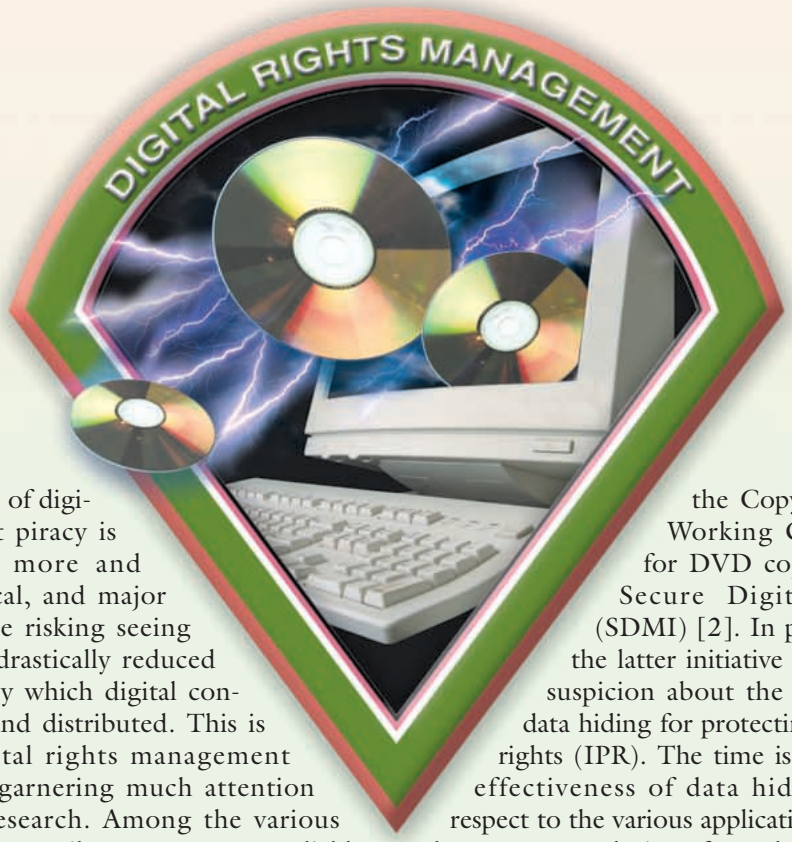
Data Hiding for Fighting Piracy

Mauro Barni and Franco Bartolini

Investigating applications, requirements, limitations, and possible future developments of watermarking technologies in DRM systems.



LOGO COMPOSITE: ©1995 PHOTODISC, INC., ©DIGITAL STOCK, ©COREL

The problem of digital content piracy is becoming more and more critical, and major content producers are risking seeing their business being drastically reduced because of the ease by which digital contents can be copied and distributed. This is the reason why digital rights management (DRM) is currently garnering much attention from industry and research. Among the various technologies that can contribute to set up a reliable DRM system, data hiding (watermarking) has found an important place, thanks to its potentiality of persistently attaching some additional information to the content itself. Many applications (ownership proofing, copy control, etc.) have been devised in this framework exploiting data hiding techniques, and many problems have emerged. Some industrial initiatives have also been carried out that tried to exploit watermarking technology for particular DRM problems: for example, the Copy Protection Technical Working Group (CPTWG) [1] for DVD copy protection and the Secure Digital Music Initiative (SDMI) [2]. In particular, the failure of the latter initiative has diffused a stronger suspicion about the actual effectiveness of data hiding for protecting intellectual property rights (IPR). The time is right for assessing the effectiveness of data hiding technology with respect to the various applications it can serve, and to draw some conclusions from the past experiences. The goal of this article is to provide an overview of watermarking principles and to analyze various applications by summarizing requirements and highlighting limitations as well as trying to foresee future developments.

DRM technology could benefit from data hiding in several ways, as is evident by the variety of watermarking-based systems addressing DRM problems proposed in the literature. In the following sections, the various aspects in which data hiding can find application in the

1053-5888/04/$20.00©2004IEEE

development of DRM systems are surveyed (e.g., ownership verification, copyright protection, item identification, etc.). Each aspect of the main characteristics the data hiding primitives should satisfy are analyzed, primary limitations are presented, and solutions and possible countermeasures are suggested.

A Primer on Data Hiding

The general model of a data hiding system is given in Figure 1. At the input of the system we find the information to be hidden and the original, nonmarked host signal A. The host signal, sometimes called the cover signal, may be an audio file, a still image, a piece of video, or a combination of the above. We assume that the information to be hidden takes the form of a binary string $\mathbf{b} = (b_1, b_2 \dots b_k)$, with b_i taking values of $\{0, 1\}$. We refer to $\mathbf{b}$ as the watermark code. The data embedding module, or simply the embedder, mixes the cover signal A and the watermark code $\mathbf{b}$ to produce a watermarked signal A_w. In order to increase the secrecy of the system, the embedding function $\mathcal{E}$ usually depends on a secret key K. Thus, in its more general form, $\mathcal{E}$ can be written as:

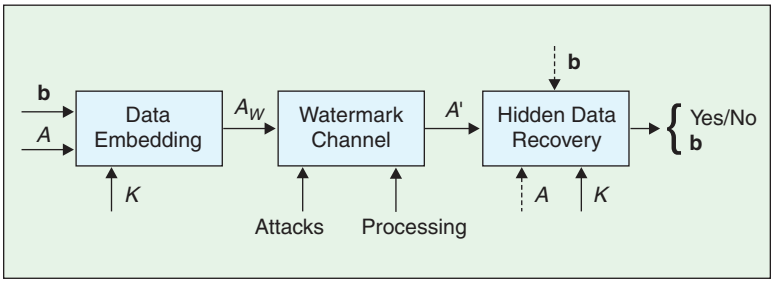
$$A_w = \mathcal{E}(A, \mathbf{b}, K). \tag{1}$$

Usually, the definition of $\mathcal{E}$ goes through the extraction from the host signal A of a set of features $\mathbf{f} = (f_1, f_2, \dots, f_n)$, called host features, that are modified according to the watermark code. Possible choices of $\mathbf{f}$ include audio or image samples, discrete Fourier transform (DFT) or discrete cosine transform (DCT) coefficients, and wavelet coefficients. In many cases it is useful to describe the embedding function by introducing a watermark signal $\mathbf{w} = (w_1, w_2, \dots, w_n)$, which is added to the host feature set. In this case, by letting $\mathbf{f}_w = (f_{w,1}, f_{w,2}, \dots, f_{w,n})$ indicate the watermarked host features, we have:

