On the Need for Signal-Coherent Watermarks

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

***Abstract*— Digital watermarking has been introduced in the 90's as a complementary technology for copyright protection. In an effort to anticipate hostile behavior of adversaries, the research community is constantly introducing new attacks to benchmark watermarking systems. In this paper, we present a generic attack strategy based on block replacement. As multimedia content is often highly repetitive, the attack exploits signal's self-similarities to replace each signal block with another, perceptually similar one. Guided by the principles of the proposed attack framework, we implemented three attack algorithms for different types of multimedia content: video shots, audio tracks and still images. Finally, considering the effectiveness of the proposed algorithms, we identify the properties that a watermark should have to counter this attacking strategy.**

***Index Terms*— Self-similarities in multimedia data, block replacement attack, signal-coherent watermarks**

I. Introduction

RECENTLY, security has become an important issue in several multimedia applications. For example, automated video surveillance is used to enforce monitoring of protected facilities; biometric systems are exploited to aid in person identification, etc. In addition, the proliferation of the Internet has triggered a substantial increase in multimedia content piracy. In particular, peer-to-peer systems have been effective in sharing music and video content world-wide without the ability to generate revenues to the copyright owners. This has put pressure on the technical community to rethink their entire content distribution framework. In particular, initiatives have been launched to deploy and standardize a Digital Rights Management (DRM) technology to protect playback, storage and distribution of multimedia items [1], [2]. The challenge is that encryption alone is not enough to ensure copyright at the client side. As soon as data is decrypted or unscrambled, the adversary obtains a plain-text copy of the multimedia item and can either copy it in its digital form or digitize it from an analog output using an A/D converter. This threat has motivated the introduction of digital watermarks in almost all modern copyright protection mechanisms [3].

Digital watermarking consists of hiding a key-dependent secret signal into digital multimedia data in a robust and imperceptible manner. The robustness of the watermark can be seen as the ability of the detector to retrieve the hidden watermark once watermarked data has been altered within perceptual bounds. For example, the embedded signal should survive D/A-A/D conversion. Watermarks can also be regarded as a communication channel whose capacity is exploited to convey information. Those parameters (capacity, imperceptibility, robustness) are conflicting and a trade-off has to be found depending on the targeted applications. Introducing watermarks in digital data can be useful to safeguard copyright. In a *content screening* scenario, content providers insert a secret watermark in their multimedia items, before releasing them on a public communication channel. On the client side, the media player blindly checks whether the watermark is present or not. In case the secret mark is detected, the player verifies whether it has an authentic and valid license to (dis)play the content. Alternatively, user-specific watermarks denoted as *fingerprints* can be embedded in the data to be protected, before being delivered to the customer. Search robots are then deployed to find illegal copies on the Internet and forensic tools are exploited to identify malicious customers who have broken their license agreement.

In such applications, the embedded watermark either limits the possible usage of multimedia content via playback or copy control, or points out at pirates' identities - an information which may be potentially used in court. As a result, malicious users are likely to attack this protection technology. The competition between attackers and security providers is common in security fields and usually results in continuously improving protection technologies. Thus, the research community makes efforts to anticipate hostile behavior and releases benchmarking tools to evaluate the robustness of watermarking techniques [4]–[6]. After a short review of already developed attacks against watermarking systems, the generic framework of the block replacement attack is presented in Section II. The basic idea consists of replacing each signal block with a perceptually similar one. In the subsequent sections, alternative types of multimedia with different levels of self-similarity, are considered: similar background in successive video frames, repetitive patterns in audio tracks, and similarities modulo some transformations in still images. The introduced generic framework of the attack is then implemented for each type of data to demonstrate its efficiency. Finally, in Section VI, an intuitive specification, referred to as signal-coherent watermarks, is introduced to establish the basic requirements for a watermarking technology to resist attacks that exploit signal self-similarities.

II. Attacks against Digital Watermarking Systems

There exists a relatively complex trade-off between conflicting parameters in digital watermarking. As a result, several

Manuscript received July 13, 2004; revised June 14, 2005. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Benoit Macq.

G. Doërr and J.-L. Dugelay are with the Eurécom Institute, Sophia-Antipolis, France (email: {doerr, dugelay}@eurecom.fr).

D. Kirovski is with Microsoft Research, Redmond, WA 98052, USA (email: darkok@microsoft.com)

Digital Object Identifier 10.1109/TMM.0000.000000

benchmarks have been released to allow a fair comparison between different algorithms. In this perspective, efficient attacks have been proposed in an attempt to anticipate malicious behavior. In content protection systems, embedded watermarks do not add any value to the multimedia items from the customer perspective. On the contrary, the hidden information can be used in court to convince a jury that a suspect has not respected the license agreement. Therefore, malicious users want to remove this hidden evidence and design efficient attacks to defeat the system. Attacks against digital watermarking systems are consequently reviewed below before introducing a new attacking strategy.

A. Signal Removal Attacks

In digital watermarking, the detector usually computes a correlation score, which is then compared to a threshold to assert whether a watermark is present or not. Consequently, a potential target for an adversary is to find an attack which brings the detection score below the detection threshold so that the embedded watermark is no longer detected. In other terms, the attack aims at decreasing the power of the hidden watermark down to a level where the detector cannot *reliably* assert that it is present. A large range of common signal processing operations can be considered as candidate removal attacks. Low-pass filtering and lossy compression are likely for instance to alter watermarks since they are usually located in high frequencies. However, many other primitives have to be considered to obtain a fair benchmark [7] including gamma correction, quantization, noise addition due for example to D/A-A/D conversion (printing and scanning), etc. Besides such blind operations, a new brand of attacks based on estimation theory has appeared. The basic idea consists in estimating either the original unwatermarked content or the embedded watermark itself. For example, denoising techniques can be exploited to remove a hidden watermark. However, such approaches have been shown to introduce blurring artifacts. Thus, a two-step strategy is usually preferred. First, the attacker estimates the embedded watermark using, for example, local median [8] or Wiener filtering [9]. Then, this estimation can be processed, e.g., high-pass filtered [8], to remove unlikely low-frequency components. Finally, the estimated watermark is remodulated either with constant [8], [9] or adaptive strength [10] which considers perceptual constraints.

