David Aucsmith, Ed., *Second workshop on information hiding*, in vol. 1525 of *Lecture Notes in Computer Science*, Portland, Oregon, USA, 14–17 April, 1998, pp. 218–238.
ISBN 3-540-65386-4.

Attacks on Copyright Marking Systems

Fabien A.P. Petitcolas*, Ross J. Anderson, and Markus G. Kuhn**

University of Cambridge, Computer Laboratory
Pembroke Street, Cambridge CB2 3QG, UK
{fapp2,rja14,mgk25}@cl.cam.ac.uk
<http://www.cl.cam.ac.uk/Research/Security/>

Abstract. In the last few years, a large number of schemes have been proposed for hiding copyright marks and other information in digital pictures, video, audio and other multimedia objects. We describe some contenders that have appeared in the research literature and in the field; we then present a number of attacks that enable the information hidden by them to be removed or otherwise rendered unusable.

1 Information Hiding Applications

The last few years have seen rapidly growing interest in ways to hide information in other information. A number of factors contributed to this. Fears that copyright would be eroded by the ease with which digital media could be copied led people to study ways of embedding hidden copyright marks and serial numbers in audio and video; concern that privacy would be eroded led to work on electronic cash, anonymous remailers, digital elections and techniques for making mobile computer users harder for third parties to trace; and there remain the traditional 'military' concerns about hiding one's own traffic while making it hard for the opponent to do likewise.

The first international workshop on information hiding [3] brought these communities together and a number of hiding schemes were presented there; more have been presented elsewhere. We formed the view that useful progress in steganography and copyright marking might come from trying to attack all these first generation schemes. In the related field of cryptology, progress was iterative: cryptographic algorithms were proposed, attacks on them were found, more algorithms were proposed, and so on. Eventually, theory emerged: fast correlation attacks on stream ciphers and differential and linear attacks on block ciphers, now help us understand the strength of cryptographic algorithms in much more detail than before. Similarly, many cryptographic protocols were proposed and almost all the early candidates were broken, leading to concepts of protocol robustness and techniques for formal verification [7].

* The first author is grateful to Intel Corporation for financial support under the grant 'Robustness of Information Hiding Systems'
** The third author is supported by a European Commission Marie-Curie grant

So in this paper, we first describe the copyright protection context in which most recent schemes have been developed; we then describe a selection of these schemes and present a number of attacks, which break most of them. We finally make some remarks on the meaning of robustness in the context of steganography in general and copyright marking in particular.

1.1 Copyright Protection Issues

Digital recording media offer many new possibilities but their uptake has been hindered by widespread fears among intellectual property owners such as Hollywood and the rock music industry that their livelihoods would be threatened if users could make unlimited perfect copies of videos, music and multimedia works.

One of the first copy protection mechanisms for digital media was the serial copy management system (SCMS) introduced by Sony and Phillips for digital audio tapes in the eighties [34]. The idea was to allow consumers to make a digital audio tape of a CD they owned in order to use it (say) in their car, but not to make a tape of somebody else's tape; thus copies would be limited to first generation only. The implementation was to include a Boolean marker in the header of each audio object. Unfortunately this failed because the hardware produced by some manufacturers did not enforce it.

More recently the Digital Video Disk, also known as Digital Versatile Disk (DVD) consortium called for proposals for a copyright marking scheme to enforce serial copy management. The idea is that the DVD players sold to consumers will allow unlimited copying of home videos and time-shifted viewing of TV programmes, but cannot easily be abused for commercial piracy [21, 46]. The proposed implementation is that videos will be unmarked, or marked 'never copy', or 'copy once only'; compliant players would not record a video marked 'never copy' and when recording one marked 'copy once only' would change its mark to 'never copy'. Commercially sold videos would be marked 'never copy', while TV broadcasts and similar material would be marked 'copy once only' and home videos would be unmarked.

Electronic copyright management schemes have also been proposed by European projects such as Imprimatur and CITED [47, 68, 69], and American projects such as the proposed by the Working Group on Intellectual Property Rights [71].

1.2 Problems

Although these schemes might become predominant in areas where they can be imposed from the beginning (such as DVD and video-on-demand), they suffer from a number of drawbacks. Firstly, they rely on the tamper-resistance of consumer electronics – a notoriously unsolved problem [5]. The tamper-resistance mechanisms being built into DVD players are fairly rudimentary and the history of satellite TV piracy leads us to expect the appearance of 'rogue' players which will copy everything. Electronic copyright management schemes also conflict with applications such as digital libraries, where 'fair use' provisions are

strongly entrenched. According to Samuelson, '*Tolerating some leakage may be in the long run of interest to publishers* [...] *For educational and research works, pay-per-use schemes may deter learning and deep scholarship*' [58]. A European legal expert put it even more strongly: that copyright laws are only tolerated because they are not enforced against the large numbers of petty offenders [35].

Similar issues are debated within the software industry; some people argue, for example, that a modest level of amateur software piracy actually enhances revenue because people may 'try out' software they have 'borrowed' from a friend and then go on to buy it (or the next update).

For all these reasons, we may expect leaks in the primary copyright protection mechanisms and wish to provide independent secondary mechanisms that can be used to trace and prove ownership of digital objects. It is here that marking techniques are expected to be most important.

2 Copyright Marks

There are two basic kinds of mark: fingerprints and watermarks. One may think of a fingerprint as an embedded serial number while a watermark is an embedded copyright message. The first enables us to trace offenders, while the second can provide some of the evidence needed to prosecute them. It may also, as in the DVD proposal, form part of the primary copy management system; but it will more often provide an independent back-up to a copy management system that uses overt mechanisms such as digital signatures.

In [8], we discussed the various applications of fingerprinting and watermarking, their interaction, and some related technologies. Here, we are concerned with the robustness of the underlying mechanisms. What sort of attacks are possible on marking schemes? What sort of resources are required to remove marks completely, or to alter them so that they are read incorrectly? What sort of effect do various possible removal techniques have on the perceptual quality of the resulting audio or video?

We will use the terminology agreed at the first international workshop on Information Hiding [54]. The information to be hidden (watermark, fingerprint, or in the general case of steganography, a secret message) is *embedded* in a *cover* object (a cover CD, a cover video, a cover text, etc.) giving a *stego* object, which in the context of copyright marking we may also call a *marked* object (CD, video, etc). The embedding is performed with the help of a *key*, a secret variable that is in general known to the object's owner. Recovery of the embedded mark may or may not require a key; if it does the key may be equal to, or derived from, the key used in the embedding process.

In the rest of this section, we will first discuss simple hiding methods and the obvious attacks on them. We will then present, as an example of the 'state of the art', robustness requirements that appeared in a recent music industry request for proposals [1]. We will then present the main contending techniques used in currently published and fielded systems. Attacks on these systems will then be presented.