$$f_{w,i} = f_i + w_i, \qquad i = 1, \dots, n. \tag{2}$$

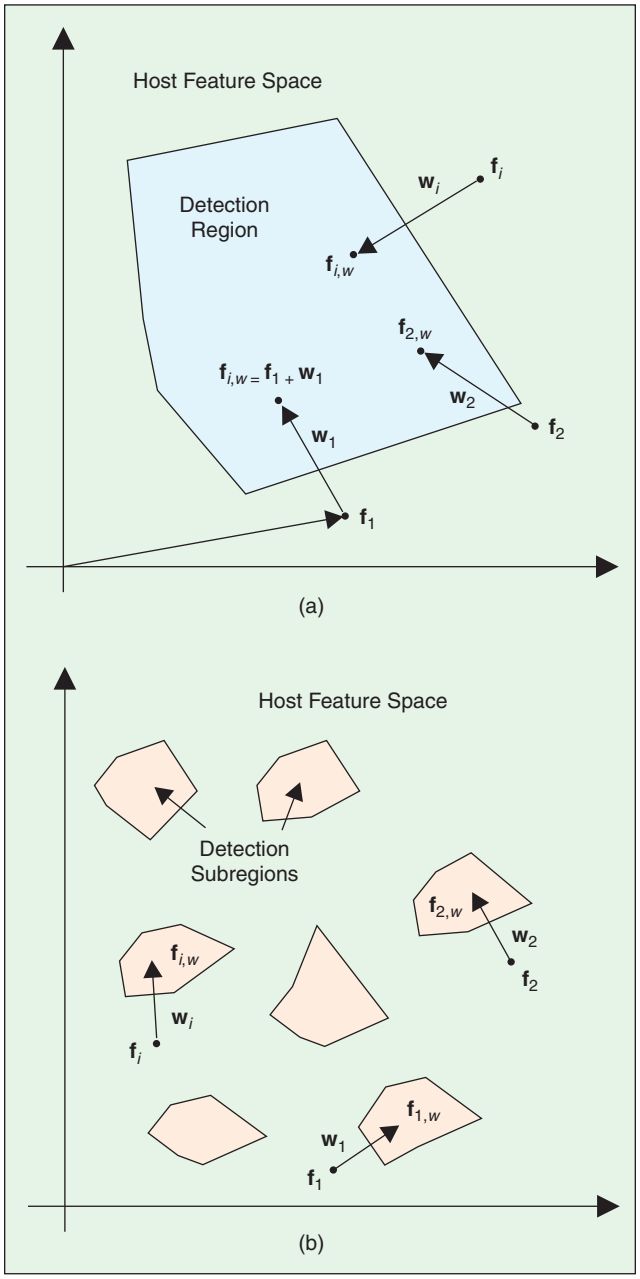
In the simplest case $\mathbf{w}$ depends only on $\mathbf{b}$ and K. For instance, according to the popular spread-spectrum approach [3], a pseudorandom sequence $\mathbf{s} = (s_1, s_2, \dots, s_n)$ is generated depending on K, then $\mathbf{s}$ is modulated by means of an antipodal version of $\mathbf{b}$. A data hiding system for which $\mathbf{w}$ does not depend on A is called a blind embedding system, since embedding is carried out blindly without taking into account the particular host signal at hand. One may guess that better results can be obtained by tailoring $\mathbf{w}$ to the host asset A. This is indeed the case, as it has been shown in a number of seminal papers that have appeared since the late 1990s [4]–[8] referring back to the theory of digital communications through channels with side information at the encoder [9], [10]. More specifically, by assuming that the detector structure and the corresponding detection regions are known to the embedder, the embedding problem may be seen as the mapping of the host signal into a point within the correct detection region [see Figure 2(a)]. (Calling the watermark extraction module a detector is somewhat imprecise, since the difference between watermark detection and decoding should be considered. We will be more precise on this aspect later on in the article.) If the role of the watermark signal has to be retained [as in (2)], we must now admit that $\mathbf{w}$ depends on the host signal A, since in this way it is possible to push the watermarked signal more inside the detection region, given a desired level of distortion [according to (2), the embedding distortion simply amounts to the norm of $\mathbf{w}$]. The informed embedding principle may be pushed further, by letting detection regions to be composed by a set of nonconnected subregions spread over all the signal space, and by deciding to map the host signal inside the subregion that results in the lowest distortion [see Figure 2(b)]. Data hiding systems obeying this strategy are collectively termed "informed watermarking," or "informed data hiding," systems [8]. A clever way to put the informed data hiding strategy to work is through the class of quantization index modulation (QIM) algorithms [6], [8]. In QIM schemes, data hiding is achieved through the quantization of the host feature vector, according to a set of predefined quantizers.

The second element in the scheme of Figure 1 is the so-called watermark channel. This accounts for all the manipulations the host signal may undergo after information embedding. Note that both intentional and nonintentional manipulations must be taken into account, with the former accounting for the possible presence of an enemy, usually called the attacker, acting with the explicit goal of damaging the hidden message, and the latter accounting for the manipulations the host signal may undergo during its normal life cycle (e.g., lossy coding, resizing, filtering). The ability to survive intentional attacks is referred to as watermark



▲ *1. Overall picture of a data hiding system. The watermark code* **b** *is embedded into the host signal A, thus producing the watermarked asset* A_W*. Due to possible attacks,* A_W *is transformed into A′. Finally, the hidden information is recovered from A′, either by extracting the hidden message* **b** *or by deciding whether A′ contains a known watermark code* **b** *or not. Watermark embedding and recovery require the knowledge of a secret key K. Watermark recovery may benefit from the knowledge of the original, nonmarked signal A.*

security, whereas resilience against nonintentional manipulations is referred to as watermark robustness. While the current state of the art of data hiding technology can provide a good degree of robustness against the most common, nonmalevolent manipulations, with the noticeable exception of geometric manipulations, watermarking security is still an open issue. For the sake of brevity, we will not delve into security details here (a comprehensive, yet simple, introduction to watermarking security is given in [11]). It is only important to point out that security requirements heavily depend on the applications. For example, scenarios where pirates can freely access the detector are by far more complex than those in which the extraction device is not publicly available.



▲ *2. (a) Informed embedding and (b) informed coding paradigms. In informed embedding, given a watermark detection region in the feature space, watermark embedding is seen as the mapping of the host feature set into a point within the detection region. The same action can be seen as the addition of a watermark signal* $\mathbf{w}_i$ *which depends on* $\mathbf{f}$*. In informed coding the watermark detection region is formed by a number of subregions scattered across the feature space. The embedder maps the to-be-marked signal into the subregion resulting in the lower distortion (or the maximum robustness).*

After the host signal has passed the watermark channel, it enters the detector, whose scope is to retrieve the hidden information. Extraction of the hidden information may follow two different approaches: the detector looks for the presence of a specific message, thus only answering yes or no, or the detector (which in this case is called a decoder) reads the information conveyed by the host signal without knowing it in advance. These two approaches lead to a distinction between algorithms embedding a message that can be read (readable or multibit watermarking) and those inserting a code that can only be detected (detectable or 1-bit watermarking). An additional distinction may be made between systems that need to know the original, nonmarked signal A in order to retrieve the hidden information and those that do not require it. In the latter case we say that the detector is blind (the term oblivious detection may also be used) whereas in the former case the detector is said to be nonblind.

In all the cases, the retrieval of $\mathbf{b}$ goes through the definition of a detection (decoding) function $\mathcal{D}$. In oblivious, detectable watermarking, $\mathcal{D}$ is a three-argument function accepting as input a digital asset A', a watermark code $\mathbf{b}$, and a secret key K. As an output $\mathcal{D}$ decides whether A' contains $\mathbf{b}$ or not, that is

$$\mathcal{D}(A', \mathbf{b}, K) = \text{yes/no}. \tag{3}$$

In the nonoblivious case, the original asset A is a further argument of $\mathcal{D}$. In blind, readable watermarking, the decoder function takes as inputs a digital asset A' a keyword K, and gives as output the string of bits $\mathbf{b}$ it reads from A':

$$\mathcal{D}(A', K) = \mathbf{b}. \tag{4}$$

In the nonoblivious case, the original asset A is a further argument of $\mathcal{D}$. Note that, in readable watermarking, the decoding process always results in a decoded bit stream; however, if the asset is not marked, decoded bits are meaningless.

The exact form of $\mathcal{D}$ depends on the particular watermarking algorithm. The most common solution with spread-spectrum systems relies on the analysis of the correlation between the spreading sequence $\mathbf{s}$ and the feature vector $\mathbf{f}'$ [3] and its comparison against a detection threshold. Alternatively, the normalized correlation may be used [3]. More recently, the optimum detector/decoder structure has been derived for several schemes, thus improving considerably their performance [12]–[14].

In the case of QIM watermarking, decoding is straightforward, since it only requires that the host features are quantized again by considering a codebook that is the union of all the possible codebooks used by the embedder. In practice this corresponds to a minimum distance decoder, which, in the case of a Gaussian noise addition attack, represents the optimum decoding strategy.

