Some General Methods for Tampering with Watermarks

Ingemar J. Cox, *Senior Member, IEEE*, and Jean-Paul M. G. Linnartz, *Member, IEEE*

***Abstract*—Watermarks allow embedded signals to be extracted from audio and video content for a variety of purposes. One application is for copyright control, where it is envisaged that digital video recorders will not permit the recording of content that is watermarked as "never copy." In such a scenario, it is important that the watermark survive both normal signal transformations and attempts to remove the watermark so that an illegal copy can be made. In this paper, we discuss to what extent a watermark can be resistant to tampering and describe a variety of possible attacks.**

I. Introduction

THE DIGITAL distribution of copyrighted content is attractive to content owners. However, the possibility of making an unlimited number of perfect digital copies is a serious concern. While it is acknowledged that professional piracy is unlikely to be prevented by technological means alone, it is hoped that the illegal casual copying that occurs in the home can be prevented by a combination of encryption and watermarking. For example, copyrighted video content intended for the digital versatile disk (DVD) will be scrambled before being placed on a disk, much like premium channels for cable TV. However, after descrambling, the content is unprotected, which is why a watermark or embedded signal will also be placed in the content. Digital video players will look for watermarks in copyrighted material and prevent playback if a "never copy" watermark is detected in material whose source is known to be a recordable disk. Similarly, digital video recorders will not record material if a "never copy" watermark is detected.

The above example is an oversimplification of the copyright protection system being designed for DVD. Nevertheless, it serves to illustrate an application in which millions of digital video players must be capable of reading signals embedded in the video content. In such a scenario, it is imperative that the watermark survive common video signal transformations, especially MPEG-2 compression and recompression and analog-to-digital and digital-to-analog conversions, since copies of content originally stored in compressed form on a DVD disc might subsequently be copied onto an analog VHS tape before being redigitized and recompressed by a writable DVD recorder. Just as importantly, it should not be trivial for an average user to circumvent the copy protection system, by for example, removing the watermark.

This paper discusses the susceptibility of watermarking algorithms to tampering. We assume that the reader is aware of typical watermark methods (e.g., [2]–[7], [13]). A comprehensive review is included in [9] Section II describes what is meant by an "unrestricted-key" watermark and Section III outlines how a public watermark will be used for copy control of DVD disks. In Section IV, we introduce some notation. In Section V, we describe how signal processing affects the detectability of the watermark. In Section VI, we then describe a series of attacks that may be used to remove a watermark.

II. Restricted and Unrestricted-Key Watermarking

The requirements for watermarking differ between applications. An important distinguishing characteristic is the level of restriction placed on the ability to read a watermark. For example, in many cases, it is desirable to embed information in audio, image, or video content such that this information is readable by many receivers. For instance, in an application such as transferring copyright ownership information by watermarking news photographs, any and all receiving users should be capable of reading the embedded information.

In the past [8], we have described such systems as "public" watermarks, drawing analogy with public key cryptography. However, this is misleading. All currently known watermarking algorithms fall into the category of secret key cryptographic algorithms and their functionality depends of the restrictions placed on the watermark key. Thus, we prefer to describe watermarks in which the key is available to a very large number of detectors as "unrestricted-key" watermarks. To the best of our knowledge, no equivalent to public key encryption is currently available for watermarking.

The key itself may simply be a pseudorandom number sequence that is embedded in all images or might be some parameters of the original unwatermarked image, such that a different key is used for each different image.

If security is an utmost concern, a content owner may desire to restrict access to information about the key, i.e., the watermark is only readable from a limited number of trusted receivers that share that secret. Or, a content owner may wish to ensure that the embedded information is most resistant to tampering. In these circumstances, one can use a restricted-key embedding method. Knowledge of this secret key is needed to embed the watermark and also to detect the watermark. It is common for such a secret key to include information about the original unwatermarked image. This can

Manuscript received August, 1997; revised October, 1997. This paper was presented in part at the IEEE International Conference on Image Processing in 1997 and appeared in part in the CD-ROM version of the same conference.

I. J. Cox is with NEC Research Institute, Princeton, NJ 08540 USA.

J.-P. M. G. Linnartz is with Natuurkundig Laboratorium, WY8, Philips Research, 5656 AA Eindhoven, The Netherlands.

Publisher Item Identifier S 0733-8716(98)01448-6.

make detection significantly more robust and consequently, the watermark becomes much more difficult for a pirate to remove. An hypothetical example of restricted-key watermarking is in the recording industry, that might choose to use watermarks to automatically monitor and log the music that is broadcast. This facilitates the transfer of airplay royalties to the music industry. In a scenario where monitoring receivers are located "in the field," the watermark embedding system as well as any and all receiving monitors can be owned and operated by the royalty collection agency.

