Collusion Issue in Video Watermarking

Gwenaël Doërr and Jean-Luc Dugelay

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

ABSTRACT

Digital watermarking has first been introduced as a possible way to ensure intellectual property (IP) protection. However, fifteen years after its infancy, it is still viewed as a young technology and digital watermarking is far from being introduced in Digital Right Management (DRM) frameworks. A possible explanation is that the research community has so far mainly focused on the robustness of the embedded watermark and has almost ignored security aspects. For IP protection applications such as fingerprinting and copyright protection, the watermark should provide means to ensure some kind of trust in a non secure environment. To this end, security against attacks from malicious users has to be considered. This paper will focus on collusion attacks to evaluate security in the context of video watermarking. In particular, security pitfalls will be exhibited when frame-by-frame embedding strategies are enforced for video watermarking. Two alternative strategies will be surveyed: either eavesdropping the watermarking channel to identify some redundant hidden structure, or jamming the watermarking channel to wash out the embedded watermark signal. Finally, the need for a new brand of watermarking schemes will be highlighted if the watermark is to be released in a hostile environment, which is typically the case for IP protection applications.

Keywords: Video watermarking, security, collusion

1. INTRODUCTION

Digital watermarking was initially introduced in the early 90's as a complementary protection technology.[1] Encryption alone is indeed not enough. 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. Digital watermarking basically consists in hiding some information into digital content in an imperceptible manner. 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. Unfortunately, the fact that digital watermarking was introduced when the *Internet bubble* was expanding exponentially really had a negative impact on the development of this technology. Industry was ready to invest considerable amounts of money in any emerging concept provided that it exhibited promising results. On the other hand, researchers were may be over-enthusiastic by the large range of possible applications and technical challenges.

In particular, 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, it appeared somewhat quickly that no method was robust enough for copyright protection. This has been demonstrated in practice by the partial failure of initiatives to launch watermarking-based copy control mechanisms.[2, 3] As a result, many industries cut down their investments devoted to watermarking for security-sensitive applications and, today, most of the research is focused on applications such as media monitoring, metadata binding, steganography, content tracking... The failure of the few attempts to introduce digital watermarking in security-sensitive applications

Send correspondence to Professor Jean-Luc Dugelay: E-mail: dugelay@eurecom.fr, Telephone: +33 4.93.00.26.41, Switchboard: +33 4.93.00.26.26, Fax: +33 4.93.00.26.27

has been mainly due to the claim that embedded watermarks would survive in a highly hostile environment. However, quite surprisingly, very few works have addressed this issue. If the survival of the watermark against common signal processing primitives - filtering, lossy compression, desynchronization - has been carefully surveyed, almost no work has considered that an attacker may exploit some knowledge on the watermarking systems to defeat it. Nevertheless, in applications such as copy control or fingerprinting, digital watermarking is usually seen as a disturbing technology. If content owners are glad to have new means to protect their high valued multimedia items and to identify malicious customers, these latter ones on the other hand do not really appreciate that some hidden signal prevent them from freely copying digital material or that an invisible watermark could possibly be used in court to identify them as a source of leakage. As a result, this protecting technology is likely to be submitted to strong hostile attacks when it is released to the public.

Researchers should consequently try to anticipate hostile behaviors from malicious customers to be able to lay claims to security concerning their watermarking systems. Since the watermarking community has still not really agreed upon what is exactly meant by security, a discussion is conducted in Section 2 to give some hints regarding this issue. In particular, the difference between security and robustness is investigated. Then, collusion attacks are introduced in Section 3 as a possible way to evaluate security. Collusion basically consists in collecting several watermarked documents and combining them to obtain unwatermarked content. Next this approach is further studied in the case of video watermarking. In Section 4, uncorrelated watermarked video frames are gathered to check whether or not these independent observations leak some knowledge on a redundant watermarking structure. In case such knowledge can be learned, it is then usually relatively easy to exploit it to remove the watermark. Alternatively, in Section 5, similar watermarked contents are inspected and combined so that the embedded watermark is washed out. Finally, conclusions are drawn in Section 6 and tracks for future work are indicated.

2. WHAT IS SECURITY?

Digital watermarking has always been regarded as a security related technology. Nevertheless, it is not really clear what the term *secure* refers to. At the very beginning, this was kind of connected with the fact that watermarking embedding and detection processes are made dependent of a secret key. A direct analogy has then been drawn with cryptography and Kerckhoffs' principles to ensure security have been considered.[4] In particular, even if the system under study is publicly known, it should not be broken down as long as the secret key is not disclosed. However, whereas *breaking down the system* means obtaining the plain text in cryptography, it might means different things in digital watermarking. For instance, unauthorized users should not be able to remove, detect, estimate, write or modify embedded watermarks.[5] But if extensive benchmarking is now performed to assert whether or not a watermarking system can cope with standard signal processing primitives, almost no work has been done to evaluate security. Thus, the remainder of this section will try to give some elements to define somewhat clearly what security is. First of all, the relationship between security and the need for trust in a hostile environment is exhibited in Subsection 2.1. Then, security is opposed to robustness in Subsection 2.2 to draw a line, even fuzzy, between both concepts. Finally, Subsection 2.3 reminds that absolute security is not required in real life. In fact, even limited security can be valuable when the risk is properly managed.

