SIGNAL COHERENT WATERMARKING IN VIDEO

Gwenaël Doërr and Jean-Luc Dugelay

Multimedia Communications Department
Eurécom Institute, Sophia-Antipolis, France

ABSTRACT

Collusion approaches in general, and block replacement attacks in particular, have been demonstrated to be a major threat against the security of frame-by-frame embedding strategies in video watermarking. These attacks exploit the redundancy of the host signal to replace each signal block with another perceptually similar one taken from another location. Such an attacking approach can be enforced both at a frame level and at a block level. Two countermeasures will consequently be introduced in this paper to combat this threat. The basic idea consists in forcing the watermark to exhibit the same spatio-temporal self-similarities as the host signal either by taking into account the camera motion or by considering the neighborhood characteristics for each sample.

1. INTRODUCTION

Digital watermarking was introduced in the early 90's as a complementary protection technology [1]. Encryption alone is indeed not enough: encrypted multimedia content is decrypted sooner or later to be presented to human beings and is thus left unprotected. It can be perfectly duplicated, manipulated and redistributed at a large scale. Digital watermarking can consequently be inserted to set a second line of defense. The basic idea consists in hiding some information into digital content in a robust and imperceptible manner. Whereas a lot of research effort has been devoted to evaluate the robustness of embedded watermarks against common signal processing primitives - such as noise addition, filtering, lossy compression - few works have addressed the impact of malicious intelligence. In other terms, even if digital watermarking was first introduced for applications to be deployed in a hostile environment (copyright protection, fingerprinting, etc), security issues have been almost ignored. In fact, robustness and security have been mixed concepts for a very long time in the watermarking community itself [2]. On one side, robustness is only concerned about regular customers who perform common blind signal processing primitives which may, or may not, degrade the watermark signal. On the other side, security has to address the behavior of malicious customers who try to learn some knowledge about the system which can be exploited to defeat or fool the system.

Nowadays, security evaluation has become a major issue in digital watermarking. As a result, researchers try to foresee hostile behaviors from malicious customers to be able to introduce countermeasures before the watermarking system is deployed. In this perspective, collusion attacks have been shown to be a serious threat [2]. Collusion basically consists in collecting several watermarked documents and combining them to obtain unwatermarked content. When video content is studied, this approach is all the more pertinent since frame-by-frame strategies are commonly used as written below:

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

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. Thus, each single video frame can be regarded as a watermarked document by itself. In Section 2, Block Replacement Attacks (BRA) are introduced as a possible mean to confuse watermark detectors. The basic idea is to exploit the redundancy of the host signal to replace each signal frame or block by another one taken at another location. Then, efficient countermeasures using motion compensation and neighborhood characterization are examined in Section 3 and conclusions are drawn in Section 4.

2. BLOCK REPLACEMENT ATTACKS ...

Previous work has shown that a redundant structure can be isolated by an hostile attacker enforcing a collusion strategy [3]. Nevertheless, completely independent watermarking is not the solution either. An attacker can indeed exploit the property that independent watermark samples usually sum to zero to design efficient attacks. In this perspective, the goal of collusion is no longer to identify some hidden structure which enables watermark removal in a second step, but rather to directly estimate the original unwatermarked content. Of course, for fidelity constraints, host contents should be quite similar so that combining them does not introduce perceptible artifacts. Hopefully, video content is redundant enough to enable such an attacking approach. Successive frames are indeed highly similar (Subsection 2.1) and even single video frames exhibit some self-similarities (Subsection 2.2).

2.1. ... At the frame level

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 [4]. From a frame-by-frame point of view, such a strategy can be seen as always embedding a different watermark. The drawback of this approach is that temporal filtering usually succeeds in confusing the watermark detector [5]. 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 $\mathbf{f}_t$. To be able to cope with dynamic content such as fast moving objects and/or camera motion,

Prof. Jean-Luc Dugelay can be reached by email at jld@eurecom.fr.
The authors acknowledge the European Commission for financial support through the IST Program under Contract IST-2002-507932 ECRYPT.