A similar scenario can be used for a service in which images are watermarked and search robots scan the Internet to find illegally posted copies of these images. In this scenario it is not a fundamental problem that the watermark detector contains sensitive secret data, i.e., a detection key, that would reveal how the watermark can be erased. Potential attackers do not, in principle, have access to a watermark detector. However, a security threat occurs if a detector may accidentally fall into the hands of a malicious user.

Different applications require different levels of robustness and security or tamper resistance. For example, the radio station application only requires that the watermark be detectable after the signal distortions caused by the normal radio transmission process, i.e., it does not need to be tamper resistant. After all, even if the radio station were able to remove these marks, they cannot do it often without being detected by random checks, because these transmissions are public. However, for the DVD "never copy" application, the pirated content may be kept private, so no such outside auditing is possible. Hence a much greater security and resistance to tampering is desirable.

Copy protection applications require that a watermark can be read by anyone, even by potential copyright pirates, but nonetheless only the sender should be able to embed and erase the watermark. An unrestricted-key watermarking is thus preferred, though other solutions are possible. For example, a restricted-key algorithm placed in a tamper-resistant box can be used. However, this approach has weaknesses and other disadvantages. An attacker may be able to reverse engineer the tamper-resistant box. For the consumer electronics and computer industry, the logistics of the manufacturing process are more complicated and less flexible if secret data has to be handled during design, prototyping, testing, debugging, and quality control. Some of the attacks to be described in Section VI exploit the very problem that algorithms which are inherently "secret key" in nature, are used in an environment where public detection properties are desired, i.e., access to the key is almost completely unrestricted.

III. Usage of Unrestricted-Key Watermarks for DVD Video Copy Protection

For consumers, the image quality of digital video disks provides a significant improvement over the quality of existing home video equipment, such as VHS recorders. However, for content providers, there is a greater risk of illegal copying, since each DVD copy is a perfect digital reproduction. The from being made. Failing that, the aim is to reduce the value of illegal copies, either by reducing their quality (hopefully to the point of being unwatchable) or by restricting their use.

Copy protection in DVD is supported by three means. First, the video material is encrypted. Thus, a digital copy of the encrypted disk will not play on compliant DVD players. This is because the disk key will not match. Clearly, encryption is very useful, but the key is less than 40 bits, in order to avoid export control restrictions. Another possible weakness is that it is important that the playback system cannot be circumvented. This is easier to achieve in a consumer electronic device that is a closed black box, but potentially significantly more difficult for personal computers.

If the video material remains encrypted, then there is no need for watermarking. However, there are several ways that a copyrighted and encrypted video might be copied as an unencrypted, in-the-clear video, thereby losing the copy protection afforded through scrambling. Ignoring the issue of compliance, which is dealt with in Section VI-D, in-the-clear copies of encrypted material are most likely to occur through subsequent recording of the descrambled video. The in-the-clear video signal is available at a variety of sources. In the analog domain it is present in the NTSC signal and/or the RGB signal. And in the very near future, uncompressed digital video is likely to be available over the IEEE 1394 "Firewire" serial interface.

To prevent analog copying, DVD players are equipped with an analog protection system (APS). This is a proprietary technology that modifies the generated NTSC signal such that most VHS video recorders cannot record a high-quality copy despite the fact that the same signal does not affect the TV display. Unfortunately, this system does not protect RGB signals, which are common to PC's, from analog recording and is therefore easily circumventable. Thus, some percentage of copyrighted video material will find its way into the analog domain.

The most likely source of a high-quality digital copy is through the digitization of this analog copy. Neither encryption nor the APS signaling prevent playback or recording of this illegal copy. The third line of defense is a watermark that is inserted into the video sequence. This watermark is intended to survive MPEG-2 compression and digital-to-analog-to-digital conversions, i.e., if the video fidelity remains high, then the watermark should remain detectable.

A watermark in the video data can be used to prevent illegal copying by telling a compliant device not to copy it. It can also reduce the value of illegal copies by preventing them from being played on compliant devices. This means that consumers will have a choice between: 1) compliant devices, that can play legal, store-bought disks that were encrypted, but cannot play pirated disks and 2) noncompliant devices, that can play pirated material, but cannot play encrypted disks.