B. Synchronization Removal Attacks

This second class of attacks does not explicitly aim at removing the embedded watermark. It rather tries to disrupt the communication between the embedder and the detector. To this end, the attacker basically performs operations which desynchronize the detector. Indeed, many detectors today are correlation based and thus expect each watermark sample to be at a predefined location in the working space. This knowledge is shared by both the embedder and the detector. If an external party disturbs this alignment, the convention shared with the detector becomes obsolete and communication is no longer possible. Consequently, spatial and temporal alterations can be performed on the watermarked data to trap the detector. Examples of such operations include image flipping, rotations, cropping, scaling, time stretching, etc. The most powerful class of desynchronization attacks are locally random geometric distortions. The most well-known implementation of such an attack is the random bending attack [11]. It exploits the fact that the human visual system is not sensitive against shifts and local affine transformations. Therefore, pixels can be locally shifted, scaled and rotated without significant visual distortions as depicted in Figure 1. Such a strategy still succeeds in confusing most of the watermark detectors today.

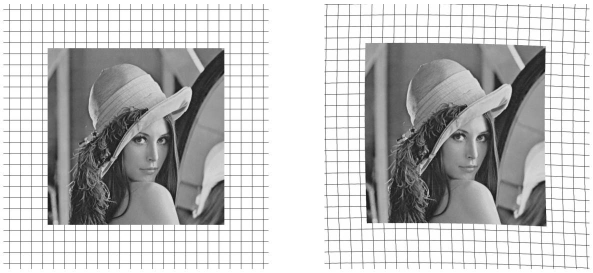


Fig. 1. Stirmark attack (left: original, right: attacked): If the central parts look really similar, distortions are obvious when the grid is considered.

C. Block Replacement Attacks

Both classes of attacks previously presented exhibit some shortcomings. On one side, although common signal processing operations decrease the correlation score, they may also introduce undesirable distortion artifacts. Furthermore, watermark estimation and remodulation attacks rely on the assumption that it is possible to obtain a relatively *accurate* estimation of the embedded watermark. While such a refined estimation can be computed when several watermarked documents are colluded [12], [13], it is relatively difficult to achieve the same goal when a single watermarked document is considered. On the other side, desynchronization attacks do not remove the hidden watermark. They simply alter the alignment shared by the embedder and the detector. However, nothing ensures that a future enhanced version of the detector is not going to be able to detect the desynchronized watermark. Moreover, the introduced misalignment prevents from using common quantitative metrics such as the Peak Signal to Noise Ratio (PSNR) to evaluate the impact of the attack.

Those limitations have consequently motivated the introduction of a novel attack. Ideally, the attack would consist of blindly restoring the original document from the watermarked one. However such a perfect attack is impossible to implement in practice. In this paper, the goal is to design an attack with the following specifications:

(i) the detector should no longer be able to detect the embedded watermark after the attack,
(ii) the attack should not introduce geometric distortions so that quantitative measures of distortion can still be used[1],

[1] It does not mean that geometrical attacks are irrelevant. The point is only to say that distortion is easier to evaluate when there is no desynchronization. In the case of StirMark for instance, it is difficult to say when the distortion is *critical* or not.

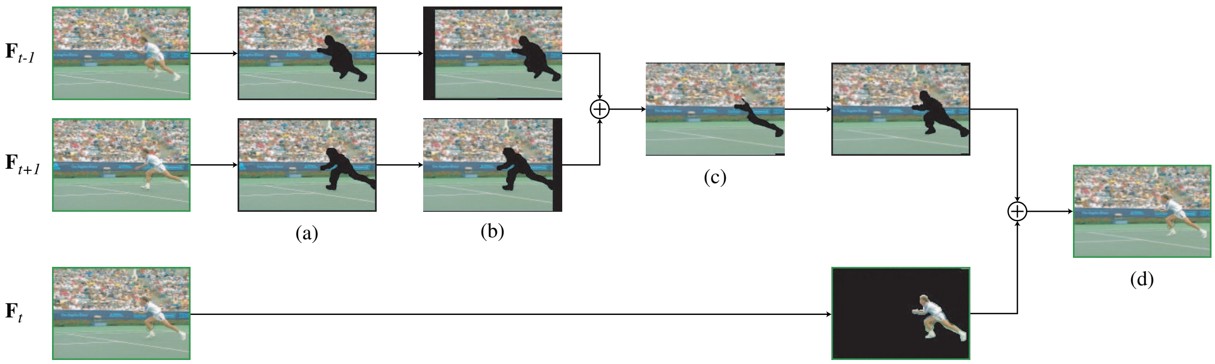


Fig. 2. Temporal Frame Averaging after Registration (TFAR): Once the video objects have been removed (a), neighbor frames are registered (b) and combined to estimate the background of the current frame (c). Next, the missing video objects are inserted back (d).

(iii) the attack should introduce a fair additional distortion, e.g., the distance between the watermarked and attacked documents should be close to the distance between the original and watermarked documents,

(iv) the attack should be designed in such a way that it is possible to adapt the strength of the attack to cope with alternative watermarking schemes using different embedding strength,

(v) the attack should ensure that a future improved version of the detector alone cannot overcome the problem. Hence, both the embedder and the retriever must be revised to re-enable content protection.