2.1 Simple Hiding Methods

The simplest schemes replace all the bits in one or more of the less significant bit planes of an image or audio sample with the 'hidden' information [12, 26, 39, 67]. This is particularly easy with pictures: even when the four least significant bits of the cover image are replaced with the four most significant bits of the embedded image, the eye cannot usually tell the difference [39]. Audio is slightly harder, as the randomisation of even the least significant bit of 8-bit audio adds noise that is audible during quiet passages of music or pauses in speech. Nonetheless, several systems have been proposed: they include embedding, in the regular channels of an audio CD, another sound channel [27, 70] and a steganographic system in which secret messages are hidden in the digitised speech of an ISDN telephone conversation [26].

However, bit-plane replacement signals are not only easy to detect. They violate Kerckhoffs' principle that the security of a protection system should not rely on its method of operation being unknown to the opponent, but rather on the choice of a secret key [36]. Better approaches use a key to select some subset of pixels or sound samples which then carry the mark.

An example of this approach is Chameleon [6], a system which enables a broadcaster to send a single ciphertext to a large population of users, each of which is supplied with a slightly different decryption key; the effect of this is to introduce a controlled number of least-significant-bit errors into the plaintext that each user decrypts. With uncompressed digital audio, the resulting noise is at an acceptably low level and then Chameleon has the advantage that the decrypted audio is fingerprinted automatically during decryption without any requirement that the consumer electronic device be tamper-resistant.

In general, schemes which use a key to choose some subset of least significant bits to tweak may provide acceptable levels of security in applications where the decrypted objects are unlikely to be tampered with. However, in many applications, a copyright pirate may be able and willing to perform significant filtering operations and these will destroy any watermark, fingerprint or other message hidden by simple bit tweaking. So we shall now consider what it means for a marking scheme to be robust.

2.2 Robustness Requirements

The basic problem is to embed a mark in the digital representation of an analogue object (such as a film or sound recording) in such a way that it will not reduce the perceived value of the object while being difficult for an unauthorised person to remove. A first pass at defining robustness in this context may be found in a recent request for proposals for audio marking technology from the International Federation for the Phonographic Industry, IFPI [1]. The goal of this exercise was to find a marking scheme that would generate evidence for anti-piracy operations, track the use of recordings by broadcasters and others and control copying. The IFPI robustness requirements are as follows:

– the marking mechanism should not affect the sonic quality of the sound recording;
– the marking information should be recoverable after a wide range of filtering and processing operations, including two successive D/A and A/D conversions, steady-state compression or expansion of 10%, compression techniques such as MPEG and multi-band nonlinear amplitude compression, adding additive or multiplicative noise, adding a second embedded signal using the same system, frequency response distortion of up to 15 dB as applied by bass, mid and treble controls, group delay distortions and notch filters;
– there should be no other way to remove or alter the embedded information without sufficient degradation of the sound quality as to render it unusable;
– given a signal-to-noise level of 20 dB or more, the embedded data channel should have a bandwidth of 20 bits per second, independent of the signal level and type (classical, pop, speech).

Similar requirements could be drawn up for marking still pictures, videos and multimedia objects in general. However, before rushing to do this, we will consider some systems recently proposed and show attacks on them that will significantly extend the range of distortions against which designers will have to provide defences, or greatly reduce the available bandwidth, or both.

2.3 General Techniques

We mentioned schemes that modify the least significant bits of digital media; by repeating such marks, or employing more robust encoding methods, we can counter some filtering attacks. We can also combine coding with various transform techniques (DCT, wavelet and so on).

The *Patchwork* algorithm [11], for instance, successively selects random pairs of pixels; it makes the brighter pixel brighter and the duller pixel duller and the contrast change in this pixel subset encodes one bit. To maintain reasonable robustness against filtering attacks, the bandwidth of such systems has to be limited to at most a few hundred bits per image [40, 41]. In a similar way, marks can be embedded in audio by increasing the amplitude contrast of many pairs of randomly chosen sound samples and using a suitable filter to minimise the introduction of high-frequency noise.

More sophisticated variants on this theme involve spread-spectrum techniques. Although these have been used since the mid-fifties in the military domain because of their anti-jamming and low-probability-of-intercept properties [61], their applicability to image watermarking has only been noticed recently by Tirkel *et al.* [66]. Since then a number of systems based on this technique have been proposed [67, 72, 73]: typically a maximal length sequence is added to the signal in the spatial domain and the watermark is detected by using the spatial cross-correlation of the sequence and the watermarked image.

Another kind of marking technique embeds the mark in a transform domain, typically one that is widely used by compression algorithms. Thus when marking sound one could add a pseudorandom sequence to the excitation signal in

an LPC or CELP coded audio signal [45] and when marking an image one could use the DCT domain. Langelaar *et al.* remove certain high frequency DCT coefficients [41]; Cox *et al.* modulate the 1000 largest DCT coefficients of an image with a random vector [19]; Koch *et al.* change the quantisation of the DCT coefficients and modify some of them in such a way that a certain property (order in size) is verified [37]; while Ó Ruanaidh *et al.* modulate the DCT coefficient with a bi-directional coding [49].

Techniques of this kind are fairly robust against various kinds of signal processing and may be combined with exploitation of the perceptual masking properties of the human auditory system in [16, 17] and of the human vision system in [28, 65, 64]. The basic idea here is to amplify the mark wherever the changes will be less noticeable and also to embed it in the *perceptually significant* components of the signal [20]. Masking may also be used to avoid placing marks in places such as the large expanses of pure colour found in cartoons; the colour histogram of such images has sharp peaks, which are split into twin peaks by some naïve marking methods as the colour value c is replaced by $c-\delta$ and $c+\delta$, thus allowing the mark to be identified and removed [44].

3 Attacks

This leads us to the topic of attacks and here we present some quite general kinds of attack that destroy, or at least reveal significant limitations of, several marking schemes: PictureMarc 1.51 [24, 56], SysCoP [37, 74, 75], JK_PGS (EPFL algorithm, part of the European TALISMAN project), SureSign [63], EIKONAmark [25, 55], Echo Hiding, and the NEC method [19]. We suspect that systems that use similar techniques are also vulnerable to our attacks.

3.1 The Jitter Attack

Our starting point in developing a systematic attack on marking technology was to consider audio marking schemes that tweak low order bits whose location is specified by a key. A simple and devastating attack on these schemes is to add jitter to the signal. In our first implementation, we split the signal into chunks of 500 samples, either duplicated or deleted a sample at random in each chunk (resulting in chunks of 499 or 501 samples long) and stuck the chunks back together. This turned out to be almost imperceptible after filtering, even in classical music; but the jitter prevents the marked bits from being located.

