Analysis of the Sensitivity Attack against Electronic Watermarks in Images

Jean-Paul M. G. Linnartz and Marten van Dijk

Both authors are with Eindhoven Philips Research Laboratories (Natlab),
Holstlaan 4, 5656 AA, Eindhoven, the Netherlands,
e-mail: {linnartz, mvandijk}@natlab.research.philips.com

Abstract. In some applications of electronic watermarks, the device that detects whether content contains a watermark or not is in public domain. Attackers can misuse such detector as an oracle that reveals up to one bit of information about the watermark in each experiment. An information-theoretical analysis of the information leakage is provided, and a method is proposed to reduce the information leakage by orders of magnitude.

keywords: Cryptanalysis, Copy Protection, Electronic Watermarks

1 Introduction

It is an open problem whether reliable and secure *public watermarks* can exists. Such public watermarks allow anyone to detect electronic watermarks, while the security and robustness are not affected by this public knowledge. By *secure* we mean that knowledge about how to detect a watermark does not reveal how the watermark can be removed or altered. We call the watermarking scheme *reliable* if it is robust to typical transmission and storage imperfections (such as lossy compression, noise addition, format conversion, bit errors) and signal processing artefacts (noise reduction, filtering), whether intentional or not. Moreover, content that has not been watermarked may not trigger a detector, or at least this probability should be negligibly small. Typical requirements for watermarking methods are

1. The watermark should be secure. Erasing the watermark should be technically difficult.
2. The watermarking scheme should be reliable.
3. An original image and its marked version should be perceptually indistinguishable. After commonly accepted processing, e.g. MPEG lossy compression, the accumulated artifacts should not be visible.

Public watermarks are desirable for copy management and embedded signalling of author's and publisher's data within the content. In innovative copy protection schemes, as for instance intended for new generation (Digital Versatile Disc) DVD systems, a consumer device performs a watermark detection as part of its judgement whether the content is original, or a legal or illegal copy. Watermarked content on discs that do not have the correct physical identifiers of the original publisher will not be played. For all systems known to the authors, the watermark detection method, i.e., its algorithm and the "keys", have to be kept secret to avoid that copyright pirates can remove the watermark. It is often assumed that the watermark detector is therefore implemented as a tamperproof box such that the attacker can not reverse-engineer critical parameters or properties of the detector from the implementation. An important class of proposed detectors is covered in Section 2.

An attacker can nonetheless learn and erase the watermark by experimenting with the content that he inputs to the detector [1]. Unless special precautions are taken, the attacker gains one bit of information about the watermark in every attempt. This implies that the attack is linear with the number of pixels in the image. This is in sharp contrast with the common belief that an attacker must do order $\mathrm{O}(exp(N))$ experiments to find a secret watermark in an image of N pixels. In Section 3 we describe the attack. An attacker is successful if he can modify a marked image such that the detector responds that it does not see a watermark, while the modifications to the image are invisible. We propose a countermeasure that increases the work load for an attacker by a several orders of magnitude in Sections 4-6.

2 Typical Watermarking Detector

Let us consider a rectangular image r of size N_1 by N_2 pixels. The coordinates of the pixels are denoted by $\boldsymbol{n} \in A = \{(n_1, n_2) : 0 \leq n_1 \leq N_1 - 1, 0 \leq n_2 \leq N_2 - 1\}$. The luminance of the pixel with coordinates $\boldsymbol{n}$ is denoted as $r(\boldsymbol{n})$. We represent the watermark as w or $w(\boldsymbol{n})$, which takes on a value in each pixel $\boldsymbol{n} \in A$. A watermark detector outputs $D = 1$ if it recognizes a watermark, otherwise $D = 0$.

The most commonly used watermark detector bases this decision on the correlation between the suspect image and (a possibly transformed version of) the watermark [2–6]. Although many authors do not explicitly mention a correlator as their detection method, many schemes published thus far are mathematically equivalent to detection by correlation, or extensions of this basic concept. Such detector, as for instance in Figure 1, extracts a decision variable y from the suspect image q through a correlation operation $R_w(q)$ with a locally stored copy of the watermark w;

$$y = R_w(q) = \sum_{\boldsymbol{n} \in A} w(\boldsymbol{n}) q(\boldsymbol{n}).$$

Then, if $y > y_{thr}$ with y_{thr} some threshold value, it decides that the watermark is present and it outputs $D = 1$, otherwise $D = 0$.