TABLE I
GENERIC DESCRIPTION OF THE BLOCK REPLACEMENT ATTACK

1	Partition the input signal into blocks	
2	For each input block,	
	(a)	Define a search window and build a codebook which contains a collection of blocks
	(b)	Compute a replacement block *similar* to the input one using the blocks in the codebook
	(c)	Replace input block by the replacement one

Multimedia digital data is highly redundant. For example, successive frames are highly similar in a video shot; similarly, most songs often contain repetitive patterns. Thus, an attacker can exploit these similarities to replace each part of the signal with a *similar* one taken from another location in the same signal. The generic block replacement strategy is further detailed in Table I. The input signal is first partitioned into a set of blocks. Then, for each input block, a search window is defined and also partitioned to obtain a codebook containing a collection of blocks. Next, this codebook is exploited to obtain a replacement block similar to the input one, but which does not directly carry the watermark signal. This is where the presented attack differs from previous works which also considered block replacement, but for copying the watermark embedded in a document to another unwatermarked one [14]. In this paper, the Mean Square Error (MSE) is used to evaluate the level of similarity between two blocks $\mathbf{B}_1$ and $\mathbf{B}_2$:

$$\mathrm{MSE}(\mathbf{B}_1, \mathbf{B}_2) = \frac{1}{N} \sum_{i=1}^{N} \big(\mathbf{B}_1(i) - \mathbf{B}_2(i)\big)^2 \qquad (1)$$

where the index i in the summation can be one-dimensional (sound) or multidimensional (image, video) and N is the number of considered signal samples. Finally, the input block is substituted by the replacement. The subsequent sections present alternative implementations of this block replacement strategy for different types of multimedia data (video, sound, images). Although the proposed algorithms share the same principles, their implementations require specific modifications to address the structural and perceptual nature of considered multimedia signals.

III. CASE STUDY WITH VIDEO SHOTS

Video content is recognized to be highly redundant. In particular, the background which usually occupies most of the scene is repeated across successive video frames. As a result, for a given video frame, neighboring frames often entail an estimation of the current one. This property has been extensively used to obtain efficient coding tools. In terms of the block replacement attack, each video frame is processed sequentially: neighboring frames are gathered and combined, e.g., averaged to obtain a replacement frame [15], [16]. To this end, video processing tools such as object-based segmentation and frame registration have to be introduced to cope with moving objects and camera motion. The block replacement attack applied to video content is depicted in Figure 2 and further detailed hereafter.

A. Temporal Frame Averaging after Registration (TFAR)

The goal is to estimate a given video frame $\mathbf{F}_t$ from its neighboring ones $\mathbf{F}_{t+\delta}$ using frame registration. The considered neighboring 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 $\mathbf{B}_t$) from useless ones (e.g., moving and varying objects $\mathbf{O}_t$):

$$\mathbf{O}_t = \mathbf{F}_t \otimes \mathbf{M}_t \quad \text{and} \quad \mathbf{B}_t = \mathbf{F}_t \otimes \bar{\mathbf{M}}_t, \qquad (2)$$

where $\otimes$ is the pixel-wise multiplication operator and $\bar{\cdot}$ is the binary negation operator. In this paper, no specific work has been done to design a novel object-based segmentation system. We use an existing algorithm based on semi-automatic initial segmentation of the first video frame, followed by automatic tracking of the selected objects [17].

Once several observations $\mathbf{B}_{t'}$ of the movie set are 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 each pixel position (x_t, y_t) in the current frame $\mathbf{F}_t$ with a corresponding position $(x_{t'}, y_{t'})$ in a neighboring frame $\mathbf{F}_{t'}$. The association can be based, for example, to minimize the MSE 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 the following first order polynomial motion model [18]:

$$\left\{ \begin{array}{l} x_{t'} = t_x + z(x_t - x_o) - z\theta(y_t - y_o) \\ y_{t'} = t_y + z(y_t - y_o) + z\theta(x_t - x_o) \end{array} \right. , \qquad (3)$$

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.

The 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. To this end, the registered backgrounds are averaged using the proper normalization factor for each pixel. For each pixel position $\mathbf{p}$, a binary mask $\mathbf{R}_t$ is also built to indicate whether a background value has been effectively estimated ($\mathbf{R}_t(\mathbf{p}) = 1$) or not ($\mathbf{R}_t(\mathbf{p}) = 0$). The moving objects are then inserted back:

$$\dot{\mathbf{F}}_t = \underbrace{\tilde{\mathbf{B}}_t \otimes \bar{\mathbf{M}}_t}_{\text{Background}} + \underbrace{\mathbf{F}_t \otimes \mathbf{M}_t}_{\text{Objects}} + \underbrace{\mathbf{F}_t \otimes (\bar{\mathbf{M}}_t \ \& \ \bar{\mathbf{R}}_t)}_{\text{Missing pixels}}, \qquad (4)$$

where $\&$ is the binary AND operator. The last term indicates 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.

B. Experiments

Today, watermarking digital video content is regarded most of the time as watermarking a sequence of still images. Without loss of generality, such a frame by frame approach is illustrated below with a simple additive spread spectrum watermark:

$$\check{\mathbf{F}}_t = \mathbf{F}_t + \alpha \mathbf{W}_t(K), \quad \mathbf{W}_t(K) \sim \mathcal{N}(0,1) \qquad (5)$$

where t is the frame index, $\mathbf{F}_t$ (resp. $\check{\mathbf{F}}_t$) is the luminance component of the original (resp. watermarked) video frame, α the embedding strength, and $\mathbf{W}_t(K)$ a key-dependent normally distributed watermark with zero mean and unit variance. Perceptual shaping can subsequently be introduced to improve the imperceptiveness of the watermark. On the detector side, the presence or absence of a watermark is checked with a correlation score:

$$\rho(\check{\mathbf{F}}_t, K) = \big(\mathbf{F}_t + \epsilon\alpha\mathbf{W}_t(K)\big) \cdot \mathbf{W}_t(K) \approx \epsilon\alpha \qquad (6)$$

where $\cdot$ is the linear correlation operator and ϵ is equal to 0 or 1 depending on whether the tested luminance component $\check{\mathbf{F}}_t$ is watermarked or not. Finally, the computed correlation score is compared to a threshold τ_{detect} to assert whether or not a watermark has been embedded. This threshold can, for instance, be set to $\alpha/2$ to have equal false positive and false negative probabilities.