2.1. Trust in a Hostile Environnement

In many watermarking applications, there is a need to trust the information which is conveyed by the watermarking channel. When fingerprinting watermarks are exploited to trace back malicious customers who have broken their license agreement, content owners basically want to rely on the information extracted from the embedded watermark. Thus, in such a scenario, it should not be possible to produce multimedia contents carrying fake valid watermarks to prevent innocent consumers from being framed. Additionally, since fingerprinting watermarks can be used to identify the source of leakage in a content distribution system, they are likely to be attacked and they should consequently survive as many removal attacks as possible. On the other hand, it is not critical if the attacker succeeds in detecting/reading the watermark. However this last point is no longer valid when digital watermarking is used for steganography. In this case, the watermarking channel is basically exploited to establish a covert communication channel between two parties whose existence remains unknown for other people. Therefore, the presence of a hidden watermark should not be even detected if the secret key is not known. Alternatively, in an authentication perspective, unauthorized watermark removal is not really important: digital content will be seen as non valid and discarded. In a completely different strategy, digital watermarks can be used to insert useful information into the cover data e.g. annotation watermarks, error recovery watermarks, metadata binding watermarks... In such cases, altering the watermark is likely to remove the associated additional information or service. In other terms, consumers have no longer interest to eliminate embedded watermarks and hostile behaviors disappear. The environment is collaborative instead of hostile and security requirements are now superfluous.

In summary, the environment in which the watermarking technology is going to be released has to be carefully studied. Depending on the targeted applications, customers will not judge digital watermarking the same way and will consequently adopt different attitudes. Generally speaking, the more the embedded watermarks *disturb* the customers, the more hostile will be their behaviors and the higher will the security specifications need to be raised. Hence, security issues basically arise when two conditions are met. On one side, content providers value the information conveyed by the watermarking channel and expect the embedded watermark to survive to provide some service e.g. copy control to prevent copying multimedia content without paying royalties, traitor tracing to identify the source of leakage in multimedia content distribution system... On the other hand, customers see the watermarking signal as a disturbing technology and deploy highly hostile strategies to defeat the protection system. In other words, the notion of security is inherently tied with the need for trust in a hostile environment. This is a key aspect to consider when a watermarking system is under study. In particular, IP protection related applications should identify which operations are critical or not in their framework. Nevertheless, it should be reminded that a large range of applications using digital watermarking do not have security specifications at all e.g. when embedded watermarks are used to convey some useful information.

2.2. Security versus Robustness

This subsection aims at drawing a line, even fuzzy, between two very important concepts in digital watermarking: security and robustness. These notions have indeed been mixed up for a long time, even in the watermarking community itself. The first noticeable difference is that security implicitly assumes a hostile environment, as reminded in the previous subsection, which is not the case for robustness. Robustness addresses the impact of regular signal processing primitives on watermarked multimedia content. Typically, a usual customer who is compressing multimedia content for storage/transmission convenience is not regarded as a security threat even if lossy compression is likely to increase the probability of missing an embedded watermark or the probability of retrieving a message with bit errors. The main point is that the customer does not intend to remove the watermark. In summary, one can say that robustness is only concerned by regular customers while security cares more specifically about hackers whose well-thought-out goal is to defeat the system. A second distinction is that security attacks usually aims at gaining some knowledge on the protection system. There is a clear gap between a customer who is blindly JPEG compressing an image and a hostile attacker who is collecting a whole database of watermarked documents to check whether some information leaks about the watermarking process or not. This knowledge can then be exploited to detect, remove or edit the embedded watermarks. A direct consequence is that robustness attacks are usually more generic/universal than security ones. Finally, a last dissimilitude between security and robustness is that the later one is only concerned by watermark removal whereas the first one also cares about unauthorized watermark detection, estimation, modification and writing. In summary, robustness deals with common consumers which perform regular signal processing operations which are likely to degrade the performances of watermark detection even if it is not their original goal. On the other hand, security handles hostile hackers who attack the protection system on purpose to gain first some knowledge about it and then to design dedicated attacks which are not limited to watermark removal.

Keeping these observations in mind, usual attacks which are relevant in digital watermarking have been separated as depicted in Figure 1 depending on whether they are a matter of security or robustness. This separation is basically an extension of previous classifications proposed in the watermarking literature.[6, 7] On the robustness side, attacks against digital watermarks have been split into synchronous and asynchronous attacks. Synchronous attacks refer to common signal processing primitives such as filtering, lossy compression, quantization, denoising which are likely to directly affect the watermark signal and thus interfere with the watermark detector. On the other hand, asynchronous attacks include all the operations which perturb the

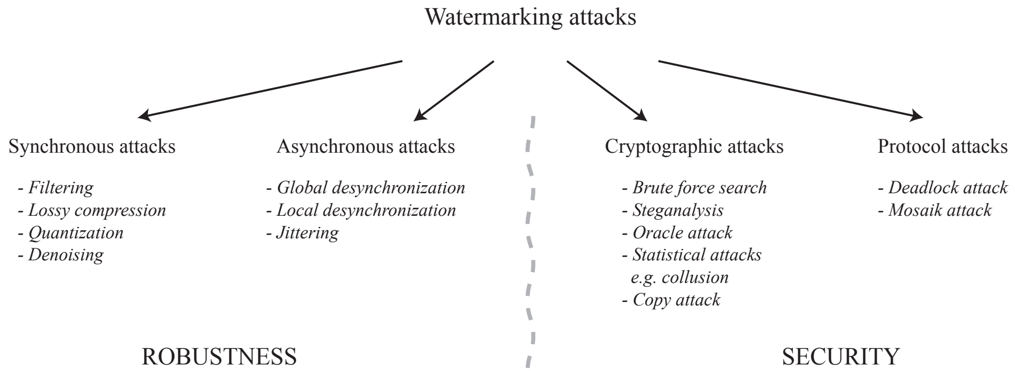

Figure 1. Watermarking attacks breakdown depending on whether they address robustness or security issues.

signal sample positions. As a result, the synchronization convention shared by the embedder and the decoder becomes obsolete. Hence, such attacks do not explicitly remove the watermark signal but still, they are considered as watermark removal attacks since the detector is no longer able to retrieve the watermark.[5, 8] A very well known example is the Random Bending Attack (RBA)[6] which is now a reference attack to evaluate robustness. It basically simulates the effects of D-A/A-D conversion: geometrical distortions are introduced to take into account lens parameters and imperfect sensors alignment, some noise is added to mimic the response of non ideal sensors and a mild JPEG compression is performed for storage convenience.