So far we have implicitly assumed that the secret key K used in the decoding/detection process is the same used for embedding. We term these kinds of algorithms as symmetric watermarking schemes. A problem with symmetric watermarking is an intrinsic lack of security, especially if the decoder/detector is implemented in publicly available consumer devices. The knowledge of K, in fact, is likely to give attackers enough information to remove the watermark from the host signal. In order to overcome the above problems, increasing attention has been given to the development of asymmetric schemes [15]. In such schemes two keys are present, a private key, K_s, used to embed the information within the host signal, and a public key, K_p, used to detect/decode the watermark (often K_p is just a subset of K_s). Knowing the public key, it should be neither possible to deduce the private key nor to remove the watermark (unlike in asymmetric cryptography, knowledge of K_s may be sufficient to derive K_p; additionally, the roles of the private key and the public key cannot be exchanged). More details about the importance of asymmetric watermarking in DRM and more generally in security-oriented applications may be found in [11].

Proof of Ownership

This is the most classical scenario served by data hiding: the author of a piece of work, say Alice, wishes to prove that she is the only legitimate owner of the work. To do so, as soon as she creates the work, she embeds within it a watermark identifying her unambiguously. In the sequel, watermark extraction can be used to verify Alice's ownership over the work since, due to the impossibility of removing the watermark, all the copies of the work will contain the watermark, thus linking them to Alice.

Apparently, the requirements to be satisfied by a watermarking algorithm to be used for rightful ownership verification are easily identified. It is obvious, in fact, that for any scheme to work, the watermark must be a secure one, given that pirates are interested in removing the watermark, possibly by means of computationally intensive procedures. As to capacity, the exact requirements depend on the number of different identification codes the system must accommodate for.

Security Threats

While robustness against nonmalevolent manipulations of the host signal is of utmost importance, we focus here on those manipulations explicitly aiming at removing the watermark. With respect to other scenarios, e.g., those encountered in copyright protection applications, the ownership verification scenario is a favorable one since we can assume that the detector (decoder) is not publicly available. This is reasonable since watermark verification is due either to the owner itself or to a trusted third party (TTP). This is a particularly important feature, since pirates cannot know whether their attacks are successful or not and learn from their previous trials how to remove the watermark.

The most dangerous attack, in this case, is the so-called average attack. By assuming that the attacker can access a number of different works belonging to the same owner, he can average all these works and obtain a good estimate of the watermark signal embedded within them. If the pirate knows the details of the watermarking system employed to mark the works, he can exploit the knowledge of $\mathbf{w}$ to unwatermark the protected works. Note that in most cases this does not require that the embedding or detection keys are known, since the attacker works directly with the embedded signal $\mathbf{w}$. The most common countermeasure against the average attack consists of adotping a host-signal-dependent watermark, where $\mathbf{w}$, or even $\mathbf{b}$, varies from one cover work to the other.

Watermark (Quasi-)Invertibility

Even if security is a crucial requirement, a closer look at the protocol level reveals that this is not the only threat to be considered. In addition to ensuring that the true watermark cannot be removed, the system must in fact guarantee that a fake watermark cannot be inserted within the protected work. If this is the case the presence within the work of two different watermarks would, in fact, make it impossible to determine the true document owner.

To be specific, let us assume that to protect Alice's work, she adds a watermark with her identification code $\mathbf{b}_A$ to it, thus producing a watermarked work $A_{w_A} = A + \mathbf{w}_A$ (the symbol + is used to indicate watermark casting since we assume, for simplicity, that the watermark is added to the host signal), then she makes A_{w_A} publicly available. For sake of brevity, we assume that a detectable watermarking scheme is used (the extension to the case of multibit watermarking being easy). To fool the ownership verification mechanism, an enemy, say Bob, adds his own watermark $\mathbf{b}_B$ to A_{w_A}, producing $A_{w_A w_B} = A + \mathbf{w}_A + \mathbf{w}_B$. Given that $A_{w_A w_B}$ contains both Alice's and Bob's watermarks, it is impossible to decide whether it belongs to Bob or Alice. To exit the deadlock, Alice and Bob can be asked to exhibit a copy of the work that contains their watermark but does not contain the watermark of the other contender. Of course, this is an easy task for Alice, since she owns the original work. Suppose, however, that the watermarking technique used by Alice is not blind. For instance, let us assume that the watermark is detected by subtracting the original signal from the watermarked one. Alice can use the true original signal to show that

Bob's copy contains her watermark and that she possesses a copy containing $\mathbf{w}_A$ but not $\mathbf{w}_B$. The problem is that Bob can do the same thing by building a fake original work A_f. It is, in fact, sufficient that Bob subtracts his watermark from A_{w_A}, maintaining that the true original signal is

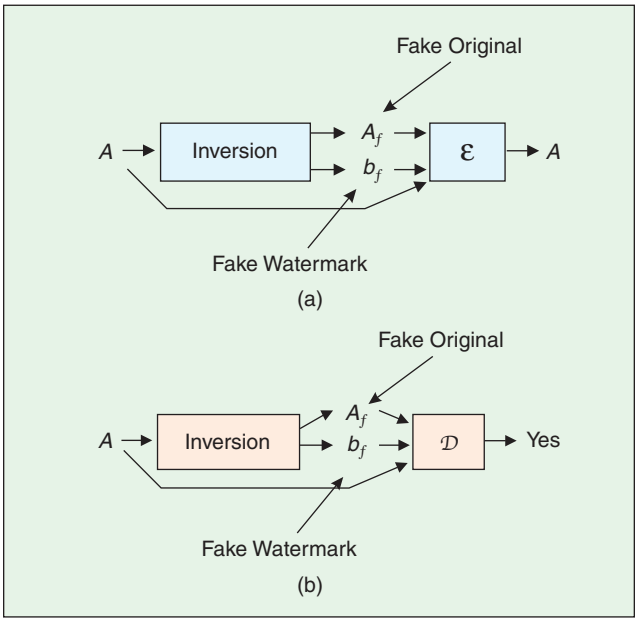
$$A_f = A_{w_A} - \mathbf{w}_B = A + \mathbf{w}_A - \mathbf{w}_B. \quad (5)$$

In this way Bob is able to prove that he possesses a copy of the work, namely the publicly available copy A_{w_A} that contains $\mathbf{w}_B$ but does not contain $\mathbf{w}_A$:

$$A_{w_A} - A_f = A + \mathbf{w}_A - (A + \mathbf{w}_A - \mathbf{w}_B) = \mathbf{w}_B. \quad (6)$$

Note that, in this way, Bob can also use the fake original work and the fake watermark to show that the original work in Alice's hands contains his identification code. As can be seen, the plain addition of a nonblind watermark to a piece of work is not sufficient to prove ownership.

At the core of the above attack, sometimes called the SWICO attack (single-watermarked-image-counterfeit-original), there is the possibility of building a fake original work (A_f) and a fake watermark ($\mathbf{b}_f$) such that the insertion of the fake watermark within the fake original produces a watermarked work that is equal to the initial one [16] [see Figure 3(a)]. Note that, generally, the SWICO attack involves two degrees of freedom, since both the fake original work and the fake watermark can be adjusted to reverse engineer the watermarking process.