In a more sophisticated implementation, we resample these chunks at a lower or higher frequency. This relies on the properties of the ear's pitch resolution:

> In pitch perception experiments in the mid-audio frequency range, subjects are able to perceive changes in frequency of pure tones of approximately 0.1%. [...] At frequencies above 4 kHz pitch discrimination reduces substantially. [...] In the case of complex signals, such as speech, it is very much less clear what the capabilities and processes of the auditory system are. [...] There is evidence that peaks in the spectrum of

the audio signal are detected more easily than features between spectral peaks. *J.N. Holmes* [33]

If n_i is the number of samples in the ith chunk, n'_i the number of samples after resampling and α the maximum relative change of frequency allowed then, in the mid-audio range, we are roughly limited, for pure tones, by $|\Delta n_i| \leq \alpha n_i$ (because α is small), where $\Delta n_i := n'_{i+1} - n'_i$. This can be simplified as $0 < k \leq \frac{\alpha n}{2}$ when the n_i are equal and when the number k of removed or added samples is constant for each chunk. This is the approach we chose; it allowed us to introduce a long jitter. Then the strategy for choosing k and n depends on the input signal. With this technique we were able to tweak up to one sample in 50 of a 44 kHz sampled voice recording without any perceptible effect.

We also applied a similar attack to SysCoP Demo 1.0. In that case we simply deleted columns of pixels and duplicated others in order to preserve the image size. Fig. 1 gives an example of this attack.

Of course, there are much more subtle distortions that can be applied. For instance, in [30], Hamdy *et al.* present a way to increase or decrease the length of a music performance without changing the pitch; this was developed to enable radio broadcasters to slightly increase or decrease the playing time of a musical track. As such tools become widely available, attacks involving sound manipulation will become easy. Most simple spread-spectrum based techniques are subject to this kind of attacks. Indeed, although spread-spectrum signal are very robust to distortion of their amplitude and to noise addition, they do not survive timing errors: synchronisation of the chip signal is very important and simple systems fail to recover this synchronisation properly.

3.2 StirMark

Following this attack and after evaluating some watermarking software, it became clear that although many of the seriously proposed schemes could survive basic manipulations – that is, manipulations that can be done easily with standard tools, such as rotation, shearing, resampling, resizing and lossy compression – they would not cope with combinations of them. This motivated us to implement StirMark.

StirMark is a generic tool developed for simple robustness testing of image marking algorithms and other steganographic techniques. In its simplest version, StirMark simulates a resampling process, i.e. it introduces the same kind of errors into an image as printing it on a high quality printer and then scanning it again with a high quality scanner. It applies a minor geometric distortion: the image is slightly stretched, sheared, shifted and/or rotated by an unnoticeable random amount[1] (Fig. 2 – middle drawing) and then resampled using either bi-linear or

[1] If A, B, C and D are the corners of the image, a point M of the said image can be expressed as $M = \alpha[\beta A + (1-\beta)D] + (1-\alpha)[\beta B + (1-\beta)C]$ where $0 \leq \alpha, \beta \leq 1$ are the coordinates of M relatively to the corners. The distortion is done by moving the corners by a small random amount in both directions. The new coordinates of M are given by the previous formula, keeping (α, β) constant.

(a)

(b)

bash$ imageread_demo watermarked.ppm

Key:

No certificate file.

————————————————————-

A valid watermark found - estimated correction percentage is : 100

Retrieved Secret Label (string) : SysCoP(TM)

(c)

bash$ imageread_demo jitter.ppm

Key:

No certificate file.

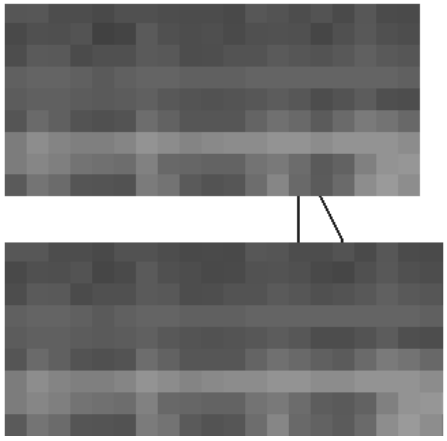
————————————————————-

Cannon find valid watermark - failed.

Image jitter.ppm has been tampered or has not been watermarked.

(d)

(e)



(f)

Fig. 1. A successful jitter attack on SysCoP. We used the demo software release 1.0 available on SysCoP's Web site [76]. (a) shows an image watermarked with SysCoP and (b) the same image but after the attack. In the first case the software detects the watermark correctly (c) but the check fails on the modified image (d). Here, the attack simply consists in deleting and duplicating some columns of pixels such that the original size of the picture is conserved. (e) shows the columns which have been deleted (-) and duplicated (+). Finally, (f) is a magnified view of the white rectangle in (e); the bottom part corresponds to the original image.

Nyquist interpolation. In addition, a transfer function that introduces a small and smoothly distributed error into all sample values is applied. This emulates the small non-linear analog/digital converter imperfection typically found in scanners and display devices. StirMark introduces a practically unnoticeable quality loss in the image if it is applied only once. However after a few iterated applications, the image degradation becomes noticeable.

With those simple geometrical distortions we could confuse most marking systems available on the market. More distortions – still unnoticeable – can be applied to a picture. We applied a global 'bending' to the image: in addition to the general bi-linear property explained previously a slight deviation is applied to each pixel, which is greatest at the center of the picture and almost null at the borders. On top of this a higher frequency displacement of the form $\lambda \sin(\omega_x x) \sin(\omega_y y) + n(x, y)$ – where n is a random number – is added. In order for these distortions to be most effective, a medium JPEG compression is applied at the end.

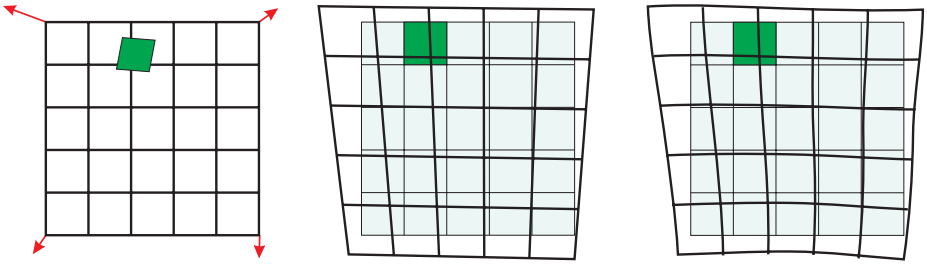


Fig. 2. We exaggerate here the distortion applied by StirMark to still pictures. The first drawing corresponds to the original picture; the others show the picture after StirMark has been applied – without and with bending and randomisation.