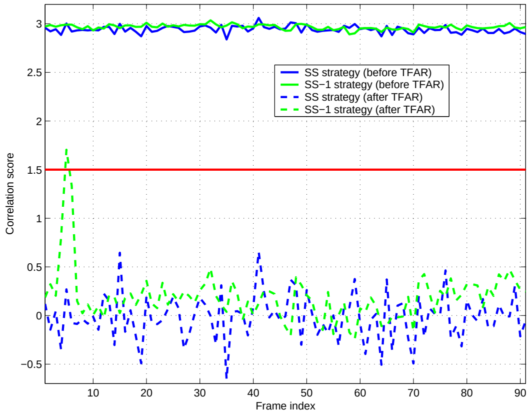


Fig. 3. Impact of the TFAR attack with usual frame-by-frame embedding strategies.

When a frame-by-frame approach is enforced, two major embedding strategies are usually observed: either a different watermark is inserted in each video frame (SS strategy [19]), or the same watermark is embedded in all video frames (SS-1 strategy [20]). The video sequence *stefan* has consequently been watermarked using both embedding strategies and using an embedding strength $\alpha = 3$. The resulting watermarked sequences have then been submitted to the TFAR attack. Figure 3 depicts the detection score computed before and after the attack. It is clear that TFAR makes the detection score drop below the threshold $\tau_{\text{detect}} = 1.5$. It should be noted that a peak around the 5th frame indicates that the SS-1 strategy survives in part the attack. At this moment, there is indeed almost no camera motion and video frames watermarked with the SS-1 strategy are thus not affected by TFAR.

IV. Case Study with Audio Tracks

Repetition is often a principal part of composing music and is a natural consequence of the fact that distinct instruments, voices and tones are used to create a soundtrack. Thus, one is likely to find similar blocks of music within a single musical piece, an album of songs from a single author, or in instrument solos. Finding similarities within musical pieces is a challenging task. First, music has a lot less redundancy than video. In addition, musical redundancies are often superimposed. Next, it is difficult to model redundancies in music as opposed to, for example, video. In particular, music performed by humans commonly exhibits repetitions with significant variance in amplitude, pitch, and timing. Detecting such similarities is a complex task which is not addressed by modern psycho-acoustic models. Although perceptually similar blocks of music can be quite distant in the Euclidean sense, the Human Auditory System (HAS) may perceive them as equivalent. On the other hand, the fact that high fidelity music tolerates significant multiplicative noise levels, makes block replacement attacks quite viable for audio content [21].

A. Watermark Estimation Remodulation

A post-processing step is added to the generic block replacement strategy defined in Table I which provides additional attack efficacy linearly proportional to the introduced noise. More formally, let $\check{\mathbf{x}}$ denote the marked signal and $\dot{\mathbf{x}}$ the signal created by the block replacement procedure. Assuming that $\dot{\mathbf{x}}$ is at a small distance from $\check{\mathbf{x}}$, a low-energy attack vector $\mathbf{d} = \dot{\mathbf{x}} - \check{\mathbf{x}}$ is created whose direction is opposite with respect to the secret watermark $\mathbf{w}$ embedded in $\check{\mathbf{x}}$. If the watermark detector yields the following expected normalized correlation scores $\rho(\check{\mathbf{x}}) = 1$ and $\rho(\dot{\mathbf{x}}) = a$, $0 < a < 1$, then $\rho(\mathbf{d}) = a - 1 < 0$ under the assumption of model linearity. For example, one large class of such correlators are matched filters used in spread-spectrum watermarking [3], [22]. This vector is subsequently scaled by β and remodulated to obtain $\mathbf{y} = \check{\mathbf{x}} + \beta\mathbf{d}$ whose correlation equals $\rho(\mathbf{y}) = 1 + \beta(a - 1)$. Clearly, for strong enough $\beta = \frac{1}{1-a}$, the adversary can set the correlation of $\mathbf{y}$ to zero.

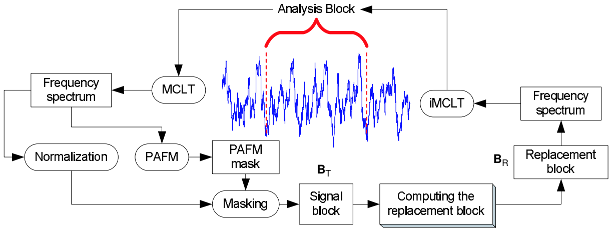


Fig. 4. Audio processing primitives performed as pre- and post-processing to the attack.