On the security side, watermarking attacks are divided into protocol and cryptographic attacks. The first set of attacks exploits some general knowledge on the watermarking framework. For instance, in a copyright protection perspective, if a multimedia item is found to carry more than a single watermark, many schemes do not provide an intrinsic way of detecting which of the watermarks was embedded first.[9] It is a deadlock and nobody can claim the ownership of the document. Alternatively, automated search robots can be used to browse the Internet and check if web sites host illegally copyrighted material. A simple way to circumvent such web crawlers is to split the images into many small pieces and to embed them in a suitable sequence in a web page.[10] The juxtaposed images appear stuck during rendering, that is to say as the original image, but the watermark detector is confused. On the other hand, cryptographic attacks aim at gaining some knowledge about the watermark signal itself i.e. the secret key or the utilized pseudo-random sequence. Brute force search basically tries all the possible keys until the correct one is found. Steganalysis objective is to isolate some characteristics of watermarking methods to permit non authorized watermark detection.[11] In another fashion, the Oracle attack uses publicly available detectors to iteratively modify watermarked content until the detector fails to retrieve the embedded watermark.[12] On their side, statistical attacks, also referred to as collusion attacks, collect several watermarked documents and combine them to obtain non watermarked documents. This attacking strategy will be further examined in the next sections of the paper. Finally, to the best knowledge of the authors, the only example of unauthorized watermark writing is the copy attack[13]: the watermark is estimated from a watermarked document and successfully inserted back into a non-protected one.

2.3. Security in the Real World

Nowadays, it is commonly admitted that a perfectly secure system does not exist. If a motivated hacker has no limit of time, computing resources and money, he/she will succeed in defeating the protection system, for instance with a brute force key search approach. This is also true for digital watermarking. Does it mean that security is useless? Not at all! Customers value even limited forms of content protection. Let us for instance examine the case of copy protection for Digital Versatile Disk (DVD) distribution. When content owners release a new movie, they know that most of the sales are going to be done within the first months. As a result, they basically want the copy protection to last few months and do not really care if a pirate hacks the protection after one year and release it on the Internet. If the protection technology longs last enough, customers who are eager to consume new released products will not wait until a pirated copy appears on Peer-to-Peer (P2P) networks. A second important point is that just because a technology can be circumvented does not necessarily implies that customers will effectively do it.[14] For example, if the case study of DVD is continued, the proposed copy/playback protection basically divides devices into compliant DVD players which cannot read illegal DVDs and noncompliant DVD players which cannot read legal DVDs.[15] The expense of owning two DVD players can then be exploited to help *keep honest people honest.* In summary, in real life, all that matters is that the cost of breaking the system (complexity, time, money...) should be higher than the cost of doing things legally.

Coming back to down-to-earth considerations, money has also to be taken into account. Who will pay to introduce a secure protection system? This is a big issue and also the point where the different concerned parties usually disagree. There are typically three actors: content owners (cinema studios, music majors...), consumer electronics manufacturers and consumers associations. Content owners want to protect their high valuable multimedia items once they are released to the public but they are most of the time not ready to bear all the cost of the protection system. From the manufacturer point of view, more security means more hardware or software, more expensive devices and consequently less sales and lower profits. And of course, consumers are not really enthusiastic about the idea of paying for some security mechanism which is going to restrict the possible usages of multimedia data. Such conflicts of interest can come to a dead end. A typical example is the introduction of digital watermarks inside DVDs which has been almost abandoned. Thus, the efficiency of the protection technology has to be balanced with the economic interests at stake. Once again insecure technologies may be worth being deployed if the risk is properly managed.[16] Although credit card networks are based on insecure magnetic stripe technology, risk management tools have been able to maintain fraud rates below 0.1% of transaction volume.

Nevertheless, even with an efficient risk management strategy, it is useful to investigate how to obtain secure watermarks. The main issue here is that watermarked content cannot be updated once released as the antivirus softwares are when a new threat appears. Watermark embedders and detectors can of course be improved but the already released items will not benefit from these enhancements. In fact, as soon as the watermark is removed from a multimedia item, this one is likely to be broadcasted on P2P networks. Today, a single leak on the Internet and everything is done. As a result it is still relevant to anticipate hostile behaviors to iteratively propose more secure algorithms i.e. a longer time for removing the watermark once the protected item is released to the public. In another fashion, research is also performed to design systems, such as conditional access for multimedia players,[17] where a defeated entity does not compromise the security of the whole system.