We refer to [4] for an evaluation of how a decision threshold y_{thr} relates to the probability of a missed detection (the watermark is present, but the detector thinks it is not) and the probability of a false alarm (no watermark is embedded,

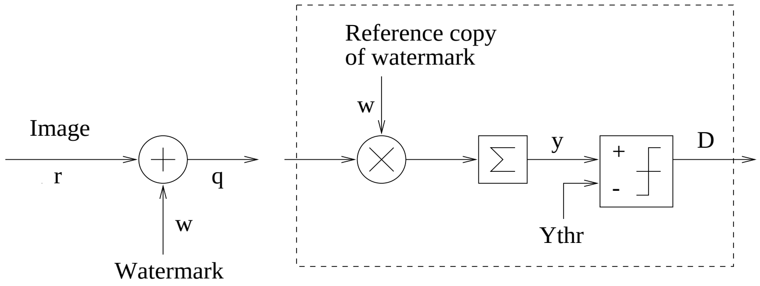

Fig. 1. *Correlator detector*

but the detector thinks one is). These probabilities measure the reliability of the watermarking scheme.

The output of the detector D can be seen as a random variable depending on y. In fact we have the Markov sequence

$$q \to y \to D,$$

where q, $y = R_w(q)$, and D are interpreted as random variables. I.e., the distribution function of random variable D, conditioned on the entire past, can be expressed exactly through conditioning only on the most recent random variable y.

Note that here we do not explicitly describe how an original image is watermarked in order to trigger a detector. In the standardization of watermarks for copy protection, it has become clear that only the detection algorithm needs to be prescribed, whereas the content owner can be given the freedom to use proprietary solutions for embedding the watermark. Particularly because of ongoing developments in perceptual modelling, such solutions tend to differ from implementation to implementation and to improve over time [7]. The reader may assume that the embedding method creates a new image q with $q = r + \eta * w$, where η is an appropriate embedding depth and $*$ is a pixelwise multiplication. The attack described in this paper is considered to be successful if the attacker manages to modify a watermark image in such a way that the detector will not be triggered. This neither implies that he recovered the original image precisely as it was before marking, nor that the new image is free of remnants from the watermark. However, one can use the r.m.s. modifications to the marked image as a first-order indication of the perceptual damage to the image.

In order to intuitively understand the concept of the attack and the countermeasures, we now present a geometrical interpretation of the correlator detector. This attack has been successfully executed against several more sophisticated watermarking methods.

Pictures are interpreted as vectors in an $N_1 N_2 = N$ dimensional vector space, see Figure 2. The vector space consists of three parts; $S_< = \{b : R_w(b) < y_{thr}\}$, $S_> = \{b : R_w(b) > y_{thr}\}$, and $S_= = \{b : R_w(b) = y_{thr}\}$. For pictures in $S_<$ the detector outputs 0. With probability close to one, a random unmarked image

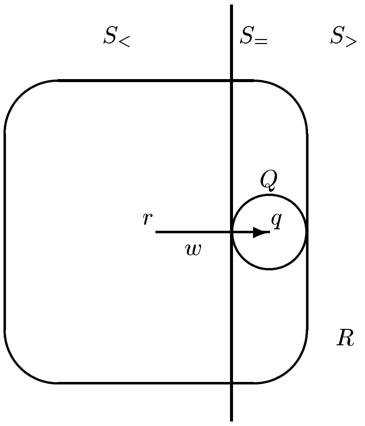

Fig. 2. *Geometrical Interpretation. Image r, Watermark w, Marked Image q*

$r \in S_<$. We will consider only those original images that do not raise a false alarm in a detector, that is we do not address the small fraction of those original images that by accident are within $S_>$. For marked pictures, which are in $S_>$ the detector outputs 1. On the separating surface $S_=$, the watermark detector also outputs $D = 0$. Area R contains all pictures which are perceptually indistinguishable from r. According to requirement 3 we have that $q \in R$.

Area Q contains modifications of q caused by typical transmission and storage imperfections and signal processing artefacts. According to requirement 3 such pictures should be perceptually indistinguishable from r as well, thus $Q \subseteq R$. The watermarking scheme should be reliable, see requirement 2, hence, $Q \subseteq S_>$. Summarizing, we have that $r \in R$, and $q \in Q \subseteq R \cap S_>$, and we assume that a watermarking method exists that allows q to be created.

The attacker's task is to find a point in $S_<$, preferably as close as possible to r. In practice, he will be satisfied with $\hat{r} \in S_<$ close to q and he hopes that $\hat{r} \in R$. We conclude this section by noting that in the figures the geometrical shape of the areas are idealized.

3 The Attack

The attacker is assumed to have a marked image q (from which he attempts to remove the watermark) and to have access to the input and output of a watermark detector. This detector can either be in a tamperproof box, or it can be a remote server algorithm on a network that allows users to submit random images for watermarks detection.

In abstract terms, the attacker operates as follows [1]:

[Select random point in $S_<$, near $S_=$] He initially searches for a random point $q_0 \in S_<$ as close as practically possible to $S_=$. At this point it does not matter whether the resulting image resembles the original or not. The only criterion is that some minor modifications to the test image cause the detector to respond with $D = 1$ while other modifications result in $D = 0$. One method is to gradually replace more and more pixels in the image by neutral grey.

