments which appear in video, when the embedding strength is increased, but not in still frames:

1. *Temporal flicker*: Embedding uncorrelated watermarks in successive video frames 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 is visually disturbing since it gives the feeling that the scene has been filmed with a camera having a dirty lens 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*.

3.2 Host Self-similarities Inheritance

For each signal block, BRA look for a linear combination of neighboring blocks resulting in a block which is similar enough to the current block so that a substitution does not introduce strong visual artifacts. Since watermarking systems do not perform today anything specific to ensure that the embedded watermark is coherent with the self-similarities of the host signal, most of them are defeated by such attacks. Intuitively, to ensure that a watermark will survive to BRA, the embedding process should guarantee that *similar signal blocks carry similar watermarks* or alternatively that *pixels with similar neighborhood carry watermark samples with close values*.

Let us assume for the moment that it is possible to associate to each pixel position $\mathbf{p} = (x, y)$ with $1 \leq x \leq X$ and $1 \leq y \leq Y$ in the image $\mathbf{i}$ a feature vector $\mathbf{f}(\mathbf{i}, \mathbf{p})$ which characterizes *in some sense* the neighborhood of the image around this specific position. Thus, this function can be defined as follows:

$$\begin{aligned} \mathbf{f} : \mathcal{I} \times \mathcal{P} &\rightarrow \mathcal{F} \\ (\mathbf{i}, \mathbf{p}) &\mapsto \mathbf{f}(\mathbf{i}, \mathbf{p}) \end{aligned} \tag{6}$$

where $\mathcal{I}$ is the image space, $\mathcal{P} = [1 \dots X] \times [1 \dots Y]$ the position space and $\mathcal{F}$ the feature space. From a very low-level perspective, generating a digital watermark can be regarded as associating a watermark value $\mathrm{w}(\mathbf{i}, \mathbf{p})$ to each pixel position in the image. However, if the embedded watermark is required to be immune against BRA, the following property should also be verified:

$$\mathbf{f}(\mathbf{i}, \mathbf{p}_0) \approx \sum_k \lambda_k \mathbf{f}(\mathbf{i}, \mathbf{p}_k) \Rightarrow \mathrm{w}(\mathbf{i}, \mathbf{p}_0) \approx \sum_k \lambda_k \mathrm{w}(\mathbf{i}, \mathbf{p}_k) \tag{7}$$

In other words, if at a given position $\mathbf{p}_0$, the local neighborhood is similar to a linear combination of neighborhoods at other locations $\mathbf{p}_k$, then the watermark sample $\mathrm{w}(\mathbf{p}_0)$ embedded at position $\mathbf{p}_0$ should be close to the linear combination (with the same mixing coefficients λ_k) of the watermark samples $\mathrm{w}(\mathbf{p}_k)$ at these locations. A simple way to obtain this property is to make the watermarking process be the composition of a feature extraction operation and a linear form φ.

Hence, one can write $\mathrm{w} = \varphi \circ \mathbf{f}$ where $\varphi : \mathcal{F} \rightarrow \mathbb{R}$ is a linear form which takes F-dimensional feature vectors in input. Next, to completely define this linear form, it is sufficient to set the values $\xi_f = \varphi(\mathbf{b}_f)$ for a given orthonormalized basis $\mathcal{B} = \{\mathbf{b}_f\}$ of the feature space $\mathcal{F}$. Without loss of generality, one can consider the canonical basis $\mathcal{O} = \{\mathbf{o}_f\}$ where $\mathbf{o}_f$ is a F-dimensional vector filled with 0's except the fth coordinate which is equal to 1. The whole secret of the algorithm is contained in the values ξ_f and they can consequently be pseudo-randomly generated using a secret key K. Now, assuming that feature vectors have an isotropic distribution, the probability density function of the linear form over the unit sphere $\mathcal{U}$ is given by [34]:

$$\mathrm{f}_{\varphi|\mathcal{U}}(w) = \frac{1}{\varXi\sqrt{\pi}} \frac{\Gamma\left(\frac{F}{2}\right)}{\Gamma\left(\frac{F-1}{2}\right)} \left[1 - \left(\frac{w}{\varXi}\right)^2\right]^{\frac{F-3}{2}} \tag{8}$$