3. COLLUSION ATTACKS

Collusion is a well-known attacking strategy which basically refers to a set of malicious customers who gather their individual knowledge of the protection system, whatever it is, to obtain unprotected multimedia content. It has been first mentioned in cryptography when protocols have been established to part a secret between different individuals. Typical examples include secret sharing, also referred to as threshold cryptography, and conference keying.[18] The idea of secret sharing is to start with a secret and then to divide it into pieces called *shares* which are distributed amongst users such that the pooled shares of specific users allow reconstruction of the original secret. These schemes can be exploited to enable shared control for critical actions. Vault deposit accounts are a good illustration of such a procedure. Both the customer key and the bank manager key are required to grant access to the account. If any part of the secret (key) is missing, the door of the vault remains closed. This idea can be extended to more than two people. Access to a top secret laboratory can for instance be controlled by access badges: admittance necessitates a security guard badge and a researcher badge. Since there are many researchers and guards in the lab, this results in two groups of badges and one badge from each group is required to enter the lab. From a more general perspective, secret sharing split knowledge between individuals so that at least u users have to merge their shares to retrieve the secret knowledge. In such a perspective, colluders are c users who try to build fake valid shares or even to reconstruct the secret despite the fact that $c < u$. On the other side, conference keying is slightly different. Whereas secret sharing can be seen as a key pre-distribution technique wherein the recovered secret is static and usually the same for all groups, conference keying protocols allow to have *session keys* which are different for different groups and which dynamically adapt to the individuals in the group. These protocols are particularly interesting for applications which need secure group communications such as telephone/video conferences, interactive meetings and Video on Demand (VoD).[19] Most concerns come from the need to manage members joining or leaving groups, which has an impact on session keys. In such scenarios, the goal of the colluders is to create some new keys to join the sessions without paying the fee. In other terms, collusion has already been studied in cryptography. The riposte which has been introduced to circumvent such behaviors is basically a dissuasive weapon. Distributed keys are build in such a way that, if some colluders combine several of them to produce an illegal key, this one contains some evidence of the pirate identities which can be used to take legal measures.[20, 21] Once there is a threat of being caught, there are far more less candidates for collusion.

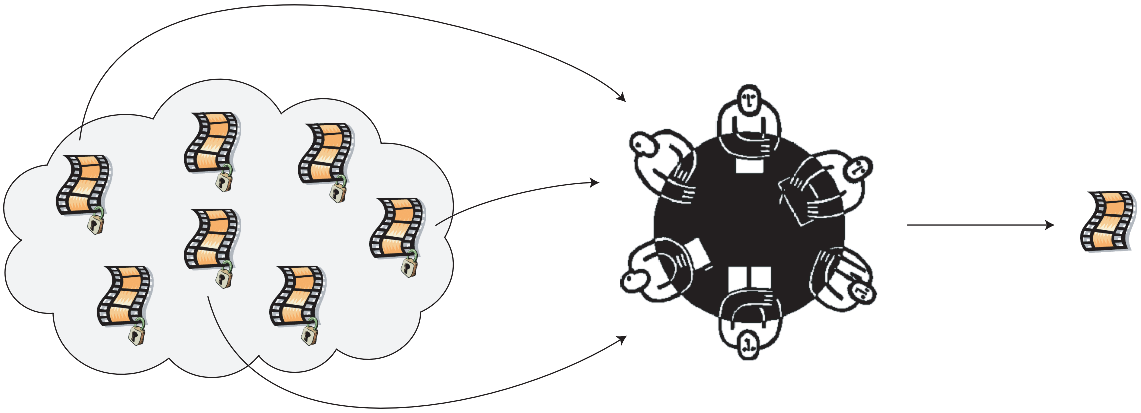

Figure 2. Collusion in watermarking: Colluders collect several watermarked documents and combine them to produce digital content without underlying watermarks.

In digital watermarking, collusion attacks were first mentioned in the context of fingerprinting.[22] In such applications, content owners want to distribute high valued items to a large audience. However, they are concerned about their copyright and want to be able to trace illegal content back to the source of leakage. To this end, instead of selling the same item to all the customers, they assign slightly different copies to each customer. As a result, each customer owns a *unique* copy carrying its own imperceptible and randomly located tracers. Thus, if a customer decides to make his/her copy freely available on the Internet, the content owner is able to find the identity of the traitor using these secret markers. In this case, colluders will typically compare several marked documents to identify where the secret markers are located and then remove them. Now, in terms of digital watermarking, collusion consists in collecting several watermarked documents and in applying a process which succeeds in producing unwatermarked content as depicted in Figure 2. Traditionally, two alternative approaches can be enforced: combining watermarked documents can either aim at estimating directly the original unwatermarked content or in estimating some properties of the watermark signal which can be exploited to remove the embedded watermark in a second step. Solutions have already been proposed in the literature. Secure codes can for instance be used to prevent the watermark to be removed when multiple customers average their protected items.[23]

Nevertheless, when video content is considered, the situation is significantly more challenging. Each frame of the video can indeed be seen as an individual watermarked content. This approach is all the more pertinent since many video watermarking schemes enforce a frame-by-frame embedding strategy.[24] An attacker can consequently collect multiple video frames and combine them to produce unwatermarked video frames. In this perspective, early studies have exhibited two main collusion strategies.[25, 26] When uncorrelated host video frames are studied, the goal is to isolate some hidden structure of the watermark signal. On the other hand, when similar video frames are collected, the objective is to fuse these frames so that the embedded watermark is no longer detectable. Both approaches will be further developed in the remaining two sections of this paper. To this end, a simple, frame-by-frame watermarking framework will be considered 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 and $\check{\mathbf{F}}_t$ its watermarked version, α the embedding strength and $\mathbf{W}_t$ the watermark embedded at instant t which is normally distributed with zero mean and unit variance. Different temporal strategies can be enforced during watermark embedding which are associated with different collusion attacks.