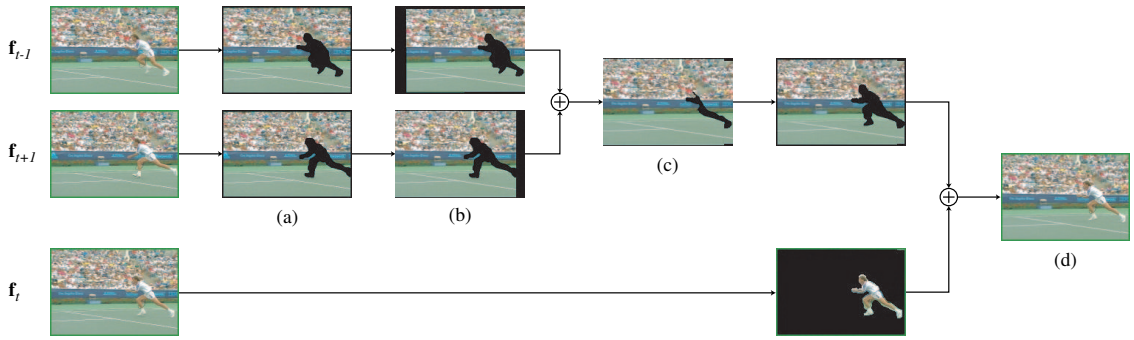

Fig. 1. BRA at the frame level: Once the video objects have been removed (a), neighbor frames are registered (b) and combined to estimate the background of the current frame (c). Next, the missing video objects are inserted back (d).

this simple attacking strategy need to be significantly improved. In particular camera motion has to be compensated to enable Temporal Frame Averaging after Registration (TFAR) [6]. As depicted in Figure 1, TFAR basically aims at estimating the current frame $\mathbf{f}_t$ using the neighbor ones. This is possible because successive video frames taken from a given video shot are different views of the same movie set or, in other words, different 2D projections of the same 3D scene. Of course, moving objects cannot be estimated and should consequently be kept. In summary, TFAR segments moving objects and leaves them untouched on one hand, and estimates the redundant background using the neighbor frames on the other hand. From a coding perspective, this comes down to encoding the background with an advanced forward-backward predictive coder e.g. B-frames in MPEG. Alternatively, it can also be seen 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, TFAR succeeds in removing the watermark. 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 [7].

2.2. ...At the block level

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 compression tools [8]. As a result, in a fractal coding fashion, an attacker can design a BRA which replaces each input signal block with another one taken within a search window and which is highly similar to the input block modulo a geometrical and photometric transformation as depicted in Figure 2. Alternatively, the attacker can also choose to combine multiple blocks to obtain a candidate block for replacement which is similar enough to be exchanged without introducing strong visible artifacts [9]. Anyway, there exists a trade-off between fidelity and attack efficiency. The more (resp. less) similar is the replacement block in comparison with the input one, the less (resp. more) efficient the attack is likely to be. As a result, an adaptive framework can be introduced to adapt to the content of the considered block and thus combine more or less blocks [10]. It is indeed necessary to combine more (resp. fewer) blocks to approximate well-enough a textured (resp. flat) block. Since most algorithms published in the literature do not consider the self-similarities of the signal to be watermarked, BRA usually succeeds in removing embedded watermarks.

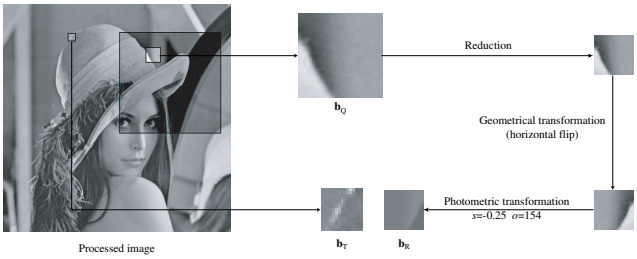

Fig. 2. BRA at the block level: Each signal block is replaced by another perceptually similar one, e.g. modulo a geometrical and photometric transformation taken at another location.

3. SIGNAL COHERENT WATERMARKING

On one hand, a redundant watermarking structure can be estimated by collecting several watermarked uncorrelated documents. On the other hand, uncorrelated watermarks can be removed by averaging similar watermarked documents. These observations intuitively lead to the well-known 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 [11], a frame-dependent binary string can be exploited to generate a watermark pattern which degrades gracefully with an increased number of bit errors [12, 13], the watermark can be embedded in some frame-dependent positions [5]. However, is it enough to ensure that embedded watermarks will survive to the attacks presented in Section 2? A video watermarking scheme should carefully consider camera motion to resist TFAR i.e. the watermark should *move* with the camera (Subsection 3.1). Furthermore, spatial self-similarities should also be examined to resist BRA at the block level: if a pattern (flower, head) is repeated in a frame, it should always carry the same watermark (Subsection 3.2).

