Published in Proc. IEEE Int. Conf. on Image Processing (1997) - available only on cdrom.

Public watermarks and resistance to tampering

Ingemar J. Cox

NEC Research Institute
4 Independence Way
Princeton, NJ 08540

Jean-Paul M.G. Linnartz

Natuurkundig Laboratorium
WY8, Philips Research
5656 AA Eindhoven, The Netherlands

Abstract

Public 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 public watermark can be resistant to tampering and describe a variety of possible attacks.

1 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, for the digital versatile disk (DVD), copyrighted video content 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 over simplification 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 re-compression 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 re-digitized and re-compressed 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.

The requirements for watermarking can differ between applications. 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. We describe such systems as "public" watermarking procedures. The embedding algorithm is private. i.e., only known to copyright owners, whereas the detection algorithm is public knowledge. There may be a fundamental asymmetry in the embedding and detection function, such that it may be computationally infeasible to derive one from the other. While a similar asymmetric concept exists in cryptography [1], it is not sure whether secure public watermarking can theoretically exist.

This paper discusses the susceptibility of public watermarking algorithms to tampering. We assume that the reader is aware typical watermark methods (e.g. [2, 3, 4, 5, 6, 7]) and of the most basic attempts to remove the watermark, such as noise addition, filtering, shifting stretching and rotating the image, etc., as we do not cover these. In Section 2 we introduce some notation. In Section 3 we then describe a series of attacks that may be used to remove a watermark.

2 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 choosen. 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 a public 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.

3 Intentional attacks

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

3.1 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.

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 watermark.

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 example:

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 on the detector for the least disturbance of the image.
5. The attacker uses the original marked image and subtracts (λ times) the estimate, such that the detector reports that no waterwark is present. (λ is found experimentally, such that λ is as small as possible.

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.

3.2 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 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 pre-distort 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 pre-distortion attack.

3.3 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 to be "cleaned".

The attacker may separate the watermark u by adding or averaging multiple images, e.g. multiple successive marked images $I_1 + u, I_2 + u, ...I_N + u$ from a video sequence. The addition of N such images results in $Nu + \sum_i I_i$, which tends to Nu for large N and sufficiently many and sufficiently independent images $I_1, I_2, ..., I_N$.

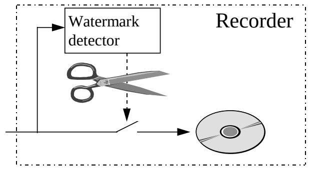

Figure 1: An attacker could modify his recorder, such that it does not check for watermarks.

A countermeasure is to use at least two different watermarks u_1 and u_2 at random, say with probability p_1 and p_2 where $p_2 = 1 - p_1$, respectively. The above attack then only produces $p_1 u_1 + (1 - p_1) u_2$, without revealing u_1 or u_2. However a refinement of the attack is to compute weighted averages, where the weight factor is determined by a (possibly unreliable but better than random) guess of whether a particular image contains one watermark or the other.

3.4 Attacks on the Copy Control Mechanism

A pirate who is interested in illegal copying may not only attempt to tamper with the watermarked image, but can also attempt to circumvent the copy control mechanism while leaving the watermarked content unchanged. The most trivial attack is to tamper with the output of the watermark detector and modify it in such a way that the copy control mechanism always sees a "no watermark" detection, even if a watermark is present in the content. Since hackers and pirates more easily can modify (their own!) recorders but not their customers' players, playback control is a mechanism that detects watermarks during the playback of discs. The resulting tape or disc can be recoginzed as an illegal copy if playback control is used.

Copy protection based on watermarking content has a further fundamental weakness. The watermark detection process is designed to detect the watermark when the video is perceptually meaningful. Thus, a user may apply a weak form of scrambling to copy protected video, e.g. inverting the pixel intensities, prior to recording. The scrambled video is unwatchable and the recorder will fail to detect a watermark and consequently allow a copy to be made. Of course, on playback, the video signal will be scrambled, but the user may then simple invert or descramble the video in order to watch a perfect and illegal copy of a video. Simple scrambling and descrambling hardware would

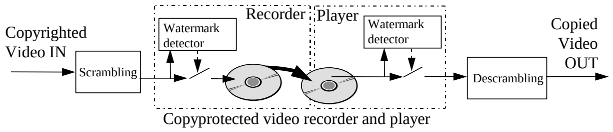

Figure 2: Scrambling as a means to defeat watermark detection.

be very inexpensive and manufacturers might argue that the devices serve a legitimate purpose in protecting a user's personal video. Similarly, digital MPEG can easily be converted into a file of seemingly random bits. One way to avoid such circumvention for digital recording is to only allow the recording of content in a recognized file format. Of course this would severely limit the functionality of the storage device.