For those unfamiliar with digital image signal processing we shall now summarise briefly the main computation steps. Apart from a few simple operations such as rotations by 90 or 180 degrees, reflection and mirroring, image manipulation usually requires resampling when destination pixels do not line up with source pixels. In theory, one first generates a continuous image from the digital one, then modifies the continuous image, finally samples this to create a new digital image. In practice, however, we compute the inverse transform of a new pixel and evaluate the reconstruction function at that point.

There are numerous reconstruction filters. In a first version of the software we simply used a linear interpolation but, as foreseen, this tended to blur the image too much, making the validity of the watermark removal arguable. Then we implemented the sinc function as a reconstruction filter, which gives theoretically perfect reconstruction for photo images and can be described as follows. If (x, y) are the coordinates of the inverse transform – which, in our case is a distortion of the picture – of a point in the new image and f the function to be reconstructed,

then, an estimate of f at (x, y) is given by $\hat{f}(x, y) = \sum_{i=-n}^{n} \sum_{j=-n}^{n} \text{sinc}(x - i)\text{sinc}(y - j) f_{i,j}$. This gives very much better results than the simple filter; an example of the removal of an NEC watermark is given in Fig. 3.

We suggest that image watermarking tools which do not survive StirMark – with default parameters – should be considered unacceptably easy to break. This immediately rules out the majority of commercial marking schemes.

(a) (b)

Fig. 3. Kings' College Chapel, courtesy of John Thompson, JetPhotographic, Cambridge. For this example we watermarked a picture with NEC's algorithm [19]. We used the default parameters suggested by their paper ($N = 1000$ and $\alpha = 0.1$). (a) is the watermarked image. We then applied StirMark (b) and tested the presence of the watermark. The similarity between the original watermark and the extracted watermark was 3.74 instead of 21.08. This is well below the decision threshold.

One might try to increase the robustness of a watermarking system by trying to foresee the possible transforms used by pirates; one might then use techniques such as embedding multiple versions of the mark under suitable inverse transforms; for instance Ó Ruanaidh and Pereira suggest to use the Fourier-Mellin transform[2] to cope with rotation and scaling [50]. However, the general theme of the attacks we have developed and described above is that given a target marking scheme, we invent a distortion (or a combination of distortions) that will remove it or at least make it unreadable, while leaving the perceptual value of the previously marked object undiminished. We are not limited in this process to the distortions produced by common analogue equipment, or considered in the IFPI request for proposals cited above.

[2] The Fourier-Mellin transform is equivalent to the Fourier transform on a log-polar map: $(x, y) \to (\mu, \theta)$ with $x = e^{\mu} \cos\theta$ and $y = e^{\mu} \sin\theta$.

As an analogy, one might consider the 'chosen protocol attack' on authentication schemes [60]. It is an open question whether there is any marking scheme for which a chosen distortion attack cannot be found.

3.3 The Mosaic Attack

This point is emphasised by a 'presentation' attack, which is of quite general applicability and which possesses the initially remarkable property that a marked image can be unmarked and yet still rendered pixel for pixel in exactly the same way as the marked image by a standard browser.

The attack was motivated by a fielded automatic system for copyright piracy detection, consisting of a watermarking scheme plus a web crawler that downloads pictures from the net and checks whether they contain a watermark.

It consists of chopping an image up into a number of smaller subimages, which are embedded in a suitable sequence in a web page. Common web browsers render juxtaposed subimages stuck together, so they appear identical to the original image (Fig. 4). This attack appears to be quite general; all marking schemes require the marked image to have some minimal size (one cannot hide a meaningful mark in just one pixel). Thus by splitting an image into sufficiently small pieces, the mark detector will be confused [53]. The best that one can hope for is that the minimal size could be quite small and the method might therefore not be very practical.

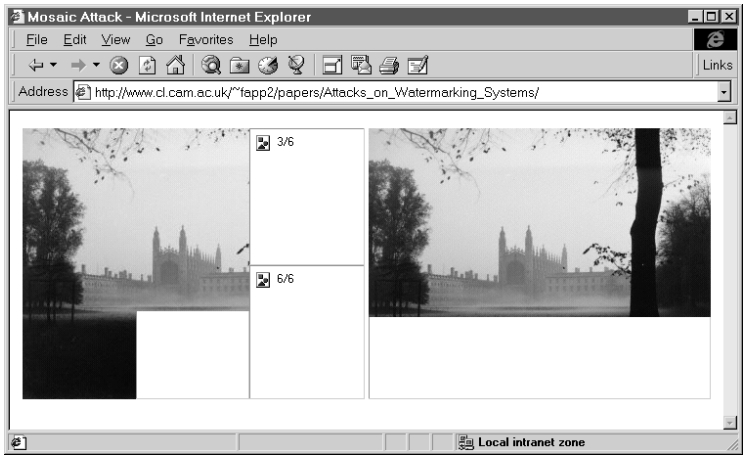


Fig. 4. Screen-shot of a web browser while downloading an image after the *mosaic attack*. This attack chops a watermarked image into smaller images which are stuck back together when the browser renders the page. We implemented software that reads a JPEG picture and produces a corresponding mosaic of small JPEG images as well as the necessary HTML code automatically [53]. In some cases downloading the mosaic is even faster than downloading the full image! In this example we used a 350×280-pixel image watermarked using PictureMarc 1.51.

There are other problems with such 'crawlers'. Java applets, ActiveX controls, etc. can be embedded to display a picture inside the browser; the applet could even de-scramble the picture in real time. Defeating such techniques would entail rendering the web page, detecting pictures and checking whether they contain a mark. An even more serious problem is that much current piracy is of pictures sold via many small services, from which the crawler would have to purchase them using a credit card before it could examine them. A crawler that provided such 'guaranteed sales' would obviously become a target.

3.4 Attack on *Echo Hiding*

One of the few marking schemes to be robust against the jitter attack is echo hiding, which hides information in sound by introducing echoes with very short delays. *Echo hiding* [29] relies on the fact that we cannot perceive short echoes (say 1 ms) and embeds data into a cover audio signal by introducing an echo characterised by its delay τ and its relative amplitude α. By using two types of echo it is possible to encode ones and zeros. For this purpose the original signal is divided into chunks separated by spaces of pseudo-random length; each of these chunks will contain one bit of information.

The echo delays are chosen between 0.5 and 2 milliseconds and the best relative amplitude of the echo is around 0.8. According to its creators, decoding involves detecting the initial delay and the auto-correlation of the cepstrum of the encoded signal is used for this purpose.