4. EAVESDROPPING THE WATERMARKING CHANNEL

Once several watermarked video frames have been collected, a typical collusion strategy consists in looking for a *unusual* statistically redundant structure. In other terms, having some a priori on the host signal, the goal is to find some suspicious discrepancies which are likely to reveal the presence of a hidden watermark. These statistical leaks are usually the mirroring footprints of the enforced temporal embedding strategy. The next subsections will in particular highlight the danger of always embedding the same reference watermark pattern (Subsection 4.1), always embedding the same set of watermarks (Section 4.2) or even the risk of using the same watermarking subspace (Section 4.3). From the attacker point of view, the most favorable position would be to have a direct access to the watermarking channel. However, in practice, it is usually not possible since the attacker does not have access to the original content and cannot consequently compute the optimal watermark estimate $\mathcal{E}_{\text{o}}(\check{\mathbf{F}}_t) = \check{\mathbf{F}}_t - \mathbf{F}_t$ at each instant t. Nevertheless, the watermarking signal can basically be seen as noise and is thus located in high frequencies. As a result, for each video frame, a rough estimation of the embedded watermark can be obtained in a blind manner by using for instance denoising techniques, or even more simply by computing the difference between the watermarked frame and its lowpass filtered version as follows[27]:

$$\mathcal{E}(\check{\mathbf{F}}_t) = \check{\mathbf{F}}_t - \mathcal{L}(\check{\mathbf{F}}_t) = \tilde{\mathbf{W}}_t \tag{2}$$

where $\tilde{\mathbf{W}}_t$ is the estimate of the embedded watermark at instant t and $\mathcal{L}(.)$ a lowpass filter e.g. a simple 5×5 spatial averaging filter. Now looking at this collection of noisy observations $\{\tilde{\mathbf{W}}_t\}$, the attacker has to identify some secret redundant structure which can be exploited to remove the embedded watermark signal.

4.1. Estimate a Redundant Watermark Pattern

Early algorithms which have been proposed to watermark digital video content basically aim at considering several successive video frames during detection to obtain more reliable detection statistics. Furthermore, to get free from temporal synchronization constraints, a straightforward solution consists in redundantly embedding the same reference watermark pattern[28]:

$$\textit{Always the same}\text{ strategy:} \qquad \forall t \quad \mathbf{W}_t = \mathbf{W}_0 \tag{3}$$

where $\mathbf{W}_0$ is a normally distributed watermark with zero mean and unit variance. The major asset of such an approach is that it does not require the detection procedure to be run for each frame. Indeed, if the detection is linear, then accumulating over time several detection scores obtained at different instant is equivalent to performing a single detection using the accumulation of frames over time. The resulting gain in computational cost enables then real-time detection. However, this advantage is counterbalanced by a critical security pitfall. Redundantly embedding the same watermark makes the reference pattern $\mathbf{W}_0$ statistically visible. If each individual estimate $\tilde{\mathbf{W}}_t = \mathcal{E}(\check{\mathbf{F}}_t)$ is not accurate enough to threaten the performances of the detector, combining over time several noisy estimates is likely to significantly enhance the estimation of the reference pattern $\mathbf{W}_0$. A simple approach is for instance to compute the average of the individual estimates as follows:

$$\tilde{\mathbf{W}}_0 = \frac{1}{T} \sum_t \tilde{\mathbf{W}}_t \tag{4}$$

where T is the number of video frames collected for collusion.[25, 26] This estimate $\tilde{\mathbf{W}}_0$ can then be remodulated to efficiently remove the embedded watermark signal.[27] The whole workflow of this Watermark Estimation Remodulation (WER) attack is depicted in Figure 3. It should be noted that the more the video frames used for collusion are different, the more each individual estimate refines the final one. As a result, WER attack is more relevant in dynamic scenes or when the key frames of a video sequence are isolated. In the same manner, the more individual estimates $\tilde{\mathbf{W}}_t$ are collected, the finer is the final watermark estimate $\tilde{\mathbf{W}}_0$. These statements will remain valid for the other estimation-based collusion strategies presented in this section.

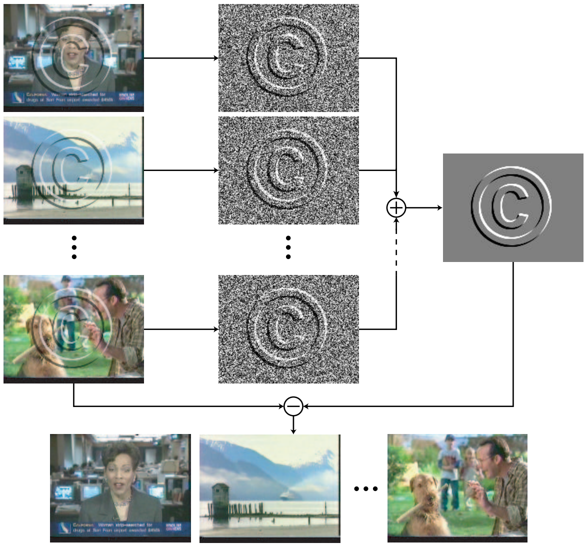

Figure 3. Watermark Estimation Remodulation (WER): Several watermark estimations obtained from different video frames are combined to refine the estimation and enable watermark removal.

4.2. Estimate a Set of Watermark Patterns

An immediate response to the threat of the WER attack is to use more than a single watermark pattern. Thus, for each video frame, the watermark to be embedded is chosen within a pool of N reference patterns $\{\mathbf{W}_i\}$ as written below:

$$\textit{1 amongst N} \text{ strategy:} \qquad \forall t \quad \mathbf{W}_t = \mathbf{W}_{\Phi(t)}, \quad \text{with} \begin{cases} \mathbf{W}_i \cdot \mathbf{W}_j = \delta_i^j, & 1 \leq i, j \leq N \\ \mathrm{P}\big(\Phi(t) = i\big) = 1/N, & 1 \leq i \leq N \end{cases} \tag{5}$$