For the DVD application, the watermark inserted into a piece of video must describe the restrictions on that video's usage. Toward this end, the Copy Protection Technical Working Group (CPTWG) of the DVD consortium has proposed that the watermark encode the following four bits shown in

TABLE I

Bits 1-2	Instructions for usage of the analog protection system (APS)
Value	Meaning
00	Don't use APS
01	Use type 1 APS
10	Use type 2 APS
11	Use type 3 APS
Bits 3-4	Copy generation management system (CGMS)
Value	Meaning
00	Video may be copied freely
01	not used
10	Video may be copied once
11	Video may never be copied

The copy generation management system (CGMS) is intended to support one generation of copying, i.e., in some circumstances, users will be able to make a digital copy, but the system should prevent copies of this copy (or subsequent copies) being made. There is no limit to the number of one generation copies that can be made. In order to implement the "copy once" functionality of CGMS, it will probably be necessary to have one or more additional bits in the watermark that can be easily changed by consumer DVD devices.

A. Technical Requirements

The requirements placed on watermarking algorithms for the above application differ from those for other applications that are currently in the market, such as identification of ownership. The application of watermarking for copy protection requires a low bit rate and allows the use of many frames for watermark detection. However, since watermark detectors must be built into millions of low-cost, consumer devices, and since these detectors must work at video rates, there is a very strong requirement that the detector be extremely simple and cheap. Furthermore, since the DVD standard employs MPEG coding, the watermarking method must work well with MPEG. These last two requirements are challenging design specifications. The requirements for the APS bits are:

1) detectable in the compressed and baseband video;
2) detector should be very inexpensive both in terms of gate count (hardware) or MIPS (software);
3) no visible artifacts, i.e., very high image fidelity;
4) tamper resistant, i.e., it should not be easily circumvented or removed;
5) watermark should survive color representation conversion from YUV to RGB;
6) data rate of 2 bits per frame;
7) permanent—the APS bits do not need to be altered.

The requirements for the playback control and copy generation system are:

1) detectable in the baseband and/or compressed video;
2) 2–5 of APS requirements;
3) data rate low (e.g., 2 bits per 100 frames);
4) field encodable for generation control—multiple watermarks using possibly different methods though detection circuitry should preferable be the same;

Both systems should also survive:

1) compression;
2) decompression;
3) digital-to-Analog;
4) analog-to-Digital;
5) standards conversion, e.g., analog video recorder (VHS), the European broadcast standard PAL, the French broadcast standard SECAM;
6) time dilation—changes in frames rate.

IV. Formulation of a Model

Mathematically, given an image I and a watermark W, the watermarked image, I', is formed by $I' = I + f(I, W)$ such that $|I - I'| < JND$ where $|I - I'|$ denotes the perceptual difference, and JND refers to just noticeable difference, i.e., the watermarked image is constrained to be visually identical (or very similar) to the original unwatermarked image.

In theory, the function f may be arbitrary, but in practice robustness requirements pose constraints on how f can be chosen. One requirement is that watermarking has to be robust to random noise addition. Therefore many watermark designers opt for a scheme in which image I will result in approximately the same watermark as a slightly altered image $I+\epsilon$ with $|\epsilon| < JND$. In such cases $f(I, W) \approx f(I+\epsilon, W)$.

For an unrestricted-key watermark, detection of the watermark, W, is typically achieved by correlating the watermark with some function, g, of the watermarked image.

Example: In its basic form, in one half of the pixels the luminance is increased by one unit step while the luminance is kept constant [3] or decreased by one unit step [2] in the other half. Detection by summing luminances in the first subset and subtracting the sum of luminances in the latter subset is a special case of a correlator. One can describe this as $I' = I + W$, with $W \in R^N$, and where $f(I, W) = W$. The detector computes $I' \cdot W$, where $\cdot$ denotes the scalar product of two vectors.

If W is chosen at random, then the distribution of $I \cdot W$ will tend to be quite small, as the random $\pm$ terms will tend to cancel themselves out, leaving only a residual variance. However, in computing $W \cdot W$ all of the terms are positive, and will thus add up. For this reason, the product $I' \cdot W = I \cdot W + W \cdot W$ will be close to $W \cdot W$. In particular, for sufficiently large images, it will be large, even if the magnitude of I is much larger than the magnitude of W. It turns out that the probability of making an incorrect detection can be expressed as the complementary error function of the square root of the ratio $W \cdot W$ over the variance in pixel luminance values. This result is very similar to expressions commonly encountered in digital transmission over noisy radio channels. A derivation is outside the scope of this paper, so we refer the interested reader to [12] for a detailed evaluation of the statistical behavior of $I \cdot W$ and $W \cdot W$.

V. Signal Transformations

The above specification may not seem difficult since it only requires the embedding of 4 bits of information in a data

the total video data is approximately $720 \times 480 \times 30 \times 10$. This is over 100 Mbytes prior to MPEG compression. However, the constraints of 1) maintaining image fidelity, and 2) surviving common signal transformations, can be severe. In particular, many signal transformations cannot be modeled by a simple linear additive noise process. Instead, such processes are highly spatially correlated and may interact with the watermark in complex ways.