▲ *3. Sketches of the (a) SWICO and (b) TWICO attacks. While for the SWICO attack it is required that embedding the fake watermark into the fake original produces the publicly available (watermarked) asset, for the TWICO attack it is only required that the fake watermark is detected in the publicly available (watermarked) asset.*

A more sophisticated version of the SWICO attack, namely the TWICO attack (twin-watermarked-images-counterfeit-original) leads to the concept of quasi-invertibility [16]. The extension relies on the observation that, for the SWICO attack to be effective, it is only needed that when the watermark detector is applied to A by using the fake original work, the presence of the fake watermark is revealed [see Figure 3(b)]:

$$\mathcal{D}(A, A_f, \mathbf{b}_f) = \text{yes}. \quad (7)$$

We can conclude that if watermarking must be used to identify the owner of a piece of work, the non(quasi-) invertibility of the watermark has to be proved. A possible way to alleviate this problem consists of adopting a blind data hiding scheme, where watermark detection is accomplished without resorting to the original nonmarked signal. The inversion of a blind watermark has only one degree of freedom, thus making it easier to prevent it by acting at a protocol level, e.g., by requiring that watermarks are assigned by a TTP, thus avoiding the use of ad-hoc fake watermarks, or by letting the watermark code depend on A. A similar strategy could be conceived in the nonblind case; however, more attention is needed, since the two degrees of freedom implicit in the inversion of a nonblind watermarking scheme could make it possible to handle a situation in which $\mathbf{b}_f$ is fixed and pirates only act on A_f.

Even if the noninvertibility (nonquasi-invertibility) of the watermark cannot be granted, data hiding technology can still be useful in less demanding applications, where demonstration of ownership in front of a court of law is not required. For example, Alice may wish to detect suspicious products existing in the distribution network. Such products could be individuated by an automated search engine looking for the watermark presence within all the works accessible through the network. At the same time, Alice may rely on more secure mechanisms to prove that she was the victim of a fraud, e.g., by depositing all her creations to a registration authority.

Copyright Protection

This has been one of the first industrial DRM applications of watermarking, surely the one that triggered the attention toward watermarking and data hiding. In the early days of watermarking research, it was thought, in fact, that simply embedding a flag within the work to be protected, e.g., stating that the cover work could not be copied, was enough to prevent fraudulent copying. It was soon realized, though, that a much deeper analysis is necessary before data hiding can be effectively used to enforce, or at least to help enforce, copyright laws. First of all, the embedded watermark must be robust against nonintentional processing and secure against intentional attacks. This turned out to be an

extremely challenging task that is far from being solved (especially with regard to security). As a matter of fact, in recent years, the failure of some attempts (e.g., SDMI [2], [17]) to develop a secure watermarking scheme has caused a boomerang effect, undermining the trust on watermarking technology in general. In the meantime, impressive progress has been made in terms of both achievable robustness and capacity; thus, even if security may still be out of reach, high capacity, robust watermarking may soon become a reality. Yet the design of a watermarking-based copy protection mechanism is not only a matter of robustness, since protocol issues must be considered as well.

In the following, we review the main watermarking-based copyright protection scenarios proposed so far. We divide them into two main categories: those aiming at discouraging illegal copying and those aiming at preventing it.

Infringement Tracking for Illegal Copying Dissuasion

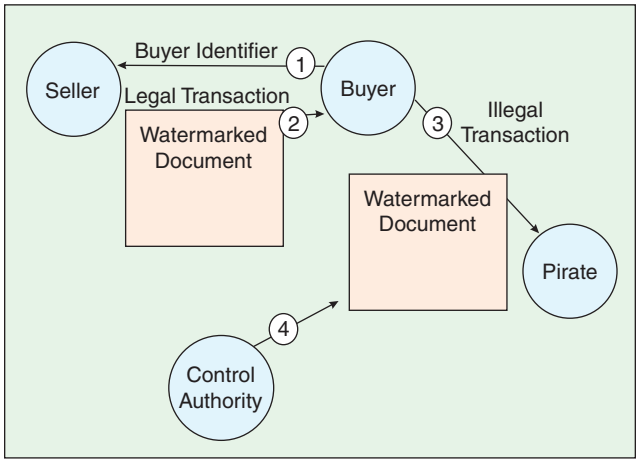
According to this scenario, a so-called copy deterrence mechanism is adopted to discourage unauthorized duplication and distribution. Copy deterrence is achieved by providing a mechanism to trace unauthorized copies to the original owner of the work or, more generally, to track the author of the infringement. In the most common case, distribution tracing is made possible by letting the seller insert a distinct watermark, which in this case is called a fingerprint, identifying the buyer, or any other addressee of the work, within any copy of data that is distributed. (It is worth noting here that the term "fingerprinting" has been recently used for another type of technology aimed at extracting from a digital document a distinctive set of unique characteristics (fingerprint) that can be later used for identifying it [18], [19].) If, later on, an unauthorized copy of the protected work is found, then its origin can be recovered by retrieving the unique watermark contained in it (Figure 4). A similar scheme has been recently proposed for avoiding illegal copying and distribution of digital cinema [20].

The main requirement set by fingerprinting applications is security since any attempt to remove the watermark or making it unreadable must be prevented. At the same time a readable watermarking scheme is preferable, since in many cases it is not possible to guess in advance the watermark content. The latter requirement may be relaxed by properly designing the infringement tracking protocol, possibly at the expense of simplicity [21].

Security Threats

A first crucial point to be considered is whether the fingerprinting protocol makes provision for public watermark decoding (publicly available decoder) or not. Of course, the former case is more difficult to treat since the attacker can exploit the decoder to infer useful information about the hidden information. This results in the sensitivity attack followed by the so-called closest-point attack. Due to the importance of such attacks in copy control scenarios, we postpone their analysis to the next section.

If the watermark decoder is not publicly available, two possible attacks are still to be considered carefully. The first one is the average attack described previously. The pirate may use a number of different works, all marked with his name, to infer the watermark signal contained in them, and, subsequently, to exploit such a knowledge to unwatermark them. A second possibility is the collusion attack. In this case a pool of pirates in possession of different copies of the same work, each marked with a different watermark, team together and average their copies to obtain a version of the host work in which all the different watermarks are so weak that they can no longer be decoded. Of course, the larger the number of colluders, the higher the probability that the watermarks are successfully removed. A deep analysis of the collusion attack [22] for the case that the watermarks consist of orthogonal Gaussian signals consisting of 10,000 samples, and that a nonoblivious detector is used, has shown that if less than 20 copies are averaged, the collusion attack cannot succeed regardless of the number of copies available, while if more than 80 copies are averaged, the attacks is always successful. As noted above, a straightforward countermeasure against the average attack consists in letting the watermark depend on the host signal. Another possibility is to use anticollusion codes [23], [24]. In this case the watermark messages $\mathbf{b}$s and the embedding strategy are chosen in such a way that averaging different watermark signals leaves certain parts of



▲ *4. Simplified scheme exemplifying a fingerprinting protocol (the numbers indicate the sequence of the operations). Before the multimedia document is legally acquired, the buyer has to provide his own identifier to the seller; the seller embeds this identifier into the document and gives it to the buyer; if the buyer illegally distributes the acquired content to another party (here the pirate) a control authority can detect his identifier into the document, and take actions against him.*

the watermark unaffected, thus permitting the recovery of some, possibly collective, information about the colluding pool. A drawback with these codes is that the number of possible messages the system can accommodate is greatly reduced.

Protocol Issues

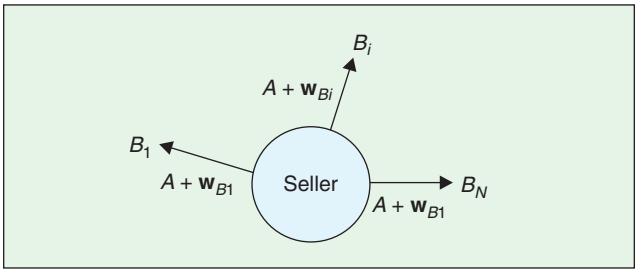
A problem with the plain fingerprinting scheme described above is that buyers' rights are not taken into account, since the watermark is inserted solely by the seller without any control. Thus, a buyer whose watermark is found in an unauthorized copy cannot be prosecuted legally since he can claim that the unauthorized copy was created and distributed by the seller. To understand why the seller could try to catch the buyer, let us consider the situation depicted in Figure 5, where the seller acts as an authorized reselling agent. The seller may distribute many copies of a work containing the fingerprint of buyer B_1 without paying the royalties he owes to the author and claim that such copies were illegally distributed by B_1. As in the case of rightful ownership demonstration, a solution consists in resorting to a TTP, e.g., by letting the TTP insert the watermark within the work to be protected and retrieve it in case a dispute resolution protocol has to be run. A problem with the presence of a TTP in practical applications is that it may easily become the bottleneck of the whole system. In addition, the protected work must be transmitted from the seller to the TTP and from the TTP to the customer, or, in an even worse case, from the TTP to the seller and from the seller to the customer, thus overloading the communication channel.