The first objective of the adversary is to obtain an accurate estimate of the original signal in $\check{\mathbf{x}}$. To this end, the block replacement strategy is exploited. For each input block $\mathbf{B}_T$, a set $\mathcal{Q}$ of N blocks $\mathbf{B}_{Q_i}$ at a *perceptually similar* distance is identified, i.e., $\text{MSE}(\mathbf{B}_{Q_i}, \mathbf{B}_T) \leq \varepsilon$. The blocks in $\mathcal{Q}$ are then combined in such a way that the similarity between the obtained replacement block $\mathbf{B}_R$ and $\mathbf{B}_T$ is maximized. In other terms, the goal is to find mixing coefficients λ_i so that $\text{MSE}(\mathbf{B}_R, \mathbf{B}_T)$ is minimized, where:

$$\mathbf{B}_R = \sum_{i=1}^{N} \lambda_i \mathbf{B}_{Q_i}. \tag{7}$$

This optimization goal can be modeled as the least-squares problem which can be solved optimally. Per sample $\mathbf{B}_R(j)$, two cases can occur:

(i) $\mathbf{B}_R(j)$ satisfies the fidelity restriction $\text{MSE}(\mathbf{B}_T(j), \mathbf{B}_R(j)) \leq \varepsilon$, in which case the replacement sample is kept,

(ii) $\mathbf{B}_R(j)$ is perceptually not similar to $\mathbf{B}_T(j)$, i.e., $\text{MSE}(\mathbf{B}_T(j), \mathbf{B}_R(j)) > \varepsilon$, in which case we propose to set $\mathbf{B}_R(j) = \mathbf{B}_T(j)$ to minimize the noise due to the attack.

In addition, after the initial approximation, the algorithm excludes all samples of type-(*ii*) from $\mathbf{B}_T$ and all blocks in $\mathcal{Q}$. Then, it recomputes the approximation of $\mathbf{B}_T$ using $\mathcal{Q}$ more accurately.

The second objective is to estimate the value of β which sets $\rho(\mathbf{y}) \approx 0$. Improved accuracy of the estimate of the original signal, lowers a, requires less amplified $\mathbf{d}$, and eventually imposes less noise in the attacked clip $\mathbf{y}$ with respect to the original. In the presence of the watermark detector, this problem is trivial and can be resolved using a simple binary search for β. Alternately, β can be estimated using the following example strategies:

(i) $\beta = 1-\gamma$, where γ is the ratio of signal samples modeled at a certain quality, e.g., $\text{MSE}(\check{\mathbf{x}}(j), \dot{\mathbf{x}}(j)) \leq \varepsilon$,

(ii) by watermarking the available marked content $\check{\mathbf{x}}$ with another watermark $\mathbf{v}$ and then attacking $\check{\mathbf{x}} + \mathbf{v}$ using the block replacement attack. Since the detector in this case is available, identifying β is straightforward under these assumptions.

For brevity, in this version of the manuscript, we do not analyze empirically this process.

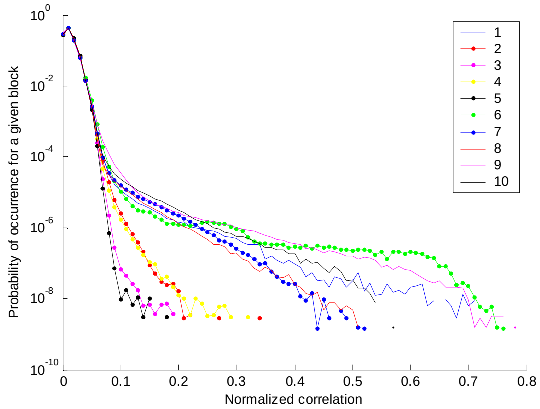


Fig. 5. Music self-similarity: probability that for a given 2048-long MCLT block, another block is found at a certain correlation within the subsequent 60 seconds of the same audio clip.

B. Experiments

Since most psycho-acoustic models operate in the frequency spectrum [23], the attack is launched in the logarithmic (dB) frequency domain. The set of signal blocks is filtered using a Modulated Complex Lapped Transform (MCLT) [24]. We consider MCLT analysis blocks with 1024 transform coefficients and a 50% overlap. Each block of coefficients is normalized and psycho-acoustically masked using an off-the-shelf masking model (PAFM) [23]. Similarity is explored exclusively in the audible part of the frequency sub-band where watermarks are hidden. In this paper, this sub-band is bounded within 2-7 kHz [22]. Figure 4 illustrates the signal processing primitives used to prepare the audio blocks for substitution.

We searched for similar blocks in the time domain. The correlation of each target block $\mathbf{B}_T$ is computed by convolving the complex conjugate of $\mathbf{B}_T$ with the corresponding search window using the Fast Fourier Transform and the overlap-add fast convolution method [25]. Both $\mathbf{B}_T$ and the search region were filtered using a band-pass within 2-7 kHz. Figure 5 illustrates the probability that one can find a similar block of certain correlation with respect to the target within a 60-second search region that follows the target block in the same audio clip. It can be noted that clips created electronically have been identified with substantially greater self-similarities than clips performed entirely by humans.

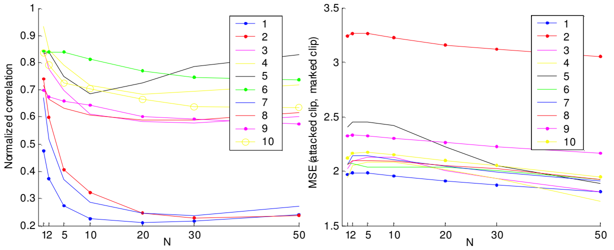


Fig. 6. Resulting correlation (left subfigure) and MSE with respect to marked clip (right subfigure) after the block replacement attack.

In order to evaluate the effect of the attack, we used Direct Sequence Spread-Spectrum (DSSS) sequences with a δ=1 dB amplitude, that spread over 240 consecutive 2048-long MCLT blocks, where only the audible frequency magnitudes in the 2-7 kHz sub-band were marked. In the experiments presented, the first 60 seconds of the following audio clips are considered:

1 Ace Of Base, *Ultimate Dance Party 1999*, Cruel Summer [Blazin' Rhythm Remix],
2 Steely Dan, *Gaucho*, Babylon Sisters,
3 Pink Floyd, *The Wall*, Comfortably Numb,
4 Dave Matthews Band, *Crash*, Crash Into Me,
5 Unidentified classical piece, produced by SONY Ent., selected because of exceptional perceptual randomness, available upon request from the authors.
6 Daft Punk, *Discovery*, One More Time,
7 The Prodigy, *The Fat of the Land*, Breathe,
8 Khaléd, *Arabica*, Didi [Didi Funk Club Remix],
9 Paul Oakenfold, *Transport*, El Niño, and
10 Groove Foundation Press Ku, *Amnesia - Ibiza 2001*, That Feeling.