There are a number of common signal transformations that a watermark should survive, e.g., affine transformations, compression/recompression, and noise. In some circumstances, it may be possible to design a watermark that is completely invariant to a particular transformation. For example, this is usually the case for translational motions. However, scale changes are often much more difficult to design for and it may be the case that a watermark algorithm is only robust to small perturbations in scale. In this case, a series of attacks may be mounted by identifying the limits of a particular watermarking scheme and subsequently finding a transformation that is outside of these limits yet maintains adequate image fidelity.

A. Attacks by Affine Transformations

Shifts over a few pixels can cause watermarking detectors to miss the presence of watermark. The problem can be illustrated by our example watermarking scheme. Suppose one shifts I' by one pixel, obtaining I'_S. Let I_S and W_S denote the similarly shifted versions of I and W. Then $I'_S \cdot W = I_S \cdot W + W_S \cdot W$. As before, the random $+/-$ terms in $I_S \cdot W$ will tend to cancel themselves out. However, the $W_S \cdot W$ terms will also cancel themselves out, if each $+/-$ value was chosen independently. Hence, $I'_S \cdot W$ will have small magnitude and the watermark will not be detected.

Typical analog VHS recorders cause shifting over a small portion of a line, but enough to cause a shift of several pixels or even a few DCT blocks. Recorder time jitter and tape wear are a significant cause of stretching of a image. Even if the effects are not disturbing to a viewer, it may completely change the alignment of the watermark with respect to pixels and DCT block boundaries.

There are a number of defenses against such attacks. Ideally, one would like to reverse the affine transformations. Given an original, a reasonable approximation to the distortion can be computed. With unrestricted-key watermarks, and in particular the "do not copy" application, no original is available. A secondary signal, i.e., a registration pattern, may be inserted into the image whose entire purpose is to assist in reversing the transformation. However, one can base attacks on this secondary signal, removing or altering it in order to block detection of the watermark. The mark components can be positioned at key visual features of the image, e.g., in patches whose average luminosity is at a local maximum. Finally, one can insert the mark into features that are transformation invariant. For example, the magnitudes of Fourier coefficients are translation invariant.

In some applications, it may be assumed that the extent of

B. Attacks by Noise Addition

A common misunderstanding is that a watermark of small amplitude can be removed by adding random noise of a similar amplitude. On the contrary, correlation detectors appear very robust against addition of a random noise term ϵ. For instance if $f(I, W) = W$ one can describe the attacked image as $I' = I + \epsilon + W$. The detector computes $I' \cdot W$. The product $I' \cdot W = I \cdot W + \epsilon \cdot W + W \cdot W$. If the watermark was designed with $W \cdot W$ largely exceeding the statistical spreading in $I \cdot W$, it will mostly also largely exceed the statistical spreading in $\epsilon \cdot W$. In practice, noise mostly is not a serious threat unless (in the frequency components of relevance) the noise is large compared to image I or if the noise is correlated with the watermark.

C. Attacks by Digital Compression: Future Digital Recorders

Digital recorders may not always make a bit exact copy. Digital recorders will, at least initially, not contain sophisticated signal processing facilities. For recording of MPEG streams onto media with limited storage capacity, the recorder may have to reduce the bit rate of the content.

This will particularly be the case for high-quality high-rate source video such as high-definition broadcasts. A commonly adopted method is to more coarsely quantize the high (spatial) frequency components in the digital representation of the frames. Since the file header structure and motion estimation can be retained, this method is substantially cheaper in implementation than to completely redo the compression, including computationally intense motion estimation. However, this form of transcoding can affect the detectability of the watermark, particularly if the significant portions of the watermark are contained in high frequencies.

For video recorders that redo compression, image quality usually degrades significantly. Usually alignment of independently coded I-frames between original and copy is important. If complete recompression occurs, quantization noise is present, typically with large high-frequency components. Moreover, at high frequencies, image and watermark components may be lost. In such cases, the watermark may be lost, though it may be that the video quality is significantly degraded.

VI. Intentional Attacks

In this section, we describe a series of attacks that can be mounted against a unrestricted-key watermark.