A clever way to avoid the above difficulties and still ensure that buyers' rights are respected relies on the joint exploitation of watermarking and cryptography, as suggested by the interactive buyer-seller (IBS) protocol described in [25]. Even in this case, the presence of a TTP is envisaged; however, its role is minimized. Data exchange is kept to a minimum as well, resulting in a very low communication overhead. The basic idea the IBS protocol relies on is that watermark embedding is performed directly in the encrypted domain. This is possible because the IBS protocol uses a cryptosystem that is a privacy homomorphism with respect to the embedding rule $\mathcal{E}$, that is:

$$E_K(\mathcal{E}(A, \mathbf{b})) = \mathcal{E}(E_K(A), E_K(\mathbf{b})), \tag{8}$$

where E_K denotes encryption. Though strange at first sight, the privacy homomorphism requirement is not difficult to satisfy. For instance, it is known that the RSA cryptosystem is a privacy homomorphism with respect to multiplication. Encryption domain watermarking permits avoiding that the seller gets to know the exact watermarked copy received by the buyer, hence avoiding that he distributes copies of the original work containing the buyer's identification watermark. In spite of this, the seller can identify the buyer from whom unauthorized copies originated and prove it by using a dispute resolution protocol. The same protocol can be used by the buyer to demonstrate his/her innocence.

With regard to the validity of fingerprinting as a proof in front of a judge, it must be said that the current state-of-the-art allows this possibility only if a trusted watermarking/certification authority is included within the copyright protection protocol. Nevertheless, it is important to stress that, even if plain fingerprinting may not be considered a proof from a legislative point of view, it may be useful in several situations; e.g., to enable the seller to identify potentially deceitful customers and break off any further business relationship with them.



▲ *5. In fingerprinting applications, the situation depicted in the figure, where several copies of the host work containing the identification code of client B_1 are distributed to other purchasers (e.g., to purchaser B_N who is deceived by the seller), must be avoided so as to take into account buyer's rights.*

Copy Control

When copy deterrence is not sufficient to effectively protect legitimate rights holders, a true copy protection mechanism must be envisaged. Having said that a comprehensive solution of the copy protection problem goes well beyond watermarking technology, we describe a mechanism that has been considered for protection of DVD video. This scenario, in fact, represents a good example of how watermarking can be integrated in a complex copy protection system and effectively contribute to its efficacy.

A possible approach to make illegal duplication and distribution difficult enough to limit the losses caused by missed revenues relies on the distinction between copyright compliant devices (CC devices) and noncompliant devices (NC devices), where CC devices are designed to refuse to make copies when they are not explicitly allowed. Then, the copy control mechanism consists of keeping the CC and the NC worlds as separate as possible (see Figure 6), for example, by allowing NC devices to play only illegal disks and CC devices to play only legal disks. In this way, users willing to play both legal and illegal disks must buy two series of devices (of course nonprotected disks would be *playable* on both kinds of devices).

The possibility that a legal disk is played on a NC device is prevented by means of a proper content scrambling system (CSS). Descrambling requires a pair

of keys, one of which is unique to the video file, while the other is unique to the DVD. Keys are stored on the lead-in area of the DVD, an area that is only read by CC devices. Note that further protection is achieved by making it impossible that the output of a CC player is connected to a NC recorder, since CC devices are not allowed to dialog with NC devices. On the other hand, recording through CC devices is governed by a copy generation management system (CGMS), which allows copying only if this is permitted for that particular disk. Simply speaking, CGMS relies on 2 bits stored in the header of the video stream, encoding one of the following three indications: copy freely, copy never, and copy once, where the result of the copy-once indication is that the video can be copied, but after copying, the CGMS bits are changed to copy never.

The above mechanisms (CSS and CGMS) hinder the flow of videos from the legal toward the NC world; however, to discourage illegal copying it must also be avoided that a CC device is used to play or record an illegal disk. To this aim, the sole CSS is not sufficient, since the pirate may decide to pass from the digital to the analog world to remove the CSS protection. This is possible, for example, by using the analog RGB output of a compliant device to make a nonencrypted copy of the video by means of an NC recorder and trying to reenter the CC world by letting CC devices mistake the illegal video for a free video without protection. This is possible because both CSS and CGMS bits do not survive digital-to-analog and analog-to-digital conversion.

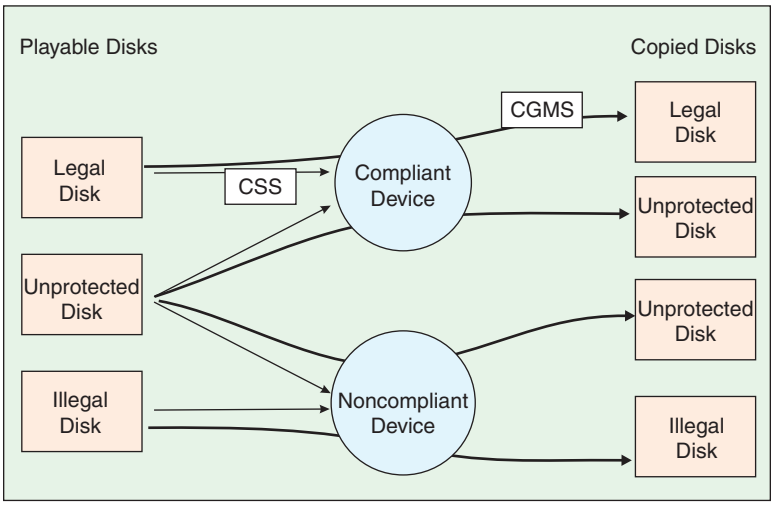
Data hiding can help solve this problem by embedding CGMS bits within the video in the form of a secure watermark. It is obvious that the presence of CGMS bits prevents video recording on a CC recorder, since, upon reading the CGMS bits, the CC devices refuse to copy the video if CGMS bits indications do not allow it. At the same time, CC players can be designed to recognize as illegal a DVD copy without CSS, yet containing the CGMS watermark, and refuse playing it. As desired, the worlds of CC and NC devices are kept separate, since illegal disks can only be managed by NC devices and legal disks by CC devices. More details about DVD copyright protection and the role of watermarking in such a framework may be found in [1].

Among the requirements to be satisfied by a watermarking system to be used for copyright protection, security is by far the most demanding one, since it is foreseeable that the system will operate in a very wide and hostile scenario, where the possibility that even a single user breaks the system compromises the effectiveness of the whole architecture. At the same time, by looking at the DVD example described above, it is readily seen that watermarking may provide a unique feature that cannot be obtained by standard means such as encryption or format standardization; i.e., the possibility of surviving a passage through the world of analog signals.

The Sensitivity Attack

Among the security threats tormenting watermark-based copy control, a central role is played by the sensitivity attack followed by the closest-point attack. Such an attack derives from the fact that, in copy control applications, the watermark decoder/detector has to be made publicly available in low-cost consumer electronic devices. The first problem, then, consists of making the devices as tamper proof as possible, since they contain enough information to remove the watermark (this is particularly true with symmetric watermarking schemes where the knowledge available at the detector/decoder coincides with that available at the embedder).