where $\cdot$ denotes the linear correlation operation and δ the Kronecker delta. It should be noted that this embedding strategy include a broad range of watermarking schemes going from periodical watermark scheduling to completely random watermark scheduling.[29] The reference patterns are orthonormalized to prevent cross-talk on the detector side and are emitted equiprobably to enable statistical invisibility against WER attacks. If multiple watermark estimates $\tilde{\mathbf{W}}_t$ are averaged, the attacker obtains the average of the reference patterns $\mathbf{W}_i$ and watermark removal is not possible. The gain in security due to this novel design basically relies on the assumption that attackers are unable to build sets of frames carrying the same reference watermark pattern. Otherwise, a simple WER attack performed on each subset succeeds in estimating the pool of secret watermarks. And indeed, even if a brute force attack can be designed to isolate such subsets, its computational complexity prevents its use when N grows large.[30] However, each individual watermark estimate $\tilde{\mathbf{W}}_t$ can be regarded as a vector in a high dimensional space which is assumed to approximate one of the reference patterns $\mathbf{W}_i$. In this perspective, vector quantization can be performed to define N clusters $\mathcal{W}_i$ whose centroids $\tilde{\mathbf{W}}_i$ are good estimates of the secret reference watermark patterns. This approach can be implemented for instance with a simple k-means algorithm and a split and merge strategy to avoid random initialization.[31] Once the reference patterns have been estimated, the attacker can easily determine which estimated watermark $\tilde{\mathbf{W}}_i$ is carried by each video frame and remodulate it to remove the watermark signal. It could be noted that the previous WER attack is a special case of this Watermark Estimates Clustering and Remodulation (WECR) approach when $N = 1$

4.3. Estimate the Watermarking Subspace

WER and WECR attacks exploit the same weak point to defeat watermarking schemes. If embedded watermarks $\mathbf{W}_t$ are regarded as vectors in a high dimensional space, both strategies *always the same* and *1 amongst N* introduce some accumulation points in this space which can be easily isolated by an attacker even if he/she has only access to noisy observations of these watermarks. To avoid inserting such security pitfalls, one can consider embedding strength modulation to mix several reference patterns in each video frames[32]:

$$\textit{Mixing of } N \text{ strategy:} \qquad \forall t \quad \mathbf{W}_t = \sum_{i=1}^{N} \frac{\lambda_i(t)}{\sqrt{\sum_{j=1}^{N} \lambda_j(t)^2}} \mathbf{W}_i \tag{6}$$

where the $\lambda_i(t)$ are N time-varying mixing parameters. Since all the embedded watermarks $\mathbf{W}_t$ have unit variance, successive watermarks can be seen as a trajectory Λ over a unit sphere which is defined by the mixing coefficients $\{\lambda_i(t), 1 \leq i \leq N, 1 \leq t \leq T\}$. If this trajectory does not exhibit any accumulation point, then WECR attacks are bound to fail. However, the attacker can still exploit one weakness: the number N of considered patterns for mixing is usually significantly smaller than the dimension D of the whole media space ($N \ll D$). In other terms, embedded watermarks are bound inside a relatively small watermarking subspace $\mathcal{W} = \text{span}(\mathbf{W}_i)$. Looking at many noisy observations $\tilde{\mathbf{W}}_t$, an attacker is then able to estimate this subspace using common space dimension reduction techniques such as Principal Component Analysis (PCA).[33] Next, each frame can be drained from any energy contained in the estimated subspace $\tilde{\mathcal{W}}$ to confuse the detector. Of course, the larger the dimension N of the watermarking subspace, the more observations are needed to finely estimate it. Furthermore, it should be noted that this Watermark Subspace Estimation Draining (WSED) attack can be further exploited to enable unauthorized writing. Once the subspace $\mathcal{W}$ has been estimated, it is indeed straightforward to *record* the trajectory Λ embedded in a watermarked video and to copy it back in another unprotected video.

5. JAMMING THE WATERMARKING CHANNEL

Even if a redundant structure can be somewhat easily found by an attacker, completely independent watermarking is not the solution. As a matter of fact, an attacker can exploit the property that independent watermark samples usually sum to zero to design efficient collusion attacks. In this perspective, the goal of collusion is no longer to identify some hidden structure which enabless 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 5.1) and even single video frames exhibit some self-similarities (Subsection 5.2).

5.1. Combine Successive Video Frames

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.[34] From a frame-by-frame point of view, such a strategy can be seen as a system which always embeds a different watermark. In other terms, the linear correlation between watermarks embedded at two different instants t and t' is likely to be almost null:

$$\textit{Always different} \text{ strategy:} \qquad \forall t \neq t' \quad \mathbf{W}_t \cdot \mathbf{W}_{t'} = 0 \tag{7}$$

However, the drawback of this approach is that temporal filtering usually succeeds in confusing the watermark detector.[26] In static scenes, video frames are effectively 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, to be able to cope with dynamic content such as fast moving objects and/or camera motion, 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).[35] As depicted in Figure 4, 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 as the background 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 encode the background with an advanced forward-backward dependent estimator (B-frame). Alternatively, it can also

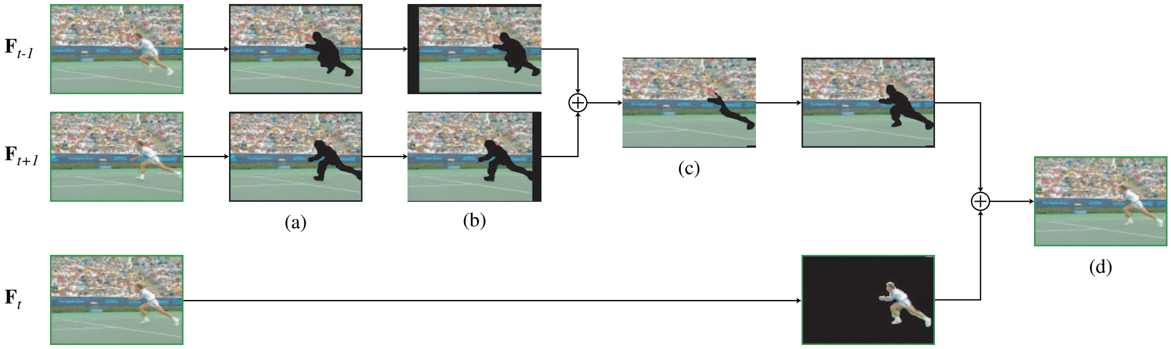

Figure 4. Temporal Frame Averaging after Registration (TFAR): Once the video objects have been removed (a), neighbor frames are registered (b) and combined to estimate the background of the current frame (c). Next, the missing video objects are inserted back (d).

be seen as temporal averaging along the motion axis. Whatever, since most watermarking algorithms do not take the evolution of the structure of the scene into account during embedding, TFAR succeeds in removing the watermark.