A. Exploiting the Presence of a Watermark Detector Device

An attacker may not have precise knowledge of the watermark. Nevertheless, he usually has access to a detector and the detector provides information about whether a certain piece of content contains a watermark or not. This information can be used to remove the watermark. This model may be particularly appropriate in copy control applications, such as for DVD. The watermark detection and consequent playback restrictions require a uniform standard to be adhered to across all brands of

a watermark pattern that reliably triggers such detectors can be chosen by the content owner according to his requirements for robustness and perceptivity. Many different patterns may all have the same effect on a standard watermark detector. An attacker may not wish to remove the very watermark that the content owner has embedded, which may have been adapted according to a particular perceptual model. He only desires to extract a pattern that cancels the effect that the present watermark has on the detector.

The aim of the attack is to experimentally deduce the behavior of the detector, and to exploit this knowledge to ensure that a particular image does not trigger the detector. For example, if the watermark detector gives a soft decision, e.g., a continuous reliability indication when detecting a watermark, the attacker can learn how minor changes to the image influence the strength of the detected watermark. That is, modifying the image pixel by pixel, he can deduce the entire correlation function or other watermark detection rule.

Interestingly, such attack can also be applied even when the detector only reveals a binary decision, i.e., present or absent. Basically the attack examines an image that is at the boundary where the detector changes its decision from "absent" to "present." For clarity the reader may consider a watermark detector of the correlator type; but this is not a necessary condition for the attack to work. For a correlator type of detector, our attack reveals the correlation coefficients used in the detector (or at least their sign) as in the following examples.

1) Starting with a watermarked image, the attacker creates a test image that is near the boundary of a watermark being detectable. At this point it does not matter whether the resulting image resembles the original or not. The only criterion is that minor modifications to the test image cause the detector to respond with "watermark" or "no watermark" with a probability that is sufficiently different from zero or one. The attacker can create the test image by modifying a watermarked image step-by-step until the detector responds "no watermark found." A variety of modifications are possible. One method is to gradually reduce the contrast in the image just enough to drop below the threshold where the detector reports the presence of the watermark. An alternative method is to replace more and more pixels in the image by neutral grey. There must be a point where the detector makes the transition from detecting a watermark to responding that the image contains no watermark. Otherwise this step would eventually result in an evenly grey colored image, and no reasonable watermark detector can claim that such image contains a watermark.
2) The attacker now increases or decreases the luminance of a particular pixel until the detector sees the watermark again. This provides the insight of whether the watermark embedder decreases or increases the luminance of that pixel.
3) This step is repeated for every pixel in the image.
4) Combining the knowledge on how sensitive the detector is to a modification of each pixel, the attacker estimates a combination of pixel values that has the largest influence
5) The attacker uses the original marked image and subtracts (λ times) the estimate, such that the detector reports that no watermark is present. λ is found experimentally, such that λ is as small as possible. Moreover, the attacker may also exploit a perceptual model to minimize the visual effect of his modifications to the image.

Our main argument here is that the effort needed to find the watermark is much less than commonly believed. If an image contains N pixels, conventional wisdom is that an attack that searches the watermark requires an exponential number of attempts of order $O(2^N)$. A brute force exhaustive search checking all combinations with positive and negative sign of the watermark in each pixel results in precisely 2^N attempts. The above method shows that most known watermarking methods can be broken much faster, namely in $O(N)$, provided a device is available that outputs a binary (present or absent) decision as to the presence of the watermark.

B. Attacks Based on the Presence of a Watermark Inserter

If the attacker has access to a watermark inserter, this provides further opportunities to break the security. Attacks of this kind are relevant to DVD copy control in which copy generation management is required, i.e., the user is permitted to make a copy from the original source disc but is not permitted to make a copy of the copied material—only one generation of copying is allowed. The recorder should change the watermark status from "one copy allowed" to "no more copies allowed." The attacker has access to the content before and after this marking. That is, he can create a difference image, by subtracting the unmarked original from the marked content. This difference image is equal to $f(I,W)$. An obvious attack is to predistort the original to undo the mark addition in the embedder. That is, the attacker computes $I - f(I,W)$ and hopes that after embedding of the watermark, the recorder stores

$$I - f(I,W) + f(I - f(I,W),W)$$

which is likely to approximate I. The reason why most watermarking methods are vulnerable to this attack is that watermarking has to be robust to random noise addition. If, for reasons discussed before

$$f(I,W) \approx f(I+\epsilon,W)$$

and because watermarks are small modifications themselves, $f(I,W) \approx f(I - f(I,W),W)$. This property enables the above predistortion attack.

C. Attacks by Statistical Averaging

An attacker may try to estimate the watermark and subtract this from a marked image. Such an attack is particularly dangerous if the attacker can find a generic watermark, for instance, one with $u = f(I,W)$ not depending significantly on the image I. Such an estimate u of the watermark can then be used to remove a watermark from any arbitrary marked image, without any further effort for each new image or frame

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.