Even by assuming that the detector/decoder is tamper proof, a very effective attack can still be conceived. Without loosing generality let us assume that a detectable watermarking scheme is used, and let us start by considering a pirate who iteratively modifies the document until the detector is no longer able to recover the watermark. This is certainly possible if the pirate can access the detector; however, with the modifications performed almost randomly, the time needed to find a successful hack within a low-quality loss is likely to be extremely high. A possibility to speed up this attack consists of first performing a learning phase in which the boundary of the detection region is estimated. By assuming that the detection region is described by a parametric function with n degrees of freedom, it only needs that n points lying on, or in the vicinity of, the border are found. This can be easily accomplished by modifying at random the marked features, and, once a feature vector judged as nonmarked



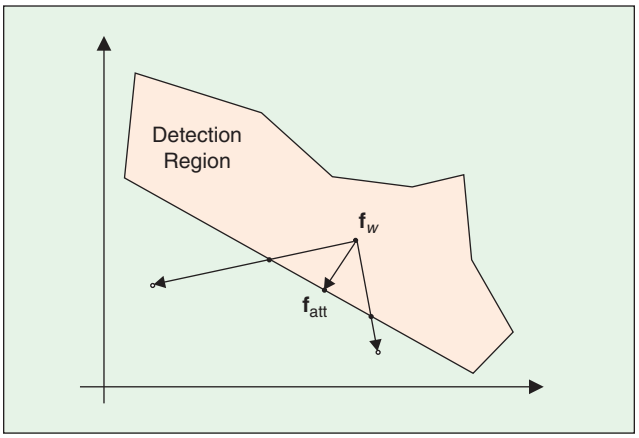
▲ *6. The figure depicts the separation between the worlds of copyright compliant and noncompliant devices: the bold lines represents possible copy flows; the normal lines represent possible playing actions. Only nonprotected disks can be played by both kinds of devices.*

by the detector is found, by getting closer to the detection region boundary by iteratively moving the host features on the line joining the original and the unmarked feature vector (see Figure 7). Once the boundary of the detection region is known, the point on such a boundary that is closest to the original marked features can be easily found ($\mathbf{f}_{att}$ in Figure 7), thus leading to a minimally distorted nonmarked host work (closest-point attack).

The sensitivity attack coupled with the closest-point attack is a very general and effective attack that cannot be easily avoided. Two possible solutions have been proposed so far: i) to design the watermarking system so that the boundary of the detection region is not a parametric one [26], and ii) to adopt an asymmetric watermarking scheme in which the key used to detect the watermark, if any, does not reveal the parameters used during the embedding phase [15]. Though these are promising solutions, their effectiveness has not been proved yet; hence, further research is needed before the sensitivity plus the closest-point attack no longer hampers the practical implementation of copy control mechanisms relying on data hiding technology. For a more detailed discussion of the sensitivity and the closest-point attacks, readers may refer to [11] and [27].

Item Identification

One of the main features of watermarking technology is that it provides a way to attach a code to a multimedia document in such a way that the code is persistent with respect to the possible changes of format the document may undergo. As an extreme case, the embedded code could be resistant even to digital-to-analog conversion. To fully exploit the potentiality of this peculiar characteristic, the concept of persistent association has been developed during the past few years. The basic idea is to associate a unique identifier (UI) to each multimedia creation. The UI is embedded inside the document itself by means of a watermarking primitive and is used for indexing a database where more detailed information (not only related to IPR) can be retrieved (Figure 8). The use of watermarking for tightly attaching a UI to a document is proposed in the Content ID Forum Specification [28]; this concept is finding wider interest in the framework of the MPEG-21 standardization process, where a list of requirements for persistent association tools (PATs) has been issued [29]. Among the many identified requirements, the most relevant from the point of view of the analysis carried out in this article are surveyed below. First of all, the watermark should be able to survive the types of transformations typically encountered in the life of a document, including conversions to and from the analog domain; i.e., the watermark should be robust. It is also required that the association created by means of watermarking is secure; i.e., it should resist transformations deliberately performed for removing it. Authorized users should be able to remove or at least change the status of the watermark, but when this is done without authorization, it should be possible to obtain sufficient evidence of the former presence of a watermark to be used in forensic examinations.



▲ *7. Sketch of the sensitivity attack followed by the closest-point attack. The boundary of the detection region is a hyperplane. The watermark is removed from* $\mathbf{f}_w$ *by first estimating the detection boundary and then looking for the point* $\mathbf{f}_{att}$ *of the boundary that is closest to* $\mathbf{f}_w$*. Boundary estimation is performed by first moving the host work outside the detection region (white dots) and then by moving on the line connecting the marked and nonmarked points (in the figure the outputs of this process are indicated by the black dots).*

A DRM protocol based on the persistent identification of the multimedia document to be enjoyed should work by querying an IPR database with the UI extracted from the document and by analyzing the returned licensing rules (MPEG-21 is also investigating the standardization of languages for the description of the rights associated with a multimedia document [30]) to decide if and how the document can be used. Such an approach allows a much higher flexibility than the simple protocols of ownership verification, copy control, or infringement tracking that we have seen above. The cost of this is major complexity of the verification process that requires a trusted archive to be queried. From the point of view of data hiding technology, this approach offers some advantages. As a first example, the TWICO attack is more difficult given that the embedded UI is provided by a TTP. Similarly, given that each document has its own unique watermark, collusion and averaging attacks are no longer feasible.

On the other hand, other attacks are feasible, for example, the so called collage attack, the copy attack, or the template removal attack. The first consists of building a copy of the watermarked document by substituting patches of it with similar patches taken from nonwatermarked or differently watermarked documents. The second uses advanced statistical estimation techniques to estimate the watermark from a document and transferring it into another document (thus creating an identification ambiguity). The last attack attempts to remove the synchronization pattern that is often used

by watermarking systems for resisting geometrical transformations. More details about these attacks can be found in [31]. Furthermore, it is likely that watermark recovery will have to be public, thus making the distribution of keys a major problem; on the other hand, if the keys are meant to be kept secrete inside the watermark recovery devices, the system would be weak against the previously analyzed sensitivity and closest-point attacks. Finally, it is worth mentioning that the concept of persistent association, as approached by MPEG-21, can also have a wider scope: it is, in fact, foreseen that the persistent association could regard not only a UI but also transactions identifiers (thus implementing a generalized fingerprinting service), users, and temporal information (i.e., a time stamp).

Providing Added Values

As we have seen above, data hiding techniques can endure many types of attacks, and up to today no technique has exhibited enough resilience against all of them. These considerations urged researchers to find novel approaches to the problem of DRM, and there have been some proposals in which watermarking can still be helpful.

The main assumption is that it is (at least today) almost impossible to design a secure watermarking technique, but it is really feasible to get a robust one. The main approach thus consists of trying to motivate users neither to attempt to remove the watermark nor to distribute the legally acquired and watermarked document for free. This is obtained by assigning the function of enhancing the host data content to the watermark [32]: as an example, the watermark could give access to discounts on other documents or to update services to a more sophisticated version of the document or to other added-value services (trials on other products, bonus programs, etc.). In this way, the users would not be motivated to remove the watermark as this would deprive them of the associated advantages; they are not motivated to distribute the watermarked document as well since this would mean giving that advantages to others. Similarly, users would be not motivated to illegally acquire nonwatermarked documents as this would not offer them the enhanced values associated with the watermark. This approach mainly requires that the watermark is resistant to the copy attack to avoid dishonest users transferring the advantages associated with a given document to another. Furthermore, the integration with effective cryptography protocols is required in order to avoid misuses. The use of a watermark associated with the added-value information, instead of the simpler format-header-based approaches, is still motivated by the characteristic of persistence of the watermark.