where $\varXi^2 = \sum_{f=1}^{F} \xi_f^2$ and $\Gamma(.)$ is the Gamma function. When the dimension F of the feature space $\mathcal{F}$ grows large, this probability density function tends towards a Gaussian distribution with zero mean and standard deviation $\varXi/\sqrt{F}$. Thus if the ξ_f's are chosen to have zero mean and unit variance, this ensures that the values of the linear form restricted to the unit sphere $\mathcal{U}$ are normally distributed with also zero mean and unit variance. Then, keeping in mind that φ is linear and that the following equation is valid,

$$\mathrm{w}(\mathbf{i}, \mathbf{p}) = \varphi\Bigg(\|\mathbf{f}(\mathbf{i}, \mathbf{p})\| \frac{\mathbf{f}(\mathbf{i}, \mathbf{p})}{\|\mathbf{f}(\mathbf{i}, \mathbf{p})\|}\Bigg) = \|\mathbf{f}(\mathbf{i}, \mathbf{p})\| \varphi\big(\mathbf{u}(\mathbf{i}, \mathbf{p})\big) \quad \text{with}\, \mathbf{u}(\mathbf{i}, \mathbf{p}) \in \mathcal{U} \tag{9}$$

it is straightforward to realize that the obtained watermark is equivalent to a Gaussian watermark with zero mean and unit variance multiplied by some local scaling factors. The more textured is the considered neighborhood, the more complicated it is to characterize it and the greater the norm $\|\mathbf{f}(\mathbf{i}, \mathbf{p})\|$ is likely to be. Looking back at Equation 9, it results that the watermark is amplified in textured area whereas it is attenuated in smooth ones. This can be regarded as some kind of perceptual shaping [35].

A practical implementation of this strategy using Gabor features has clearly demonstrated its superiority with respect to BRA in comparison to common SS watermarks [36]. Furthermore, this implementation exhibited an unexpected relationship with earlier multiplicative watermarking schemes in the frequency domain. The watermark sample obtained at position $\mathbf{p}$ is simply given by:

$$\mathrm{w}(\mathbf{i}, \mathbf{p}) = \sum_{f=1}^{F} \xi_f \mathbf{g}_f(\mathbf{i}, \mathbf{p}) \tag{10}$$

where $\mathbf{g}_f(\mathbf{i}, \mathbf{p})$ is the f-th coordinate of the F-dimensional Gabor feature vector $\mathbf{g}(\mathbf{i}, \mathbf{p})$. In other words, the watermark is a linear combination of different Gabor responses $\mathbf{g}_f$. However, when the number of filters in the Gabor filterbank grows, more and more Gabor responses need to be computed which can be quickly computationally prohibitive. Hopefully, when the Fourier domain is considered, the watermark can be computed as follows:

$$\begin{aligned}\mathbf{W}(\mathbf{i},\mathbf{q}) &= \sum_{\mathbf{p}\in\mathcal{P}} \left(\sum_{f=1}^{F} \xi_f \, \mathbf{g}_f(\mathbf{i},\mathbf{p}) \right) \omega_{\mathbf{p},\mathbf{q}} \\ &= \sum_{f=1}^{F} \xi_f \left(\sum_{\mathbf{p}\in\mathcal{P}} \mathbf{g}_f(\mathbf{i},\mathbf{p}) \, \omega_{\mathbf{p},\mathbf{q}} \right) = \sum_{f=1}^{F} \xi_f \, \mathbf{G}_f(\mathbf{i},\mathbf{q}) \\ &= \sum_{f=1}^{F} \xi_f \, \mathbf{H}_f(\mathbf{q}) \, \mathbf{I}(\mathbf{q}) = \mathbf{H}(K,\mathbf{q}) \, \mathbf{I}(\mathbf{q}) \qquad (11) \\ &\text{with} \quad \mathbf{H}(K,\mathbf{q}) = \sum_{f=1}^{F} \xi_f \, \mathbf{H}_f(\mathbf{q})\end{aligned}$$

where $\omega_{\mathbf{p},\mathbf{q}} = \exp\left[-j2\pi\left((x-1)(u-1)/X + (y-1)(v-1)/Y\right)\right]$, capital letters indicate FFT-transformed variables and $\mathbf{q} = (u, v)$ denotes a frequency position with $1 \leq u \leq U$ and $1 \leq v \leq V$. The Gabor response $\mathbf{G}_f$ is given in the frequency domain by the multiplication of the image spectrum $\mathbf{I}$ with some filter $\mathbf{H}_f$. In summary, Equation 11 means that the watermark can be generated in one row in the Fourier domain by computing $\mathbf{H}$. It is now straightforward to realize that the watermark generation process comes down to a simple multiplication between the image spectrum $\mathbf{I}$ and some pseudo-random signal $\mathbf{H}(K)$. Following this track, multiplicative watermarks in the FFT [37] and the DCT [38] domains have been shown to be also resilient against BRA. At this point, it is interesting to note that multiplicative watermarking in the frequency domain was initially motivated by contrast masking properties: larger coefficients can convey a larger watermark value without compromising invisibility [39]. This can be related with the natural perceptual shaping of signal coherent watermarks exhibited in Equation (9).