In Figure 6, we present, for all clips from our benchmark suite, the resulting correlation and MSE with respect to the marked clip after the block replacement attack. We used $1 \leq N \leq 50$. The achieved correlations are higher than in the case of previous work [21] at the benefit of significantly decreased MSE noise. This result has been exploited to drive β such that the resulting normalized correlation score equals $\rho = 0.5$ and $\rho = 0$. For both cases, in Figure 7 we present the resulting MSE after such β is applied. One can observe that in all but one case (**4**, $\beta|\rho = 0$), the desired normalized correlation score was achieved with MSE lower than 3.5 dB. For most listeners such noise is typically imperceptible in the considered sub-band. The result can be further improved by considering larger audio databases than the 60 second search region we used.

V. Case Study with Still Images

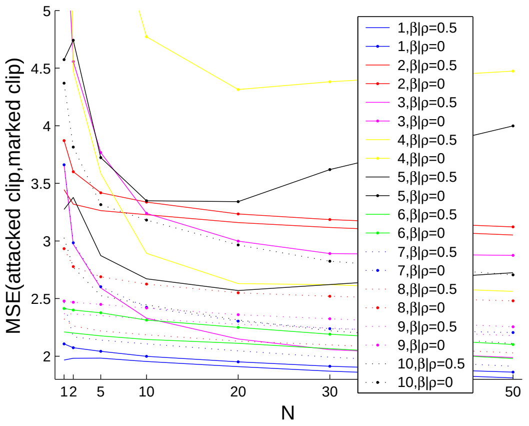


Fig. 7. Resulting noise after the β-amplified watermark estimate is subtracted. Two cases are considered for each clip: β such that resulting correlation is halved and zeroed out.

In the context of still images, self-similarities have been previously exploited to obtain efficient coding tools [26]. On the coder side, for each image block, the most similar block in the search window modulo a geometrical (scaling, rotation, flip) and a photometric transform is found as depicted in Figure 8. Then, starting from any image, the decoder applies those transformations several times and converges towards the original image. This strategy is similar to the presented block replacement strategy. In particular, early work has shown that the candidate block could not be obtained using a single block [27] for fidelity issues. A straightforward improvement consists then in combining several blocks optimally, i.e., in a least square sense [28]. This can be seen as projecting each input block on a specific subspace. However the basis vectors of this subspace - in this case, some blocks of the codebook - are not orthogonal which increases the computational cost. As a result, an alternative approach using Principal Component Analysis (PCA) [29] is described below.

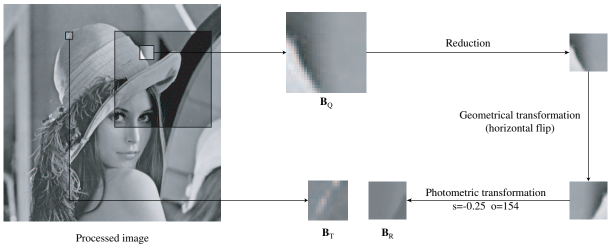


Fig. 8. Fractal-based image coding: for each image block, the coder finds in the search window the most similar block modulo a geometrical and photometric transformation.

A. Block Projection on a PCA Subspace

For each block $\mathbf{B}_T$ of the input image, a search window is defined and partitioned to define a codebook $\mathcal{Q}$. Photometric compensation is then performed to make the blocks $\mathbf{B}_{Q_i}$ of the codebook more similar to the target block $\mathbf{B}_T$. To this end, each block $\mathbf{B}_{Q_i}$ 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:

$$s = \frac{(\mathbf{B}_T - \mathrm{m}_T\mathbf{1}) \cdot (\mathbf{B}_{Q_i} - \mathrm{m}_{Q_i}\mathbf{1})}{|\mathbf{B}_{Q_i} - \mathrm{m}_{Q_i}\mathbf{1}|^2} \tag{8}$$

$$o = \mathrm{m}_T - s.\mathrm{m}_{Q_i} \tag{9}$$

where m_T (resp. m_{Q_i}) is the mean value of block $\mathbf{B}_T$ (resp. $\mathbf{B}_{Q_i}$), $\cdot$ is the linear correlation defined as:

$$\mathbf{B} \cdot \mathbf{B}' = \frac{1}{S_T} \sum_{i=1}^{S_T} \mathbf{B}(i)\mathbf{B}'(i) \tag{10}$$

and $|\mathbf{B}|$ is the norm defined as $\sqrt{\mathbf{B} \cdot \mathbf{B}}$. At this point, the codebook $\mathcal{Q}$ contains a collection of blocks similar to the input block $\mathbf{B}_T$. However, they are usually not similar enough to allow replacement without introducing strong visual distortions. Several blocks need to be combined to obtain valid candidate blocks for replacement. PCA is consequently performed considering all the blocks $\mathbf{B}_{Q_i}$ to obtain an orthogonal basis describing the variations of the blocks in the codebook. This gives a centroid $\mathbf{C}$ defined as follows:

$$\mathbf{C} = \frac{1}{|\mathcal{Q}|} \sum_{\mathbf{B}_{Q_i} \in \mathcal{Q}} \mathbf{B}_{Q_i} \tag{11}$$

and a set of eigenblocks $\mathbf{E}_i$ associated with their eigenvalues e_i which indicate how much the codebook $\mathcal{Q}$ varies in this direction. Next, those eigenblocks are sorted by descending eigenvalues, i.e., the direction $\mathbf{E}_1$ contains more information than any other one in the basis. A candidate block for replacement $\mathbf{B}_R$ is then computed using the N first eigenblocks so that the distortion with the target block $\mathbf{B}_T$ is minimized. In other terms, the block $\mathbf{B}_T - \mathbf{C}$ is projected onto the subspace spanned by the N first eigenblocks and the replacement block can be written:

$$\mathbf{B}_R = \mathbf{C} + \sum_{i=1}^{N} \frac{(\mathbf{B}_T - \mathbf{C}) \cdot \mathbf{E}_i}{|\mathbf{E}_i|^2} \mathbf{E}_i \tag{12}$$