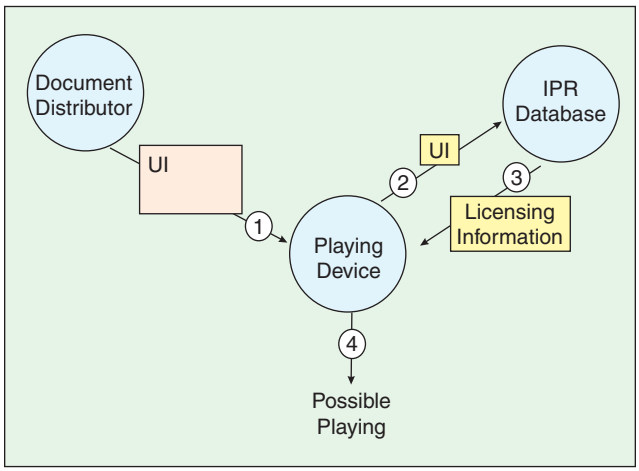
Indeed, this approach does not appear to be very effective: it would be really successful only if the added value services associated with the watermark were really of high value. On the contrary, what is of most interest for the widest part of the users is the document itself and not the possibly associated added values.

New Business Models

In this section we try to analyze whether watermarking technology can be of help in novel business models. It is not our goal to evaluate the efficiency of the alternative models but just to investigate their feasibility from a data hiding point of view.

We still start from the assumption that it is quite difficult (at least today) to devise a watermarking tool satisfying the security requirements related to its use in a highly hostile environment. On the other hand, we observe that a really successful application of data hiding technology is represented by the monitoring of broadcast transmissions [33]. Finally, we take into account that, due to the difficulty to devise effective countermeasures against the illegal distribution of copyrighted materials, in particular on peer-to-peer networks [34], it has been proposed to substitute the main source of revenues of multimedia document producers, which is now the direct selling of the goods, with a share on some levies added to the prices of recording material (e.g., CD-R, and CR-RW) and devices. This is indeed already done in some European countries for compensating producers for the losses caused by piracy. The distribution of these levies among the producers could be made on the basis of lump sums. We deem that watermarking-based monitoring of recording sessions could make this latter approach more efficient, fair, and (hopefully) attractive.

In practice (see Figure 9) every recording device (be it hardware or software) could be able to extract the hidden identification information from the document to be recorded and send it to a trusted organization monitoring and reporting on the recording activities, thus allowing the precise sharing of the levies. Then,

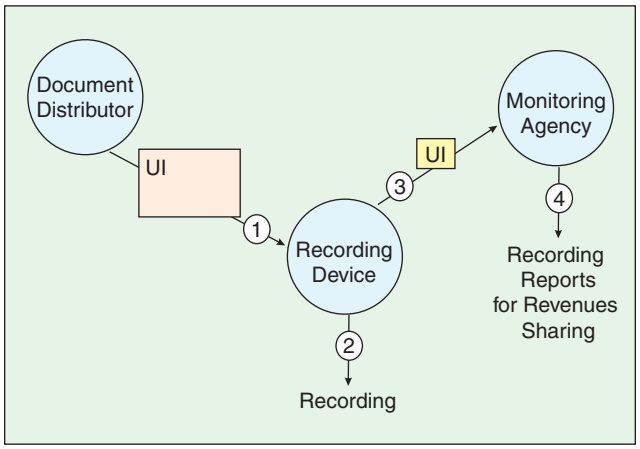


▲ *8. Thanks to the persistent association of a UI to each multimedia document, a player can request the appropriate licensing information from an IPR database and, as a consequence, apply the corresponding copyright policy to the document. The numbers indicate the sequence of the operations.*

instead of forbidding the Internet-based exchange of documents, this exchange could be exploited to increase the diffusion of the documents themselves. In this way, no user would be motivated to remove the watermark from the document as this does not give him any advantage, and watermark robustness would be enough. The problem of security is translated from the users to the producers' side, as it should be granted that they are not allowed to deceive the monitoring service. As an example, a multimedia document producer could set up false recording servers that simulate recording sessions of their productions, by sending fake monitoring information to the trusted organizations (i.e., by indirectly forging the monitoring reports). This could be avoided, for example, by letting the validation of a recording session depend on the combination of the document identifier and of a unique number associated to each recording support (e.g., to each CD-R): recording sessions duplicating this combination could be then discharged by the monitoring service. In general, anyway, it is likely that controlling the correct behavior of a few multimedia document producers could be easier than trying to develop an effective DRM system in a world populated by several millions of potentially hostile users.

The main advantage of the above approach resides in the fact that the problems raised by peer-to-peer networks allowing an extremely rapid diffusion of copyrighted material all around the world can be transformed into a great opportunity for enhancing the market of the products, and, in parallel, reducing the costs of distribution. Of course, it remains to be evaluated if these increased business opportunities are enough to compensate the reduction of income that a levies-based model causes on a single recording.

A further requirement of the proposed model is that recording devices must always be able to communicate the recording information to the monitoring organizations. However, with the diffusion of home DSL, of wireless communications, and (in general) of "always on" systems, this does not seem to be a major limitation. It is also worth highlighting that some privacy issues could emerge, even if, at least in principle, the system would require that only the information related to the multimedia document is transmitted to the monitoring agencies, and no private information regarding the document user is required.



▲ *9. The persistently associated UI could be exploited for monitoring documents' recording operations, thus allowing a more efficient sharing of revenues. The numbers indicate the sequence of the operations: it is worth noting that, in this case, differently from Figure 8, the TTP (here the monitoring agency) does not need to be contacted before the document is used.*

The investigation we carried out about the advantages and limitations of this alternative business model is certainly not (and did not want to be) exhaustive or able to grant that such a system could be really successful. Our aim was only to demonstrate that alternative models can be devised, at least from a technological point of view, and that data hiding technology can be helpful. Further investigations are surely needed, and we deem that other models (more effective and easier to implement) will emerge.

Conclusions

In this article we analyzed the possible use of data hiding technology in DRM systems.We gave a brief survey of the main characteristics of the most common data hiding methods. We then investigated the different approaches whereby data hiding tools can be used for DRM, by highlighting the critical points of each approach, in particular from the point of view of hostile attacks. What has emerged is that, generally, current data hiding technologies present security weaknesses making their actual use in DRM systems difficult. Consequently, we suggested that the more reliable features of data hiding technology can be successfully exploited for the profitable distribution of multimedia documents in the framework of alternative business models.

Finally, it is worth mentioning that data hiding is still a young research field and important advances are made every day. Thus, it is possible that the security problems that are still open today will be solved in the future. On the other hand, DRM needs an urgent solution and hence reasonable compromises are expected.

Mauro Barni graduated in electronic engineering at the University of Florence in 1991. He received the Ph.D. in informatics and telecommunications in 1995. From 1991 through 1998 he was with the Department of Electronic Engineering, University of Florence, Italy, where he was a postdoc researcher. Since 1998, he has been with the Department of Information Engineering, of the University of Siena, Italy, where he is an associate professor. His main interests are in the field of digital image processing and computer vision. He is author/coauthor of more than 120 papers. He is a Member of the IEEE, where he serves as member of the Multimedia Signal Processing Technical Committee (MMSP-TC).

Franco Bartolini graduated (cum laude) in electronic engineering from the University of Florence, Italy, in 1991. In 1996, he received the Ph.D. degree in informatics and telecommunications from the University of Florence. Since 2001 he has been an assistant professor at the University of Florence. His research interests include digital image sequence processing, still and moving image compression, nonlinear filtering techniques, image protection and authentication (watermarking), image processing applications for the cultural heritage field; signal compression by neural networks, and secure communication protocols. He has published more than 130 papers. He holds one European and three Italian patents in the field of digital watermarking. He is a member of the Program Committee of the SPIE/IST Workshop on Security, Steganography, and Watermarking of Multimedia Contents, and technical program cochair for IEEE MMSP Workshop 2004. He is a member of IEEE, SPIE and IAPR.