5.2. Combine Similar Signal Blocks

Multimedia digital data is highly redundant: successive video frames are similar in a movie clip, popular songs usually contain some repetitive patterns... An attacker can consequently exploit these similarities to replace each part of the signal with a perceptually *similar* one taken from another location in the same signal. Such approaches have already been investigated to obtain efficient compression tools.[36] If there exists obvious similarities in successive video frames as noticed in the previous subsection, there are also some at a lower resolution level: the block level. As a result, in a fractal coding fashion, an attacker can design a Block Replacement Attack (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 5. 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.[37] 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.[38] Since most algorithms published in the literature do not care about the self similarities of the signal to be watermarked, BRA usually succeeds in removing embedded watermarks.

6. CONCLUSION

After being mixed concepts for years, security and robustness begin to be distinguished and security evaluation is even now a raising issue. It is indeed necessary to investigate how the watermarking technology will resist against wily attackers once it is released in a hostile environment, which is usually the case for IP protection related application. In an effort to anticipate hostile behaviors, a collection of collusion attacks has been presented in this paper. Although collusion originally refers to a group of users who combine their individual knowledge, the presented attacks can be performed by a single person. Nevertheless, the *collusion spirit* remains. The baseline of the attacks consists in *combining* several watermarked documents to produce unwatermarked content. The only variation is that different documents are now different parts of the video signal, should they be video frames or signal blocks. Two alternative approaches have been successively studied: estimating some structure of the watermark signal and estimating directly the original non-watermarked signal. Furthermore, these attacks have been shown to defeat many proposed video watermarking schemes relying on a frame-by-frame approach.

These pitfalls should then be considered carefully to design more secure video watermarking schemes. Intuitively, and with simple words, one can say that similar blocks (or frames) should carry similar watermarks while

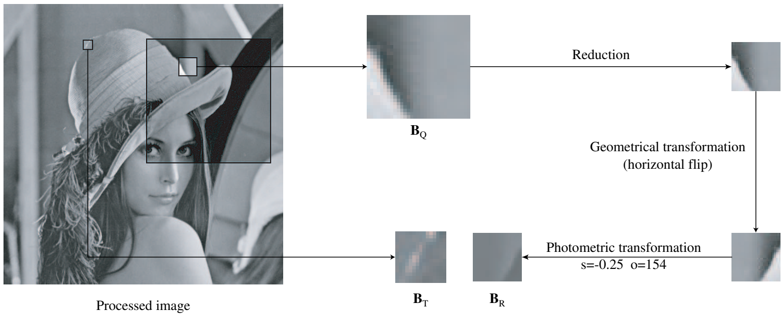

Figure 5. Block Replacement Attack (BRA): Each signal block is replaced by another perceptually similar one, e.g. modulo a geometrical and photometric transformation taken at another location.

uncorrelated blocks should carry uncorrelated watermarks. In other terms, two signal blocks should carry watermarks which are as similar as the blocks themselves i.e. the self-similarities of the host signal should be taken into account during watermark embedding. This can be seen as informed watermarking. Let us consider that digital watermarking basically consists in introducing a small displacement in a random direction in a high dimensional media space. Then, taking the signal self-similarities into account simply forbids some directions. In practice, video mosaicing can be introduced to enforce motion-compensated watermarking.[35] This embedding strategy forces each physical point of the movie set to always carry the same watermark sample whenever its projection appears in the video shot. As a result, such watermarks are not altered by TFAR attacks. Furthermore, recent works have also proposed solutions to generate watermarks which are coherent with the block self-similarities and which are thus immune to BRA.[39] Nevertheless, all these additional security constraints are likely to have an impact on the achievable embedding rate and this trade-off has still to be studied.

ACKNOWLEDGMENTS

The work described in this paper has been supported (in part) by the European Commission through the IST Programme under Contract IST-2002-507932 ECRYPT. The authors want to thank the members of the WAVILA mailing list for their unexpected contributions, at the time this article was written, to precise the term *security* in the context of digital watermarking. Furthermore, they also want to thank Dr. Teddy Furon from IRISA, France, for his valuable comments on collusion in cryptography.

REFERENCES

1. I. Cox, M. Miller, and J. Bloom, *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. A. Kerckhoffs, “La cryptographie militaire,” *Journal des sciences militaires* **IX**, pp. 5–83, January 1883.
5. T. Kalker, “Considerations on watermarking security,” in *Proceedings of the IEEE Workshop on Multimedia Signal Processing*, pp. 201–206, October 2001.
6. F. Petitcolas, R. Anderson, and M. Kuhn, “Attacks on copyright marking systems,” in *Proceedings of the Second International Workshop on Information Hiding*, *Lecture Notes in Computer Science* **1525**, pp. 219–239, April 1998.
7. S. Voloshynovskiy, S. Pereira, V. Iquise, and T. Pun, “Attack modeling: Towards a second generation watermarking benchmark,” *Signal Processing* **81**, pp. 1177–1214, June 2001.
8. Stirmark, “http://www.petitcolas.net/fabien/watermarking/stirmark.”
9. S. Craver, N. Memon, B.-L. Yeo, and M. Yeung, “Resolving rightful ownerships with invisible watermarking techniques: Limitations, attacks, and implications,” *Journal on Selected Areas in Communications* **16**, pp. 573–586, May 1998.
10. 2Mosaic, “http://www.petitcolas.net/fabien/watermarking/2mosaic.”
11. R. Chandramouli, M. Kharrazi, and N. Memon, “Image steganography and steganalysis: Concepts and practice,” in *Proceedings of the Second International Workshop on Digital Watermarking*, *Lecture Notes in Computer Science* **2939**, pp. 35–49, March 2004.