Of course, the distortion $\Delta = \mathrm{MSE}(\mathbf{B}_T, \mathbf{B}_R)$ gracefully decreases as the number N of combined eigenblocks increases. An adaptive framework can be introduced to dynamically adjust the value N so that $\sqrt{\Delta}$ is as near as possible to a target value $\sqrt{\tau_{\text{target}}}$. In other terms, N is iteratively increased and the value which gives the candidate block whose distortion minimizes $|\sqrt{\Delta} - \sqrt{\tau_{\text{target}}}|$ is retained. This parameter can be used as an attacking strength: the higher τ_{target}, the stronger the attack. It should be noted that the underlying assumption is that most of the watermark energy is 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. Furthermore, for visibility reasons, overlapping blocks are considered during the attack to prevent blocking artifacts.

B. Experiments

A database of 500 images of size 512×512 has been considered for experiments. It contains snapshots, synthetic images, drawings and cartoons. Each image has been watermarked using a simple additive spread-spectrum scheme with an embedding strength $\alpha = 3$ and then submitted to three alternative attacks (Gaussian blurring, JPEG compression and block projection on a PCA-defined subspace). For each attack, several predefined values have been used for the parameter settings (filter width, JPEG quality factor, target threshold τ_{target}). At this point, for each image in the database, a distortion vs. correlation curve can be drawn for each attack. All the curves associated with a given attack are then *averaged* to depict the statistical behavior of the image database for this particular attack. The three resulting curves have been gathered in Figure 9. It is obvious that the block replacement strategy first succeeds in removing the embedded watermark, second outperforms both surveyed image processing operations. In other terms, from an attacker perspective, it allows to improve the trade-off distortion vs. correlation. For instance, the correlation score drops below the detection threshold $\tau_{\text{detect}} = 1.5$ around 40 dB with the block replacement strategy while it is necessary to introduce a distortion around 36 dB to obtain the same result with the other attacks. Furthermore, assuming that the parameters of the attacks are set so that the introduced distortion is similar to the one due to the embedding process (38 dB), the block replacement strategy traps the detector while watermarks submitted to either Gaussian blurring or JPEG compression can still be detected.

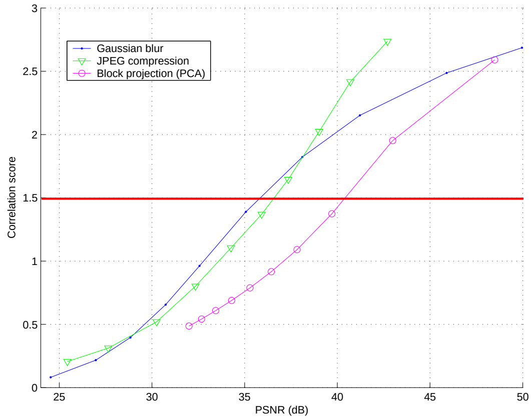


Fig. 9. Correlation score vs. distortion curves for Gaussian blurring, JPEG compression and block projection on a PCA-defined subspace.

VI. Conclusion

In some applications, digital watermarks are embedded to reduce the potential illegal usage of protected data or to identify customers who have broken their license agreement. In such situations, malicious users may try to remove the hidden data which can be used against them. Since attackers are likely to follow best-effort attacks, resistance to strong hostile attacks has to be considered, not only survival after *usual* signal processing operations such as filtering or lossy compression. In security fields, improvements usually come up from the competition between technology providers and attackers. Consequently, in this paper, a novel attacking strategy has been proposed in an effort to anticipate the possible behavior of malicious users. The basic idea consists of exploiting multimedia self-similarity to replace each signal block with another, perceptually similar one. More precisely, several blocks carrying different watermarks are combined to obtain a candidate block for replacement which is expected not to carry the same watermark as the input signal block. Thus, this attack can be considered as a form of intra-signal collusion: several watermarked signal-blocks are combined to obtain unwatermarked blocks. Nevertheless, even if this attacking strategy is somewhat generic, it needs to be adjusted to cope with different nature of multimedia content. In this paper, practical implementations of this attack have been presented for video shots, audio tracks, and still images. In each case, experimental results have demonstrated the efficiency of the proposed block replacement attacking strategy.

The attacker basically exploits the fact that the embedding algorithm does not take into account the self-similarities of the signal. It is possible to build some sets of similar blocks, which on the other hand are not assumed to carry similar watermark samples. This is a weak link of most watermarking schemes today and an informed attacker can exploit it to defeat the protection system. Of course, this brings up an interesting question: which countermeasures can be introduced by technology providers to disable, or at least decrease the impact, of such an attack? Intuitively, if *similar signal blocks carry similar watermarks*, the presented block replacement strategy is likely to be ineffective. In other terms, the introduced watermark has to be coherent with the self-similarities of the host signal. This can be seen as an intermediary specification between the security requirements for steganography - the embedded watermark should be statistically invisible [30] so that an attacker cannot even detect the *presence* of the hidden watermark - and the absence of any one for non-secure applications such as data hiding or broadcast monitoring. Unfortunately this intuitive statement does not point to straightforward constructive ideas on how to obtain such *coherent watermarks* in practice. An early study in video has demonstrated that security can be improved by making the watermark coherent with camera motion [16], so that a given 3D physical point carries the same watermark sample along a video scene. Then, temporal frame averaging after registration becomes useless. However, a generic approach has still to be designed to solve the problem in the general case. In particular, this new specification raises many interesting questions. Is it possible to obtain such a coherent watermark for an arbitrary host signal? If not, which strategy should be enforced to minimize the impact of block replacement attacks? How many bits can be reliably embedded? Does the achievable capacity depend on the host signal or not?