[Find tangent e_l] He then estimates the tangent e_l to the surface $S_=$ by taking a random vector t_j and searches the values γ_j for which $q_l + \gamma_j t_j$ changes the decision of the detector. Typically, one only needs a single small positive or negative value for γ_j, e.g. $\gamma_j \in \{-1, +1\}$. A useful choice for t_j is zero for all pixels except for a single pixel $\boldsymbol{n_j}$. That is, $q_l + \gamma_j t_j$ slightly increases or decreases the luminance of that pixel just enough to ensure to trigger the detector ($q_l + \gamma_j t_j \in S_>$). This provides the insight of whether $w(\boldsymbol{n_j}) > 0$ or < 0. In a more sophisticated version, one can also estimate the value of $w(\boldsymbol{n_j})$.

This test is repeated for a complete set of independent vectors t_j, $j = 0$, $1, \ldots, N-1$. At the end the attacker has gained knowledge about w and, hence, about the shape of the surface $S_=$ near q_l. Using this knowledge he estimates the tangent e_l to the surface $S_=$ near q_l.

[Create a point q_{l+1} in $S_<$ near $S_=$] 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 (expected) influence on the detector decision variable. The attacker uses the original marked image q (or q_l) and subtracts $\lambda_l * e_l$ resulting in a new point q_{l+1} in $S_<$ near $S_=$, such that the detector reports that no watermark is present. Parameter e_l is the tangent vector constructed in the previous step. Parameter λ_l may be found experimentally, such that λ_l may have the smallest perceptual effect on the image. A sophisticated attacker also exploits a perceptual model that makes the value of λ_l dependent on the pixel location. This is the final step for watermarking schemes with a simple correlator. If the surface $S_=$ is not a hyper plane, e.g., if the threshold value depends on the variance in the image, or if the surface is a collection of parts of hyperplanes, the attacker may iterate.

[Iterate] If the attacker is dissatisfied with the perceptual damage to the image, he may treat this image q_{l+1} again as a test image to estimate the local sensitivities. That is, he repeats the procedure for $l + 1$ (find tangent e_{l+1} and create a point q_{l+2} in $S_<$ on or very close to the separating surface $S_=$) until he finds a point q_n appropriately close to q.

If the surface $S_=$ in not a perfect plane, he may need to invoke more sophisticated searching algorithms, possibly including simulated annealing. However, for most correlator-based detection methods the attack only needs a single round of the above iterative process. For intuitive understanding we analyse the attack against a simple correlator/threshold detector with an idealised perceptual model. In this case a single round of iteration is sufficient. For ease of analysis we focus on the special case $w(\boldsymbol{n}) \in \{-k, k\}$ where $k > 0$, i.e. similar to proposals as in for instance in [2, 5].

4 Countermeasure

It appears possible to make the watermark detector substantially less vulnerable to the attack by randomizing the transition point y_{thr} of the detector. If the transition area $S_=$ is not a perfect plane, but a fuzzy area with random decisions by the detector if $y \approx y_{thr}$, an attacker will get much less information from each experiment performed in Step 2. If the randomization only occurs in a limited range of the decision value, the effect on the reliability is small.

For instance, instead of using one threshold y_{thr}, the detector uses two thresholds y_1 and y_2 with $y_2 > y_1$. If $y < y_1$, $D = 0$ and if $y > y_2$, $D = 1$. In the range $y_1 < y < y_2$, the detector chooses $D = 1$ with probability $p(y)$, where $p(y)$ is smoothly increasing in y.

4.1 Reliability

For reliability reasons the detector must respond $D = 0$ with very high probability for unmarked images and with $D = 1$ for marked images. Random responses are acceptable only in a transition range: y_1 is taken large enough such that the probability for a random, unmarked image not to generate $D = 0$ is small enough (probability of a false alarm). Similarly, y_2 is taken small enough such that the probability for a watermarked image not to generate $D = 1$ is small enough (probability of a missed detection). To satisfy the reliability requirements, the system designer should select the decision interval $[y_1, y_2]$ small enough such that the reliability of the detector stays within acceptable range. On the other hand, the length of the transition interval $[y_1, y_2]$ is taken large enough to ensure that for small changes to the image (resulting in small changes to y), the gradient of the decision probability $p(y)$ is only noticeable to an attacker after taking many samples and statistically processing these. It has been shown that the decision variable is a Gaussian random variable. Its mean value corresponds to the energy in the watermark, defined as $E_w = \sum_{\boldsymbol{n} \in A} w(\boldsymbol{n})^2$. The variance is determined by the variance of pixel luminance values, thus $\sigma^2 = \mathrm{E}r^2 - \mathrm{E}^2 r$ and other parameters. Erroneous detections occur with a probability that is determined by the energy in the watermark, the threshold setting and the variance of the random cross correlation between the original image and the reference watermark. If, in a detector without a countermeasure, a threshold of y_{thr} would be chosen, one could include the countermeasure by taking $y_1 = y_{thr}$ and $y_2 = \mathit{\Gamma} y_{thr}$. This would require the watermark to be embedded with a slight increase in energy, determined by $\mathit{\Gamma}$. This increase can be limited to a few dB, however, an detailed evaluation is outside the scope of this paper.