4 Conclusion

The partial failure of initiatives to launch copy control mechanisms using digital watermarking has recently triggered an effort in the watermarking community to evaluate security. Security is basically related with the fact that, in many applications, consumers do not benefit from the introduction of digital watermarks: they can be used to identify customers, to prevent playback of illegal content, etc. As a result, customers are likely to attack the protection system. In this perspective, researchers try to anticipate their hostile behaviors to propose efficient countermeasures. In this paper, two collusion attacks have been introduced which exploit the redundancy of the host signal to remove the embedded watermark. In order to circumvent those threats, two remedies have been proposed to make the embedded watermark coherent with the spatio-temporal redundancy of the host video signal. The first one considers camera motion during embedding to ensure immunity against TFAR. The second one takes the self-similarities of the host signal into account to cope with BRA at the block level. However, at this stage it is not possible to assert how secure the obtained schemes are. One can only claim that they resist BRA but nothing ensures that another attack will not defeat them. Recent studies have defined some kind of *security metric* to determine how much information leaks when a redundant watermarking structure is used [6]. It could be interesting to investigate in the near future whether this approach can be extended to also consider the case when non redundant watermarks are used. The resulting metric would then be useful to quantify the security level of signal coherent watermarking.

5 Acknowledgment

This work has been supported in part by the European Commission through the IST Program under Contract IST-2002-507932 ECRYPT.

References

1. Cox, I., Miller, M., Bloom, J.: Digital Watermarking. Morgan Kaufmann Publishers (2001)
2. DVD Copy Control Association: (http://www.dvdcca.org)
3. Secure Digital Music Initiative: (http://www.sdmi.org)
4. Su, K., Kundur, D., Hatzinakos, D.: A novel approach to collusion resistant video watermarking. In: Security and Watermarking of Multimedia Contents IV. Volume 4675 of Proceedings of SPIE. (2002) 491–502
5. Doërr, G., Dugelay, J.L.: Collusion issue in video watermarking. In: Security, Steganography and Watermarking of Multimedia Contents VII. Volume 5681 of Proceedings of SPIE. (2005) 685–696
6. Cayre, F., Fontaine, C., Furon, T.: Watermarking security, part I: Theory. In: Security, Steganography and Watermarking of Multimedia Contents VII. Volume 5681 of Proceedings of SPIE. (2005) 746–757
7. Cayre, F., Fontaine, C., Furon, T.: Watermarking security, part II: Practice. In: Security, Steganography and Watermarking of Multimedia Contents VII. Volume 5681 of Proceedings of SPIE. (2005) 758–768
8. Doërr, G., Dugelay, J.L.: Security pitfalls of frame-by-frame approaches to video watermarking. IEEE Transactions on Signal Processing, Supplement on Secure Media **52** (2004) 2955–2964
9. Hartung, F., Girod, B.: Watermarking of uncompressed and compressed video. Signal Processing **66** (1998) 283–301
10. Doërr, G., Dugelay, J.L.: A guide tour of video watermarking. Signal Processing: Image Communication, Special Issue on Technologies for Image Security **18** (2003) 263–282
11. 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