REFERENCES

[1] DVD Copy Control Association, “http://www.dvdcca.org.”

[2] Secure Digital Music Initiative, “http://www.sdmi.org.”

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

[4] Certimark, “http://www.certimark.org.”

[5] Checkmark, “http://watermarking.unige.ch/checkmark.”

[6] Stirmark, “http://www.petitcolas.net/fabien/watermarking/stirmark.”

[7] M. Kutter and F. Petitcolas, “A fair benchmark for image watermarking systems,” in *Security and Watermarking of Multimedia Contents*, ser. Proceedings of SPIE, vol. 3657, January 1999, pp. 226–239.

[8] G. Langelaar, R. Lagendijk, and J. Biemond, “Removing spatial spread spectrum watermarks by nonlinear filtering,” in *Proceedings of the European Signal Processing Conference*, vol. IV, September 1998, pp. 2281–2284.

[9] J. Su and B. Girod, “Power-spectrum condition for energy-efficient watermarking,” *IEEE Transactions on Multimedia*, vol. 4, no. 4, pp. 551–560, December 2002.

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

[11] F. Petitcolas, R. Anderson, and M. Kuhn, “Attacks on copyright marking systems,” in *Proceedings of the Second International Workshop on Information Hiding*, ser. Lecture Notes in Computer Science, vol. 1525, April 1998, pp. 219–239.

[12] 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.

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

[14] M. Holliman and N. Memon, “Counterfeiting attack on oblivious blockwise independent invisible watermarking schemes,” *IEEE Transactions on Image Processing*, vol. 9, no. 3, pp. 432–441, March 2000.

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

[16] ——, “Secure background watermarking based on video mosaicing,” in *Security, Steganography and Watermarking of Multimedia Contents VI*, ser. Proceedings of SPIE, vol. 5306, January 2004, pp. 304–314.

[17] 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.

[18] 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*, vol. 6, no. 4, pp. 303–316, December 1995.

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

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

[21] D. Kirovski and F. Petitcolas, “Replacement attack on arbitrary watermarking systems,” in *Proceedings of the ACM Digital Rights Management Workshop*, ser. Lecture Notes in Computer Science, vol. 2696, July 2003, pp. 177–189.

[22] D. Kirovski and H. Malvar, “Robust covert communication over a public audio channel using spread spectrum,” in *Proceedings of the Fourth International Workshop on Information Hiding*, ser. Lecture Notes in Computer Science, vol. 2137, April 2001, pp. 354–368.

[23] K. Brandenburg, “Coding of high quality digital audio,” in *Applications of Digital Signal Processing to Audio and Acoustics*, M. Kahrs and K. Brandenburg, Eds. Boston, MA: Kluwer, 1998.

[24] H. Malvar, “A modulated complex lapped transform and its application to audio processing,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1999.

[25] A. Oppenheim and R. Schafer, *Discrete-time Signal Processing*. Prentice-Hall, 1989.

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

[27] C. Rey, G. Doërr, J.-L. Dugelay, and G. Csurka, “Toward generic image dewatermarking?” in *Proceedings of the IEEE International Conference on Image Processing*, vol. III, September 2002, pp. 633–636.

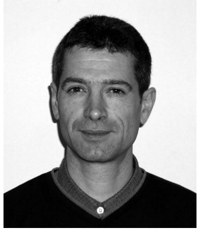
[28] G. Doërr, J.-L. Dugelay, and L. Grangé, “Exploiting self-similarities to defeat digital watermarking systems - a case study on still images,” in *Proceedings of the ACM Multimedia and Security Workshop*, September 2004, pp. 133–142.

[29] I. Jolliffe, *Principal Component Analysis*. Springer-Verlag, 1986.

[30] P. Sallee, “Model-based steganography,” in *Proceedings of the Second International Workshop on Digital Watermarking*, ser. Lecture Notes in Computer Science, vol. 2939, October 2003, pp. 154–167.

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

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



Jean-Luc Dugelay (Ph.D. 92, IEEE M’94-SM’02) joined the Eurécom Institute (Sophia Antipolis, France) in 1992. He is currently a Professor in the Department of Multimedia Communications and is in charge of the Image and Video Group for Multimedia Communications and Applications. His research interests include security imaging (watermarking and biometrics), image/video coding, facial image analysis, face cloning and talking heads. He has published over 80 technical papers and holds three international patents. He has in particular contributed to the first book on digital watermarking [31] and has given several tutorials on this topic co-authored with F. Petitcolas (Microsoft Research, Cambridge, England). In addition to national French projects, his group is involved in the European Network of Excellence *E-Crypt*. He is also serving as a Consultant in digital watermarking for France Télécom R&D and STMicroelectronics. He is an Associate Editor for the IEEE Transactions on Multimedia, the IEEE Transactions on Image Processing, the EURASIP Journal on Applied Signal Processing and the Kluwer Multimedia Tools and Applications.

Darko Kirovski (Ph.D. 01) has been a researcher at Microsoft Research since April 2000. His research interests include: system security, multimedia processing, and embedded system design. He has received the 1999 Microsoft Graduate Research Fellowship, the 2000 ACM/IEEE Design Automation Conference Graduate Scholarship, the 2001 ACM Outstanding Ph.D. Dissertation Award in Electronic Design Automation, and the Best Paper Award at the ACM Multimedia 2002.