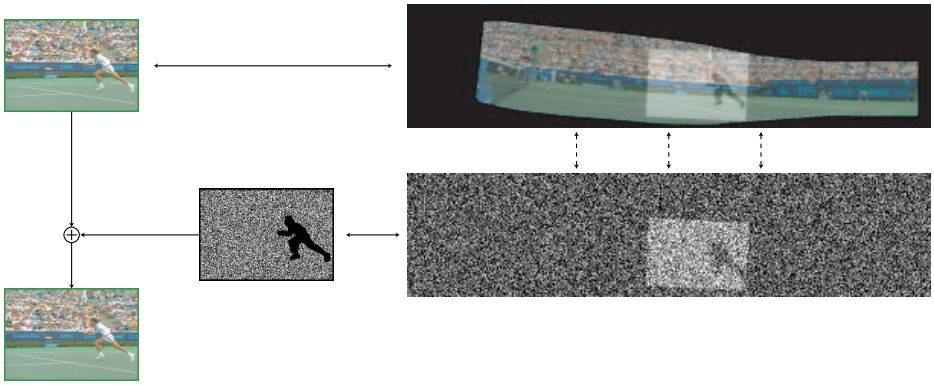

Fig. 3. Embedding procedure for camera motion coherent watermarking: The part of the watermark pattern which is associated with the current video frame is retrieved and registered back. Next, it is embedded in the background portion of the video frame.

3.1. Motion Compensated Watermarking

For a given scene, backgrounds of video frames can be considered as several 2D projections of the same 3D set. TFAR basically exploits the fact that common watermarking schemes do not consider camera motion at all. As a result, a given 3D point which is projected in different locations in different video frames is associated with uncorrelated watermark samples and averaging registered video frames succeeds in confusing the watermark detector. A possible way to circumvent this pitfall is 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. Video mosaicing can be exploited to this end as depicted in Figure 3 [6]. First, for each video frame, some warping parameters θ_t are computed to associate the frame background with a portion of the mosaic. Next, a key-dependent pattern $\mathbf{p}$ is generated which as the same dimensions as the mosaic representation of the video shot and the portion $\mathbf{p}_t$ associated the current frame is retrieved. Finally, $\mathbf{p}_t$ is registered back using the warping parameters to obtain the watermark to be embedded in the current video frame. Furthermore, moving objects do not host watermark samples to follow the underlying strategy: *a 3D point carries the same watermark sample all along the video scene*. The embedded watermark can consequently be written:

$$\mathbf{w}_t = \mathbf{m}_t \otimes \mathbf{p}_t^{(\theta_t)} \tag{2}$$

where $\mathbf{m}_t$ is a binary mask which discriminates moving objects from the background and $\otimes$ denotes the pixelwise multiplication. On the receiver side, the detector only checks whether the expected watermark $\mathbf{w}_t$ has been effectively embedded or not using for instance a correlation score. In summary, motion compensated watermarking simulates an utopian world where the movie set would already be watermarked. It can consequently be related with other works which use texture watermarking to protect the usage of 3D objects rather than their 3D structure [14]. Such watermarking strategies have been shown to resist TFAR and have also exhibited interesting properties in terms of imperceptibility [6].

3.2. Self-similar watermarks

BRA exploit the fact that watermarking algorithms do not consider the self-similarities of the host signal during embedding. As a result, similar signal blocks are likely to carry uncorrelated watermark samples and exchanging them confuses the detector. Intuitively, if *similar signal blocks carry similar watermarks*, BRA are likely to be ineffective. In other terms, the embedded watermark should inherit the self-similarities of the host signal and the next subsections present two alternative ways to meet this specification.

3.2.1. From linear watermarking in Gabor space ...

The requirement *similar signal blocks should carry similar watermarks* can also be rephrased as *pixels with similar neighborhood should carry watermark samples with close values*. Under this new light, the very first task is to compute some features which can be used to characterize neighborhoods and Gabor features are a good candidate [15]. A Gabor Elementary Function (GEF) $\mathbf{h}_{\rho,\theta}$ is basically a complex 2D sinusoid whose orientation and frequency are given be (θ, ρ) restricted by a Gaussian envelope and the response $\mathbf{g}_t^{\rho,\theta}$ of a video frame to a GEF is given by:

$$\mathbf{g}_t^{\rho,\theta} = \mathbf{f}_t * \mathbf{h}_{\rho,\theta} \tag{3}$$

where $*$ denotes convolution. For computational complexity reasons, Gabor filtering is usually performed is the FFT domain since it then comes down to a simple multiplication and GEF are paired ($\mathbf{h}_{\rho,\theta} \leftarrow \mathbf{h}_{\rho,\theta} + \mathbf{h}_{\rho,\theta+\pi}$) to obtain real-valued features. Then, defining a linear form $\varphi(.)$ on the Gabor space is enough to obtain the desired property [16]. In other terms, if M frequencies and N orientations are considered in the Gabor filter bank, the secret key can be used to generate the MN values $\psi_{i,j}$ that $\varphi(.)$ takes on the canonical basis and the embedded watermark can be written:

$$\mathbf{w}_t \propto \sum_{i=1}^{M}\sum_{j=1}^{N} \psi_{i,j}\mathbf{g}_t^{\rho_i,\theta_j} \tag{4}$$

where the watermark $\mathbf{w}_t$ is normalized to have unit variance. Such linear watermarks have been demonstrated to be almost immune to BRA at the block level and the influence of the number MN of GEF in the filterbank has little influence. Nevertheless, it is still required to have a large enough number of GEF to ensure that watermarks generated with different keys are not correlated. Of course, increasing the number of GEF also raises the computational load and a trade-off has to be found.