Moreover, it does not make sense because a more subtle circumvention of the copy control mechanism can be used. This method exploits the technique of data hiding to bypass the watermark detector in the recorder. The method of attack is similar to a technique used in countries where the private use of cryptographic encryption is outlawed [9]. The copyrighted work is hidden in an innocent-looking file of a known recognized format. For instance the digital MPEG video representation allows the user to embed additional user data or stuff bit without any significant limitation. Stuff bits may be misused by a pirate to embed a complete MPEG video film. During playback, the playback platform, e.g. the PC must perform a few additional functions, but this does not need to cause significant performance problems.

4 Conclusions

In this short paper we summarized a series of attacks that are all independent of the underlying algorithm used for watermarking. In addition, there are numerous other attacks that can be made to specific classes of algorithms. For example, in many watermarking schemes for video and images, a registration pattern is embedded in the image to provide tolerance to geometric distortions. When a registration pattern is used, this is often the Archille's heel of such a scheme, i.e. if correct registration can be prevented, then watermark detection will fail.

Legal, economic and technological efforts are all needed to prevent and/or deter piracy. Public watermarking is a promising technology but one that cannot be absolutely secure. Nevertheless, we believe it is a useful technology that both compliments the protection afforded by encryption and can be applied in the analog and the digital domains.

Acknowledgments

The authors thank Joe Kilian for many useful discussions and comments.

References

[1] R.L. Rivest, A. Shamir, and L.M. Adleman, A method for obtaing Digital Signatures and Public-Key Cryptosystems, *Communications of the ACM*, Vol. 21, No. 2, Feb. 1978, pp. 120-126.

[2] W. Bender, D. Gruhl, N. Morimoto, "Techniques for Data Hiding", Proceedings of the SPIE, 2420:40, San Jose CA, USA, February 1995

[3] I. Pitas, T. Kaskalis, "Signature Casting on Digital Images", Proceedings IEEE Workshop on Nonlinear Signal and Image Processing, Neos Marmaras, June 1995

[4] E. Koch, J. Zhao, "Towards Robust and Hidden Image Copyright Labeling". Proceedings IEEE Workshop on Nonlinear Signal and Image Processing, Neos Marmaras, June, 1995

[5] F.M. Boland, J.J.K. O Ruanaidh, C. Dautzenberg, "Watermarking Digital images for Copyright Protection", Proceedings of the 5th IEE International Conference on Image Processing and its Applications, no. 410, Edinburgh, July, 1995, pp. 326-330.

[6] J. Zhao, E. Koch, "Embedding Robust Labels into Images for Copyright Protection", Proceedings of the International Congress on Intellectual Property Rights for Specialized Information, Knowledge and New Technologies, Vienna, Austria, August 1995

[7] I. Cox, J. Kilian, T. Leighton and T. Shamoon, "A secure, robust watermark for multimedia", in Proc. Workshop on Information Hiding, Univ. of Cambridge, U.K., May 30 - June 1, 1996, pp. 175-190

[8] J.R. Smith, B. O. Comiskey, "Modulation and Information Hiding in Images", in Proc. Workshop on Information Hiding, Univ. of Cambridge, U.K., May 30 - June 1, 1996, pp. 191 - 201

[9] E. Franz, A. Jerichow, S. Moeller, A. Pfitzmann and I. Stierand, "Computer-based steganography: How it works and why therefore any restrictions on cryptography are nonsense, at best", in: Information Hiding, First Int. workshop, Cambridge, UK, May, June, 1996, Springer, Lecture Notes in Computer Science, No. 1174, pp. 7-21