12. Doërr, G., Dugelay, J.L.: Secure background watermarking based on video mosaicing. In: Security, Steganography and Watermarking of Multimedia Contents VI. Volume 5306 of Proceedings of SPIE. (2004) 304–314
13. Smolic, A., Lorei, M., Sikora, T.: Adaptive kalman-filtering for prediction and global motion parameter tracking of segments in video. In: Proceedings of the Picture Coding Symposium. (1996)
14. Nicolas, H., Labit, C.: Motion and illumination variation estimation using a hierarchy of models: Application to image sequence coding. Journal of Visual Communication and Image Representation **6** (1995) 303–316
15. Szeliski, R., Shum, H.Y.: Creating full view panoramic image mosaics and environment maps. In: Proceedings of the International Conference on Computer Graphics and Interactive Techniques. (1997) 251–258
16. Sun, Z., Tekalp, M.: Trifocal motion modeling for object-based video compression and manipulation. IEEE Journal on Circuits and Systems for Video Technology **8** (1998) 667–685
17. Koenen, R.: MPEG-4 overview. In: JTC1/SC29/WG11 N4668, ISO/IEC (2002)
18. Fisher, Y.: Fractal Image Compression: Theory and Applications. Springer-Verlag (1994)
19. Rey, C., Doërr, G., Dugelay, J.L., Csurka, G.: Toward generic image dewatermarking? In: Proceedings of the IEEE International Conference on Image Processing. Volume III. (2002) 633–636
20. Petitcolas, F., Kirovski, D.: The blind pattern matching attack on watermarking systems. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. Volume IV. (2002) 3740–3743
21. Kirovski, D., Petitcolas, F.: Blind pattern matching attack on watermarking systems. IEEE Transactions on Signal Processing **51** (2003) 1045–1053
22. Kirovski, D., Petitcolas, F.: Replacement attack on arbitrary watermarking systems. In: Proceedings of the ACM Digital Rights Management Workshop. Volume 2696 of Lecture Notes in Computer Science. (2003) 177–189
23. Doërr, G., Dugelay, J.L., Grangé, L.: Exploiting self-similarities to defeat digital watermarking systems - a case study on still images. In: Proceedings of the ACM Multimedia and Security Workshop. (2004) 133–142
24. Holliman, M., Macy, W., Yeung, M.: Robust frame-dependent video watermarking. In: Security and Watermarking of Multimedia Contents II. Volume 3971 of Proceedings of SPIE. (2000) 186–197
25. Fridrich, J., Goljan, M.: Robust hash functions for digital watermarking. In: Proceedings of the International Conference on Information Technology: Coding and Computing. (2000) 178–183
26. Delannay, D., Macq, B.: A method for hiding synchronization marks in scale and rotation resilient watermarking schemes. In: Security and Watermarking of Multimedia Contents IV. Volume 4675 of Proceedings of SPIE. (2002) 548–554
27. Irani, M., Anandan, P., Bergen, J., Kumar, R., Hsu, S.: Mosaic representations of video sequences and their applications. Signal Processing: Image Communication **8** (1996) 327–351
28. Bas, P., Macq, B.: A new video-object watermarking scheme robust to object manipulation. In: Proceedings of the IEEE International Conference on Image Processing. Volume II. (2001) 526–529
29. Hartung, F., Eisert, P., Girod, B.: Digital watermarking of MPEG-4 facial animation parameters. Computers & Graphics **22** (1998) 425–435
30. Garcia, E., Dugelay, J.L.: Texture-based watermarking of 3D video objects. IEEE Transactions on Circuits and Systems for Video Technology **13** (2003) 853–866

31. Visual Quality Expert Group (VQEG): (http://www.vqeg.org)
32. Macy, W., Holliman, M.: Quality evaluation of watermarked video. In: Security and Watermarking of Multimedia Contents II. Volume 3971 of Proceedings of SPIE. (2000) 486–500
33. Winkler, S., Gelasca, E., Ebrahimi, T.: Towards perceptual metrics for video watermark evaluation. In: Applications of Digital Image Processing. Volume 5203 of Proceedings of SPIE. (2003) 371–378
34. Doërr, G.: Security Issue and Collusion Attacks in Video Watermarking. PhD thesis, Université de Nice Sophia-Antipolis, France (2005)
35. 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. Volume 1768 of Lecture Notes in Computer Science. (1999) 211–236
36. Doërr, G., Dugelay, J.L.: How to combat block replacement attacks? In: Accepted for publication in the 7th Information Hiding Workshop. (2005)
37. Barni, M., Bartolini, F., De Rosa, A., Piva, A.: A new decoder for optimum recovery of nonadditive watermarks. IEEE Transactions on Image Processing **10** (2001) 755–766
38. Cox, I., Kilian, J., Leighton, T., Shamoon, T.: Secure spread spectrum watermarking for multimedia. IEEE Transactions on Image Processing **6** (1997) 1673–1687
39. Foley, J., Legge, G.: Contrast masking in human vision. Journal of the Optical Society of America **70** (1980) 1458–1470