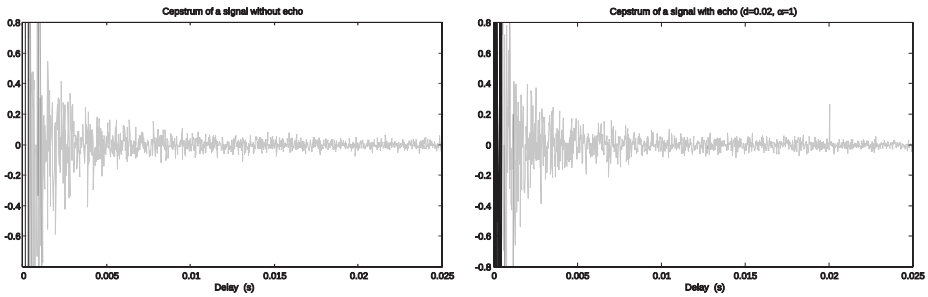
The 'obvious' attack on this scheme is to detect the echo and then remove it by simply inverting the convolution formula; the problem is to detect the echo without knowledge of either the original object or the echo parameters. This is known as 'blind echo cancellation' in the signal processing literature and is known to be a hard problem in general.

We tried several methods to remove the echo. Frequency invariant filtering [51, 59] was not very successful. Instead we used a combination of cepstrum analysis and 'brute force' search.

The underlying idea of cepstrum analysis is presented in [15]. Suppose that we are given a signal $y(t)$ which contains a simple single echo, i.e. $y(t) = x(t) + \alpha x(t-\tau)$. If we note Φ_{xx} the power spectrum of x then $\Phi_{yy}(f) = \Phi_{xx}(f)[1 + 2\alpha\cos(2\pi f\tau) + \alpha^2]$ whose logarithm is approximately $\log\Phi_{yy}(f) \approx \log\Phi_{xx}(f) + 2\alpha\cos(2\pi f\tau)$. This is a function of the frequency f and taking its power spectrum raises its 'quefrency' τ, that is the frequency of $\cos(2\pi\tau f)$. The auto-covariance of this later function emphasises the peak that appears at 'quefrency' τ (Fig. 5).

To remove the echos, we need a method to detect the echo delay τ. For this, we used a slightly modified version of the cepstrum: $\mathrm{C}\circ\Phi\circ\ln\circ\Phi$ where C is the auto-covariance function[3], Φ the power spectrum density function and $\circ$ the composition operator. Experiments on random signals as well as on music show that this method returns quite accurate estimators of the delay (Fig. 6) when an artificial echo has been added to the signal. In the detection function we only

[3] $\mathrm{C}(x) = \mathrm{E}[(x-\overline{x})(x-\overline{x})^*]$.



(a) (b)

Fig. 5. Graph (a) represents the cepstrum of a signal without echo. Graph (b) is the cepstrum of the same signal with a 20 ms echo which is emphasised by the very clear peak at 0.02 s.

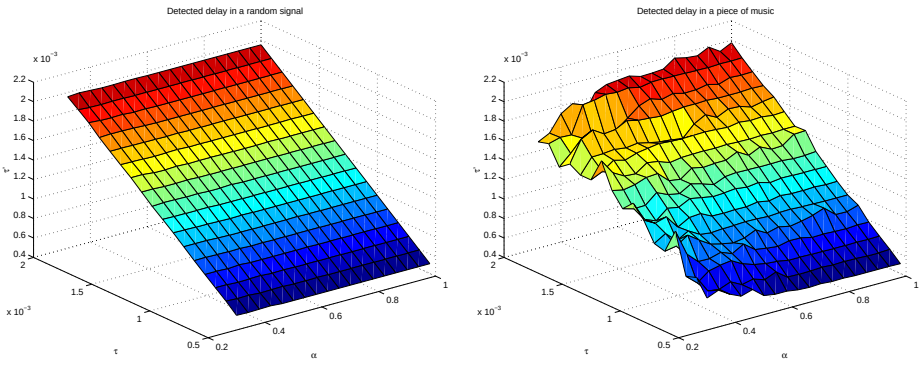
consider echo delays between 0.5 and 3 milliseconds. Below 0.5 ms the function does not work properly and above 3 ms the echo becomes too audible.

Our first attack was to remove an echo with random relative amplitude, expecting that this would introduce enough modification in the signal to prevent watermark recovery. Since echo hiding gives best results for α greater than 0.7 we could use $\tilde{\alpha}$ – an estimation of α – drawn from, say a normal distribution centred on 0.8. It was not really successful so our next attack was to iterate: we re-apply the detection function and vary $\tilde{\alpha}$ to minimise the residual echo. We could obtain successively better estimators of the echo parameters and then remove this echo. When the detection function cannot detect any more echo, we have got the correct value of $\tilde{\alpha}$ (as this gives the lowest output value of the detection function). Results obtained using this algorithm are presented in Fig. 6.

3.5 Protocol Considerations

The main threat addressed in the literature is an attack by a pirate who tries to remove the watermark directly. As a consequence, the definition commonly used for robustness includes only resistance to signal manipulation (cropping, scaling, resampling, etc.). Craver *et al.* show that this is not enough by exhibiting a 'protocol' level attack [22].

The basic idea is that many schemes provide no intrinsic way of detecting which of two watermarks was added first: the process of marking is often additive, or at least commutative. So if the owner of the document d encodes a watermark w and publishes the marked version $d + w$ and has no other proof of ownership, a pirate who has registered his watermark as w' can claim that the document is his and that the original unmarked version of it was $d + w - w'$. Their paper ([23]) extends this idea to defeat a scheme which is non-invertible (an inverse needs only be approximated).



(a) (b)

Fig. 6. Performances of the echo detector. We added different echoes characterised by their relative amplitude α and their delay τ to a signal and each time we used our echo detector to find an estimation $\hat{\tau}$ of τ. These graphs show the detected echo delay as a function of α and τ for random signals (a) and for a piece of music (b).

Craver *et al.* argue for the use of information-losing marking schemes whose inverses cannot be approximated closely enough. However, our alternative interpretation of their attack is that watermarking and fingerprinting methods must be used in the context of a larger system that may use mechanisms such as timestamping and notarisation to prevent attacks of this kind.

Registration mechanisms have not received very much attention in the copyright marking literature to date. The existing references such as [18, 32, 31, 52] mainly focus on protecting the copyright holder and do not fully address the rights of the consumers who might be fooled by a crooked reseller.

3.6 Implementation Considerations

The robustness of embedding and retrieving techniques is not the only issue. Most attacks on fielded cryptographic systems have come from the opportunistic exploitation of loopholes that were found by accident; cryptanalysis was rarely used, even against systems that were vulnerable to it [2].