12. J.-P. Linnartz and M. van Dijk, “Analysis of the sensitivity attack against electronic watermarks in images,” in *Proceedings of the Second International Workshop on Information Hiding*, *Lecture Notes in Computer Science* **1525**, pp. 258–272, April 1998.
13. M. Kutter, S. Voloshynovskiy, and A. Herrigel, “Watermark copy attack,” in *Security and Watermarking of Multimedia Contents II*, *Proceedings of SPIE* **3971**, pp. 371–380, January 2000.
14. M. Barni, “What is the future for watermarking? (part I),” *IEEE Signal Processing Magazine* **20**, pp. 55–59, September 2003.
15. J. Bloom, I. Cox, T. Kalker, J.-P. Linnartz, M. Miller, and C. Traw, “Copy protection for DVD video,” *Proceedings of the IEEE* **87**, pp. 1267–1276, July 1999.
16. P. Kocher, J. Jaffe, B. Jun, C. Laren, and N. Lawson, “Self-protecting digital content.” Cryptography Research Inc. White Paper, April 2003.
17. D. Kirovski, H. Malvar, and Y. Yacobi, “Multimedia content screening using a dual watermarking and fingerprinting system,” in *Proceedings of the Tenth ACM International Conference on Multimedia*, pp. 372–381, November 2002.
18. A. Menezes, P. van Oorschot, and S. Vanstone, *Handbook of Applied Cryptography*, CRC Press, 1996.
19. A. Eskicioglu, “Multimedia security in group communications: Recent progress in key management, authentication and watermarking,” *ACM Multimedia Systems, Special Issue on Multimedia Security* **9**, pp. 239–248, September 2003.
20. B. Chor, A. Fiat, and M. Naor, “Tracing traitors,” in *Proceedings of the 14th Annual International Cryptology Conference on Advances in Cryptology*, *Lecture Notes in Computer Science* **839**, pp. 257–270, August 1994.
21. B. Chor, A. Fiat, M. Naor, and B. Pinkas, “Tracing traitors,” *IEEE Transactions on Information Theory* **46**, pp. 893–910, May 2000.
22. M. Wu, W. Trappe, J. Wang, and R. Liu, “Collusion-resistant fingerprinting for multimedia,” *IEEE Signal Processing Magazine,* **21**, pp. 15–27, March 2004.
23. D. Boneh and J. Shaw, “Collusion secure fingerprinting for digital data,” *IEEE Transaction on Information Theory* **44**, pp. 1897–1905, September 1998.
24. G. Doërr and J.-L. Dugelay, “A guide tour of video watermarking,” *Signal Processing: Image Communication, Special Issue on Technologies for Image Security* **18**, pp. 263–282, April 2003.
25. M. Holliman, W. Macy, and M. Yeung, “Robust frame-dependent video watermarking,” in *Security and Watermarking of Multimedia Contents II*, *Proceedings of SPIE* **3971**, pp. 186–197, January 2000.
26. K. Su, D. Kundur, and D. Hatzinakos, “A novel approach to collusion resistant video watermarking,” in *Security and Watermarking of Multimedia Contents IV*, *Proceedings of SPIE* **4675**, pp. 491–502, January 2002.
27. 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*, *Proceedings of SPIE* **3971**, pp. 358–370, January 2000.
28. T. Kalker, G. Depovere, J. Haitsma, and M. Maes, “A video watermarking system for broadcast monitoring,” in *Security and Watermarking of Multimedia Contents*, *Proceedings of SPIE* **3657**, pp. 103–112, January 1999.
29. E. Lin and E. Delp, “Temporal synchronization in video watermarking,” *IEEE Transactions on Signal Processing, Supplement on Secure Media* **52**, pp. 3007–3022, October 2004.
30. G. Doërr and J.-L. Dugelay, “Switching between orthogonal watermarks for enhanced security against collusion in video,” Tech. Rep. RR-03-080, Eurécom Institute, July 2003.
31. 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* **52**, pp. 2955–2964, October 2004.
32. G. Doërr and J.-L. Dugelay, “Secure video watermarking via embedding strength modulation,” in *Proceedings of the Second International Workshop on Digital Watermarking*, *Lecture Notes in Computer Science* **2939**, pp. 340–354, March 2004.
33. G. Doërr and J.-L. Dugelay, “Danger of low-dimensional watermarking subspaces,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, **III**, pp. 93–96, May 2004.
34. F. Hartung and B. Girod, “Watermarking of uncompressed and compressed video,” *Signal Processing* **66**, pp. 283–301, May 1998.
35. G. Doërr and J.-L. Dugelay, “Secure background watermarking based on video mosaicing,” in *Security, Steganography and Watermarking of Multimedia Contents VI*, *Proceedings of SPIE* **5306**, pp. 304–314, January 2004.
36. Y. Fisher, *Fractal Image Compression: Theory and Applications*, Springer-Verlag, 1994.
37. D. Kirovski and F. Petitcolas, “Blind pattern matching attack on watermarking systems,” *IEEE Transactions on Signal Processing* **51**, pp. 1045–1053, April 2003.
38. 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*, pp. 133–142, September 2004.
39. 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.