3.2.2. . . . To multiplicative watermarking

When Equation (4) is rewritten in the FFT domain, due to the linearity of the Fourier transform, the following formula is obtained:

$$\mathbf{W}_t \propto \Big(\sum_{i=1}^{M} \sum_{j=1}^{N} \psi_{i,j} \mathbf{H}_{\rho_i,\theta_j} \Big) \mathbf{F}_t = \mathbf{H}\mathbf{F}_t \qquad (5)$$

where capital letters are used for variables in the FFT domain. In other terms, the watermark generation process can be seen as a multiplication in the frequency domain [17]. In the frequency domain, $\mathbf{H}_{\rho,\theta}$ can be seen as two 2D Gaussian shifted by $\pm\rho$ frequency units and rotated by an angle θ. The bandwidth of the GEF is regulated by the variances σ_ρ and σ_θ and the more GEF there are in the filterbank, the tighter is the bandwidth. In the limit case, the GEF is limited to a Dirac impulse and the watermark generation process comes down to a multiplication in the FFT domain with a symmetric watermark [18]. Furthermore, keeping in mind that DCT coefficients can be considered as FFT coefficients [19], multiplicative watermarking in the DCT domain should also produce signal coherent watermarks and this statement has been confirmed experimentally [17].

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, BRA have been introduced as the possible result of an attacking strategy based on collusion and two countermeasures have been proposed to circumvent this threat. 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 [20]. It could be interesting to investigate in the near future whether this approach can be extended to also consider the case when uncorrelated watermarks are used.

5. REFERENCES

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

[2] G. Doërr and J.-L. Dugelay, "Collusion issue in video watermarking," in *Security, Steganography and Watermarking of Multimedia Contents VII*, January 2005, vol. 5681 of *Proceedings of SPIE*.

[3] G. Doërr and J.-L. Dugelay, "Security pitfalls of frame-by-frame approaches to video watermarking," *IEEE Transactions on Signal Processing, Supplement on Secure Media*, vol. 52, no. 10, pp. 2955–2964, October 2004.

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

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

[6] G. Doërr and J.-L. Dugelay, "Secure background watermarking based on video mosaicing," in *Security, Steganography and Watermarking of Multimedia Contents VI*, January 2004, vol. 5306 of *Proceedings of SPIE*, pp. 304–314.

[7] R. Koenen, "MPEG-4 overview," in *JTC1/SC29/WG11 N4668*. ISO/IEC, March 2002.

[8] Y. Fisher, *Fractal Image Compression: Theory and Applications*, Springer-Verlag, 1994.

[9] D. Kirovski and F. Petitcolas, "Blind pattern matching attack on watermarking systems," *IEEE Transactions on Signal Processing*, vol. 51, no. 4, pp. 1045–1053, April 2003.

[10] G. Doërr, J.-L. Dugelay, and L. Grangé, "Exploiting self-similarities to defeat digital watermarking systems - a case study on still images," in *Proceedings of the ACM Multimedia and Security Workshop*, September 2004, pp. 133–142.

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

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

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

[14] E. Garcia and J.-L. Dugelay, "Texture-based watermarking of 3D video objects," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 8, pp. 853–866, August 2003.

[15] J. Daugman, "Complete discrete 2-D Gabor transforms by neural network for image analysis and compression," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 36, no. 7, pp. 1169–1179, July 1988.

[16] G. Doërr and J.-L. Dugelay, "A countermeasure to resist block replacement attacks," in *Submitted for publication to the IEEE International Conference on Image Processing*, September 2005.

[17] G. Doërr and J.-L. Dugelay, "How to combat block replacement attacks?," in *Submitted for publication to the 7th Information Hiding Workshop*, June 2005.

[18] M. Barni, F. Bartolini, A. De Rosa, and A. Piva, "A new decoder for optimum recovery of nonadditive watermarks," *IEEE Transactions on Image Processing*, vol. 10, no. 5, pp. 755–766, May 2001.

[19] J. Lim, *Two-Dimensional Signal and Image Processing*, Prentice Hall PTR, 1989.

[20] F. Cayre, C. Fontaine, and T. Furon, "Watermarking security, part I: Theory," in *Security, Steganography and Watermarking of Multimedia Contents VII*, January 2005, vol. 5681 of *Proceedings of SPIE*.