We cannot expect copyright marking systems to be any different and the pattern was followed in the first attack to be made available on the Internet against the most widely used picture marking scheme, PictureMarc, which is bundled with Adobe Photoshop and Corel Draw. This attack [13] exploited weaknesses in the implementation rather than the underlying marking algorithms, even although these are weak (the marks can be removed using StirMark).

Each user has an ID and a two-digit password, which are issued when she registers with Digimarc and pays for a subscription. The correspondence between IDs and passwords is checked using obscure software in the implementation and although the passwords are short enough to be found by trial and error, the

attack first uses a debugger to break into the software and disable the password checking mechanism.

We note in passing that IDs are public, so either password search or disassembly can enable any user to be impersonated.

A deeper examination of the program also allows a villain to change the ID, thus the copyright, of an already marked image as well as the type of use (such as adult versus general public content). Before embedding a mark, the program checks whether there is already a mark in the picture, but this check can be bypassed fairly easily using the debugger with the result that it is possible to overwrite any existing mark and replace it with another one.

Exhaustive search for the personal code can be prevented by making it longer, but there is no obvious solution to the disassembly attack. If tamper resistant software [9] cannot give enough protection, then one can always have an online system in which each user shares a secret embedding key with a trusted party and uses this key to embed some kind of digital signature. Observe that there are two separate keyed operations here; the authentication (which can be done with a signature) and the embedding or hiding operation.

Although we can do public-key steganography – hiding information so that only someone with a certain private key can detect its existence [4] – we still do not know how to do the hiding equivalent of a digital signature; that is, to enable someone with a private key to embed marks in such a way that anyone with the corresponding public key can read them but not remove them. One problem is that a public decoder can be used by the attacker; he can remove a mark by applying small changes to the image until the decoder cannot find it anymore. This was first suggested by Perrig in [52]. In [42] a more theoretical analysis of this attack is presented as well as a possible countermeasure: randomisating the detection process. One could also make the decoding process computationally expensive. However neither approach is really satisfactory in the absence of tamper-resistant hardware.

Unless a breakthrough is made, applications that require the public verifiability of a mark (such as DVD) appear doomed to operate within the constraints of the available tamper resistance technology, or to use a central 'mark reading' service. This is evocative of cryptographic key management prior to the invention of public key techniques.

4 Conclusion

We have demonstrated that the majority of copyright marking schemes in the literature are vulnerable to attacks involving the introduction of sub-perceptual levels of distortion. In particular, many of the marking schemes in the marketplace provide only a limited measure of protection against attacks. Most of them are defeated by StirMark, a simple piece of software that we have placed in the public domain [38]. We have also shown a specific attack on the one serious exception to this rule (echo hiding).

This experience confirms our hypothesis that steganography would go through the same process of evolutionary development as cryptography, with an iterative process in which attacks lead to more robust systems.

Our experience in attacking the existing marking schemes has convinced us that any system which attempted to meet all the accepted requirements for marking (such as those set out by IFPI) would fail: if it met the robustness requirements then its bandwidth would be quite insufficient. This is hardly surprising when one considers that the information content of many music recording is only a few bits per second, so to expect to embed 20 bits per second against an opponent who can introduce arbitrary distortions is very ambitious.

Our more general conclusion from this work is that the 'marking problem' has been over-abstracted; there is not one 'marking problem' but a whole constellation of them. We do not believe that any general solution will be found. The trade-offs and in particular the critical one between bandwidth and robustness, will be critical to designing a specific system.

We already remarked in [8] on the importance of whether the warden was active or passive – that is, whether the mark needed to be robust against distortion. In general, we observe that most real applications do not require all of the properties in the IFPI list. For example, when auditing radio transmissions, we only require enough resistance to distortion to deal with naturally occurring effects such as multipath. Many applications will also require supporting protocol features, such as the timestamping service that we mentioned in the context of reversible marks.

So we do not believe that the intractability of the 'marking problem' is a reason to abandon this field of research. On the contrary; practical schemes for most realistic application requirements are probably feasible and the continuing process of inventing schemes and breaking them will enable us to advance the state of the art rapidly.

Finally, we suggest that the real problem is not so much inserting the marks as recognising them afterwards. Thus progress may come not just from devising new marking schemes, but in developing ways to recognise marks that have been embedded using the obvious combinations of statistical and transform techniques and thereafter subjected to distortion. The considerable literature on signal recognition may provide useful starting points.

Acknowledgements

Some of the ideas presented here were clarified by discussion with Roger Needham, David Wheeler, John Daugman, Peter Rayner, David Aucsmith, Stewart Lee, Scott Craver, Brian Moore, Mike Roe, Peter Wayner, Jon Honeyball, Scott Moskowitz and Matt Blaze.

References

1. Request for proposals – Embedded signalling systems issue 1.0. International Federation of the Phonographic Industry, 54 Regent Street, London W1R 5PJ, June 1997.
2. Ross J. Anderson. Why cryptosystems fail. *Communications of the ACM*, 37(11):32–40, November 1994.
3. Ross J. Anderson, editor. *Information hiding: first international workshop*, volume 1174 of *Lecture Notes in Computer Science*, Isaac Newton Institute, Cambridge, England, May 1996. Springer-Verlag, Berlin, Germany.
4. Ross J. Anderson. Stretching the limits of steganography. In IH96 [3], pages 39–48.
5. Ross J. Anderson and Markus G. Kuhn. Tamper resistance – A cautionary note. In *Second USENIX Workshop on Electronic Commerce*, pages 1–11, Oakland, CA, USA, November 1996.
6. Ross J. Anderson and Charalampos Manifavas. Chameleon – a new kind of stream cipher. In Biham [14], pages 107–113.
7. Ross J. Anderson and Roger M. Needham. Programming satan's computer. In J.van Leeuwen, editor, *Computer Science Today – Commemorative Issue*, volume 1000 of *Lecture Notes in Computer Science*, pages 426–441. Springer-Verlag, Berlin, Germany, 1995.
8. Ross J. Anderson and Fabien A. P. Petitcolas. On the limits of steganography. *IEEE Journal of Selected Areas in Communications*, 16(4):474–481, May 1998. Special Issue on Copyright & Privacy Protection.
9. David Aucsmith. Tamper resistant software: An implementation. In Anderson [3], pages 317–333.
10. David Aucsmith, editor. *Information Hiding: Second International Workshop*, volume 1525 of *Lecture Notes in Computer Science*, Portland, Oregon, USA, 1998. Springer-Verlag, Berlin, Germany.
11. Walter Bender, Daniel Gruhl, and Norishige Morimoto. Techniques for data hiding. In Niblack and Jain [48], pages 164–173.
12. Walter Bender, Daniel Gruhl, Norishige Morimoto, and Anthony Lu. Techniques for data hiding. *IBM Systems Journal*, 35(3 & 4):313–336, 1996.
13. Anonymous (<zguan.bbs@bbs.ntu.edu.tw>). Learn cracking IV – another weakness of PictureMarc. <news:tw.bbs.comp.hacker> mirrored on <http://www.cl.cam.ac.uk/~fapp2/watermarking/image_watermarking/digimarc_crack.html>, August 1997. Includes instructions to override any Digimarc watermark using PictureMarc.
14. Eli Biham, editor. *Fast Software Encryption – 4th International Workshop, FSE'97*, volume 1267 of *Lecture Notes in Computer Science*, Haifa, Israel, January 1997. Springer-Verlag, Germany.
15. Bruce P. Bogert, M.J.R. Healy, and John W. TEnglandey. The quefrency alanysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-ceptstrum and saphe cracking. In M. Rosenblatt, editor, *Symposium on Time Series Analysis*, pages 209–243, New York, NY, USA, 1963. John Wiley & Sons, Inc.
16. Laurence Boney, Ahmed H. Tewfik, and Khaled N. Hamdy. Digital watermarks for audio signals. In *European Signal Processing Conference, EUSIPCO '96*, Trieste, Italy, September 1996.
17. Laurence Boney, Ahmed H. Tewfik, and Khaled N. Hamdy. Digital watermarks for audio signals. In *International Conference on Multimedia Computing and Systems*, pages 473–480, Hiroshima, Japan, 17–23 June 1996. IEEE.