References

[1] J.A. Bloom, I.J. Cox, T. Kalker, J.-P. Linnartz, M.L. Miller, and C.B.S. Traw, "Copy protection for DVD video," *Proc. IEEE,* vol. 87, no. 7, pp. 1267–1276, Jul. 1999.

[2] S. Craver and J.P. Stern, "Lessons learned from SDMI," in *Proc. IEEE Work. Multimedia Signal Processing, MMSP'01,* Cannes, France, Oct. 2001, pp. 213–218.

[3] I.J. Cox, M.L. Miller, and J.A. Bloom, *Digital Watermarking.* San Mateo, CA: Morgan Kaufmann, 2001.

[4] I.J. Cox, M.L. Miller, and A.L. McKellips, "Watermarking as communications with side information," *Proc. IEEE,* vol. 87, no. 7, pp. 1127–1141, Jul. 1999.

[5] J. Chou, S.S. Pradhan, and K. Ramchandran, "On the duality between distributed source coding and data hiding," in *Proc. 33rd Asilomar Conf. Signals, Systems, and Computers,* vol. II, Pacific Grove, CA, USA, Oct. 1999, pp. 1503–1507.

[6] B. Chen and G. Wornell, "Quantization index modulation: a class of provably good methods for digital watermarking and information embedding," *IEEE Trans. Inform. Theory,* vol. 47, no. 4, pp. 1423–1443, May 2001.

[7] A.S. Cohen and A. Lapidoth, "The Gaussian watermarking game," *IEEE Trans. Inform. Theory,* vol. 48, no. 6, pp. 1639–1667, Jun. 2002.

[8] J.J. Eggers and B. Girod, *Informed Watermarking.* Norwell, MA: Kluwer, 2002.

[9] M.H.M. Costa, "Writing on dirty paper," *IEEE Trans. Inform. Theory,* vol. 29, no. 3, pp. 439–441, May 1983.

[10] S.I. Gelf and M.S. Pinsker, "Coding for channel with random parameters," *Probl. Control Inform. Theory,* vol. 9, no. 1, pp. 19–31, 1980.

[11] M. Barni, F. Bartolini, and T. Furon, "A general framework for robust watermarking security," *Signal Process.,* vol. 83, no. 10, pp. 2069–2084, Oct. 2003.

[12] J.R. Hernandez, M. Amado, and F. Perez-Gonzales, "DCT-domain watermarking techniques for still images: Detector performance analysis and a new structure," *IEEE Trans. on Image Processing,* vol. 9, no. 1, pp. 55–68, Jan. 2000.

[13] M. Barni, F. Bartolini, A. De Rosa, and A. Piva, "A new decoder for the optimum recovery of non-additive watermarks," *IEEE Trans. Image Process.,* vol. 10, no. 5, pp. 755–766, May 2001.

[14] Q. Cheng and T.S. Huang, "Robust optimum detection of trasform domain multiplicative watermarks," *IEEE Trans. Signal Process.,* vol. 51, no. 4, pp. 906–924, Apr. 2003.

[15] T. Furon, I. Venturini, and P. Duhamel, "An unified approach of asymmetric watermarking schemes," in *Security and Watermarking of Multimedia Contents III, Proc. SPIE,* vol. 4314, San Jose, CA, Jan. 2001, pp. 269–279.

[16] T. Furon, I. Venturini, and P. Duhamel, "Resolving rightful ownership with invisible watermarking techniques: limitations, attacks and implications," *IEEE J. Select. Areas Commun.,* vol. 4, no. 16, pp. 573–586, May 1998.

[17] M. Wu, S.A. Craver, E.W. Felten, and B. Liu, "Analysis of attacks on SDMI audio watermarks," in *Proc. IEEE Int. Conf. Acoustic Speech and Signal Processing, ICASSP'01,* vol. III, Salt Lake City, UT, USA, May 2001, pp. 1369–1372.

[18] J. Haitsma, T. Kalker, and J. Oostveen, "Robust audio hashing for content identifaction," in *2nd International Workshop Content Based Multimedia Indexing, CBMI 2001,* Brescia, Italy, Sept. 2001, pp. 19–21.

[19] H. Neuschmied, H. Mayer, and E. Balle, "Content-based identification of audio titles on the internet," in *Proc. 21st Int. Conf. WEB Delivering of Music,* Firenze, Italy, Nov. 2001, pp. 96–100.

[20] J. Haitsma and T. Kalker, "A watermarking scheme for digital cinema," in *Proc. 8th IEEE Int. Conf. Image Processing, ICIP'01,* vol. II, Thessaloniki, Greece, Oct. 2001, pp. 487–489.

[21] F. Bartolini, A. Piva, and M. Barni, "Managing copyright in open networks," *IEEE Internet Comput.,* vol. 6, no. 3, pp. 18–26, 2002.

[22] Z. Wang, M. Wu, H. Zhao, K. Liu, and W. Trappe, "Resistance of orthogonal gaussian fingerprints to collusion attacks," in *Proc. IEEE Int. Conf. Acoustic Speech and Signal Processing, ICASSP'03,* Hong-Kong, Apr. 2003, pp. 724–727.

[23] W. Trappe, M. Wu, and K.J.R. Liu, "Joint coding and embedding for collusion-resistant fingerprinting," in *Proc. XI Europ. Signal Processing Conf., EUSIPCO'02,* Toulouse, France, Sept. 2002.

[24] W. Trappe, M. Wu, Z.J. Wang, and K. Liu, "Anti-collusion fingerprinting for multimedia," *IEEE Trans. Signal Process.,* vol. 41, no. 4, pp. 1069–1087, Apr. 2003.

[25] N. Memon and P.W. Wong, "A buyer-seller watermarking protocol," *IEEE Trans. Image Process.,* vol. 10, no. 4, pp. 643–649, Apr. 2001.

[26] M.F. Mansour and A.H. Tewfik, "Secure detection of public watermarks with fractal decision boundary," in *Proc. XI Europ. Signal Processing. Conf., EUSIPCO'02,* Toulouse, France, Sept. 2002.

[27] M.L. Miller, "Is asymmetric watermarking necessary or sufficient?" in *Proc. XI Europ. Signal Processing Conf., EUSIPCO'02,* vol. I, Toulose, France, Sept. 2002, pp. 291–294.

[28] The Content ID Forum, *cIDf Specification v. 1.1,* Tokyo, Japan, Sept. 2002.

[29] *N5229—Requirements for the Persistent Association of Identification and Description of Digital Items,* ISO/IEC, JTC1/SC29/WG11, 2002 (MPEG-21—Requirements).

[30] *FCD21000-5—Information technology—Multimedia framework—Part 5: Rights Expression Language,* ISO/IEC JTC1/SC2/WG11, 2002 (MPEG21).

[31] S. Voloshynovskiy, S. Pereira, V. Iquise, and T. Pun, "Attack modelling: Towards a second generation watermarking benchmark," *Signal Process.,* vol. 81, no. 6, pp. 1177–1214, Jun. 2001.

[32] J. Dittmann, M. Steinebach, P. Wohlmacher, and R. Ackermann, "Digital watermarking enabling e-commerce strategies: conditional and user specific access to services and resources," *EURASIP Jo. Applied Signal Proces.,* vol. 2002, no. 2, pp. 174–184, Feb. 2002.

[33] T. Kalker, G. Depovere, J. Haitsma, and M. Maes, "A video watermarking system for broadcast monitoring," in *Security and Watermarking of Multimedia Contents, Proc. SPIE,* vol. 3657, San Jose, CA, Jan. 1999, pp. 103–112.

[34] P. Biddle, P. England, M. Peinado, and B. Willman, "The darknet and the future of content distribution," in *2002 ACM Workshop on Digital Rights Management, DRM 2002,* (Lecture Notes in Computer Science, vol. 2696) Washington, DC, USA: Springer Verlag, Nov. 2002.