18. Marc Cooperman and Scott A. Moskowitz. Steganographic method and device. US Patent 5,613,004, March 1995.
19. Ingemar J. Cox, Joe Kilian, Tom Leighton, and Talal Shamoon. A secure, robust watermark for multimedia. In Anderson [3], pages 183–206.
20. Ingemar J. Cox and Matt L. Miller. A review of watermarking and the importance of perceptual modeling. In Rogowitz and Pappas [57].
21. Ingemar J. Cox and Kazuyoshi Tanaka. NEC data hiding proposal. Technical report, NEC Copy Protection Technical Working Group, July 1997. Response to call for proposal issued by the Data Hiding SubGroup.
22. Scott Craver, Nasir Memon, Boon-Lock Yeo, and Minerva M. Yeung. Can invisible watermark resolve rightful ownerships? In Sethin and Jain [62], pages 310–321.
23. Scott Craver, Nasir Memon, Boon-Lock Yeo, and Minerva M. Yeung. Resolving rightful ownerships with invisible watermarking techniques: Limitations, attacks, and implications. *IEEE Journal of Selected Areas in Communications*, 16(4):573–586, May 1998. Special Issue on Copyright & Privacy Protection. ISSN 0733-8716.
24. Digimarc home page. `<http://www.digimarc.com/>`, April 1997.
25. Eikonamark. Alpha Tec Ltd., `<http://www.generation.net/~pitas/sign.html>`, October 1997.
26. Elke Franz, Anja Jerichow, Steffen Möller, Andreas Pfitzmann, and Ingo Stierand. Computer based steganography: how it works and why therefore any restriction on cryptography are nonsense, at best. In Anderson [3], pages 7–21.
27. Michael A. Gerzon and Peter G. Graven. A high-rate buried-data channel for audio CD. *Journal of the Audio Engineering Society*, 43(1/2):3–22, January–February 1995.
28. François Goffin, Jean-François Delaigle, Christophe De Vleeschouwer, Benoît Macq, and Jean-Jacques Quisquater. A low cost perceptive digital picture watermarking method. In Sethin and Jain [62], pages 264–277.
29. Daniel Gruhl, Walter Bender, and Anthony Lu. Echo hiding. In Anderson [3], pages 295–315.
30. Khaled N. Hamdy, Ahmed H. Tewfik, Ting Chen, and Satoshi Takagi. Time-scale modification of audio signals with combined harmonic and wavelet representations. In *International Conference on Acoustics, Speech and Signal Processing – ICASSP '97*, volume 1, pages 439–442, Munich, Germany, April 1997. IEEE, IEEE Press. Session on Hearing Aids and Computer Music.
31. Alexander Herrigel, Joseph J. K. Ó Ruanaidh, Holger Petersen, Shelby Pereira, and Thierry Pun. Secure copyright protection techniques for digital images. In Aucsmith [10], pages 169–190.
32. Alexander Herrigel, Adrian Perrig, and Joseph J. K. Ó Ruanaidh. A copyright protection environment for digital images. In *Verläßliche IT-Systeme '97*, Albert-Ludwigs Universität, Freiburg, Germany, October 1997.
33. J.N. Holmes. *Speech Synthesis and Recognition*, chapter 3.6 Analysis of simple and complex signals, pages 47–48. Aspects of Information Technology. Chapman & Hall, London, England, 1988.
34. International Electrotechnical Commission, Geneva, Switzerland. *Digital audio interface, IEC 60958*, February 1989.
35. Alastair Kelman. Electronic copyright management – the way ahead. Security Seminars, University of Cambridge, February 1997.
36. A. Kerckhoffs. La Cryptographie Militaire. *Journal des Sciences Militaires*, 9:5–38, January 1883.

37. E. Koch and J. Zhao. Towards robust and hidden image copyright labeling. In *Workshop on Nonlinear Signal and Image Processing*, pages 452–455, Neos Marmaras, Greece, June 1995. IEEE.
38. Markus G. Kuhn and Fabien A. P. Petitcolas. StirMark. `<http://www.cl.cam.ac.uk/~fapp2/watermarking/stirmark/>`, November 1997.
39. Charles Kurak and John McHugh. A cautionary note on image downgrading. In *Computer Security Applications Conference*, pages 153–159, San Antonio, TX, USA, December 1992.
40. Gerrit C. Langelaar, Jan C.A van der Lubbe, and J. Biemond. Copy protection for multimedia data based on labeling techniques. In *17th Symposium on Information Theory in the Benelux*, Enschede, The Netherlands, May 1996.
41. Gerrit C. Langelaar, Jan C.A. van der Lubbe, and Reginald L. Lagendijk. Robust labeling methods for copy protection of images. In Sethin and Jain [62], pages 298–309.
42. Jean-Paul M.G. Linnartz and Marten van Dijk. Analysis of the sensitivity attack against electronic watermarks in images. In Aucsmith [10], pages 258–272.
43. Mark Lomas, Bruno Crispo, Bruce Christianson, and Mike Roe, editors. *Security Protocols: Proceeding of the 5th International Workshop*, volume 1361 of *Lecture Notes in Computer Science*, École Normale Supérieure, Paris, France, April 1997. University of Cambridge, Isaac Newton Institute, Springer-Verlag, Berlin, Germany.
44. Maurice Maes. Twin peaks: The histogram attack on fixed depth image watermarks. In Aucsmith [10], pages 290–305.
45. Kineo Matsui and Kiyoshi Tanaka. Video-steganography: How to secretly embed a signature in a picture. *Journal of the Interactive Multimedia Association Intellectual Property Project*, 1(1):187–205, January 1994.
46. Norishige Morimoto and Daniel Sullivan. IBM DataHiding proposal. Technical report, IBM Corporation, September 1997. Response to call for proposal issued by the Data Hiding SubGroup.
47. Peter Nancarrow. Digital technology – Bane or boon for copyright? Computer Laboratory Seminars, University of Cambridge, November 1997.
48. Wayne Niblack and Ramesh C. Jain, editors. *Storage and Retrieval for Image and Video Database III*, volume 2420, San Jose, California, USA, February 1995. IS&T, The Society for Imaging Science and Technology and SPIE, The International Society for Optical Engineering, SPIE.
49. Joseph J. K. Ó Ruanaidh, W. J. Dowling, and F. M. Boland. Watermarking digital images for copyright protection. *IEE Proceedings on Vision, Signal and Image Processing*, 143(4):250–256, August 1996.
50. Joseph J. K. Ó Ruanaidh and Shelby Pereira. A secure robust digital image watermark. In *International Symposium on Advanced Imaging and Network Technologies – Conference on Electronic Imaging: Processing, Printing and Publishing in Colour*, Europto, Zürich, Switzerland, May 1998. International Society for Optical Engineering, European Optical Society, Commission of the European Union, Directorate General XII.
51. Alan V. Oppenheim and Ronald W. Schafer. *Discrete-Time Signal Processing*, chapter 12, pages 768–834. Prentice-Hall International, Inc., Englewood Cliffs, NJ, USA, international edition, 1989.
52. Adrian Perrig. A copyright protection environment for digital images. Diploma dissertation, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, February 1997.

53. Fabien A. P. Petitcolas. Weakness of existing watermarking schemes. <http://www.cl.cam.ac.uk/~fapp2/watermarking/image_watermarking/>, October 1997.
54. Birgit Pfitzmann. Information hiding terminology. In Anderson [3], pages 347–350. Results of an informal plenary meeting and additional proposals.
55. I. Pitas. A method for signature casting on digital images. In *International Conference on Image Processing*, volume 3, pages 215–218, September 1996.
56. Geoffrey B. Rhoads. Steganography methods employing embedded calibration data. US Patent 5,636,292, June 1997.
57. Bernice E. Rogowitz and Thrasyvoulos N. Pappas, editors. *Human Vision and Electrtonic Imaging II*, volume 3016, San Jose, CA, USA, February 1997. IS&T, The Society for Imaging Science and Technology and SPIE, The International Society for Optical Engineering, SPIE.
58. Pamela Samuelson. Copyright and digital libraries. *Communications of the ACM*, 38(4):15–21, 110, April 1995.
59. Ronald W. Schafer. Echo removal by discrete generalized linear filtering. Technical Report 466, Massachusetts Institute of Technology, February 1969.
60. Bruce Schneier. Protocol interactions and the chosen protocol attack. In Lomas et al. [43], pages 91–104.
61. Robert A. Scholtz. The origins of spread-spectrum communications. *IEEE Transactions on Communications*, 30(5):822–853, May 1982.
62. Ishwar K. Sethin and Ramesh C. Jain, editors. *Storage and Retrieval for Image and Video Database V*, volume 3022, San Jose, CA, USA, February 1997. IS&T, The Society for Imaging Science and Technology and SPIE, The International Society for Optical Engineering, SPIE.
63. Signum Technologies – SureSign digital fingerprinting. <http://www.signumtech.com/>, October 1997.
64. Mitchell D. Swanson, Bin Zhu, and Ahmed H. Tewfik. Transparent robust image watermarking. In *International Conference on Image Processing*, volume III, pages 211–214. IEEE, 1996.
65. Mitchell D. Swanson, Bin Zu, and Ahmed H. Tewfik. Robust data hiding for images. In *7th Digital Signal Processing Workshop (DSP 96)*, pages 37–40, Loen, Norway, September 1996. IEEE.
66. A.Z. Tirkel, G.A. Rankin, R.M. van Schyndel, W.J. Ho, N.R.A. Mee, and C.F. Osborne. Electronic watermark. In *Digital Image Computing, Technology and Applications – DICTA '93*, pages 666–673, Macquarie University, Sidney, 1993.
67. R.G. van Schyndel, A.Z. Tirkel, and C.F. Osborne. A digital watermark. In *International Conference on Image Processing*, volume 2, pages 86–90, Austin, Texas, USA, 1994. IEEE.
68. Georges Van Slype. Natural language version of the generic CITED model – ECMS (Electronic Copyright Management System) design for computer based applications. Report 2, European Commission, ESPRIT II Project, Bureau Vam Dijk, Brussel, Belgium, May 1995.
69. Georges Van Slype. Natural language version of the generic CITED model – Presentation of the generic model. Report 1, European Commission, ESPRIT II Project, Bureau Vam Dijk, Brussel, Belgium, May 1995.
70. A. Werner, J. Oomen, Marc E. Groenewegen, Robbert G. van der Waal, and Raymond N.J. Veldhuis. A variable-bit-rate buried-data channel for compact disc. *Journal of the Audio Engineering Society*, 43(1/2):23–28, January–February 1995.
71. The Working Group on Intellectual Property Rights is part of the US Information Infrastructure Task Force, formed in February 1993.

72. Raymond B. Wolfgang and Edward J. Delp. A watermark for digital images. In *International Conference on Images Processing*, pages 219–222, Lausanne, Switzerland, September 1996. IEEE.
73. Raymond B. Wolfgang and Edward J. Delp. A watermarking technique for digital imagery: further studies. In *International Conference on Imaging, Systems, and Technology*, pages 279–287, Las Vegas, NV, USA, 30 June–3 July 1997. IEEE.
74. J. Zhao and E. Koch. Embedding robust labels into images for copyright protection. In *International Congress on Intellectual Property Rights for Specialised Information, Knowledge and New Technologies*, Vienna, Austria, August 1995.
75. Jian Zhao. A WWW service to embed and prove digital copyright watermarks. In *European Conference on Multimedia Applications, Services and Techniques*, pages 695–710, Louvain-la-Neuve, Belgium, May 1996.
76. Jian Zhao. The syscop home page. `<http://syscop.igd.fhg.de/>` or `<http://www.crcg.edu/syscop/>`, February 1997.