4.2 Sophisticating the attack

Despite the random responses, an attacker can nonetheless extract information if he manages to estimate $p(q_0)$ and $p(q_0 + \gamma_j t_j)$. He could estimate these probabilities by repeated trials. Particularly, if $p(y)$ has a pronounced discontinuity at y_d, he could launch the attack near y_d. If for instance the detector would flip an unbiased coin when $y_1 < y < y_2$, the attacker launches the attack either at $y \approx y_1$ or $y \approx y_2$. In a few attempts he will learn whether the probability is 0, 0.5 or 1 for each $q_0 + \gamma_j t_j$.

There appears to be an optimum shape for $p(y)$ which minimizes the leakage of information, independent of the value of $R_w(q_0)$ at which the attack is executed. In the coming sections we will construct, study, and analyse this optimal shape.

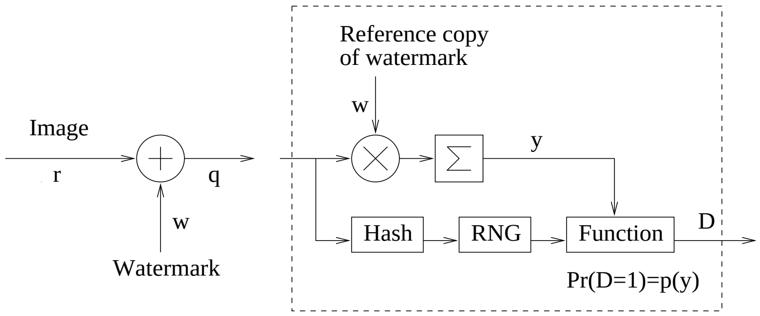

Fig. 3. *Improved detector using countermeasure*

Figure 3 gives an example of a possible implementation. For $y_1 < y < y_2$, the behavior is determined by the cryptographic Hash value generator, the Random Number generator and the Function. We notice that the implementation of Figure 3 results in a deterministic machine. That is if a fixed image q is input in the detector then either it always detects the watermark or it never detects the watermark. This avoids that an attacker estimates $p(y)$ by inputting the same image q in the detector over and over.

To the attacker not aware of the internal behaviour of the hash and random function generator,

$$Pr(D(q) = 1 | R_w(q) = y) = p(y).$$

Let us consider an attacker eager to find $p(y)$, who therefore manages to find small modifications $q+\epsilon$ to the image q, where $y = R_w(q) = R_w(q+\epsilon)$ ($R_w(\epsilon) = 0$ if the detector is a linear correlator). The output of the Random Number generator for these modifications $q+\epsilon$ differ in an unpredictable manner. Hence, for a fraction $p(y)$ of all small modifications $q+\epsilon$ we have $D(q+\epsilon) = 1$. Thus by interpreting q as a fixed picture and ϵ as a uniformly distributed random variable representing small modifications with $R_w(q) = R_w(q+\epsilon)$,

$$Pr(D(q+\epsilon) = 1 | R(q) = y) = p(y). \tag{1}$$

5 Probabilistic Behavior

As we argued before, preferably $p(y)$ is a smooth function. The attacker can still estimate the sensitivity of $p(y)$ to his intentional modifications $\gamma_j t_j$ of the image. Hence he will learn the relation between y and $\gamma_j t_j$. We will determine the optimum relation between $p(y)$ and y to protect against this attack.

Assume that the pirate has created a test image q_l in step 3 or initially in step 1 of the attack. In the following analysis we focus on step 2 (estimating the tangent). More specifically we investigate how the attacker can find $p(q_l + \gamma_j t_j)$ by making second-oder small modifications ϵ_i [1].

Let $R_w(q_l) = y$. For small modifications $q_l + \epsilon, |\epsilon| << |\gamma_j t_j|$ to the original image q_l approximation

$$y = R(q_l) \approx R(q_l + \epsilon)$$

holds. Thus the detector returns $D = 1$ with probability $p(y)$ for these small modifications, see (1). Henceforward, through many experiments with different ϵ_i's the attacker is able to estimate the value of $p(y) = p(R_w(q_l))$.

Let $q_{l,j,i} = q_l + \gamma_j t_j + \epsilon_i$ such that

$$R_w(q_{l,j,i}) \approx R_w(q_l + \gamma_j t_j).$$

For ease of notation we write q_i instead of $q_{l,j,i}$. Image $\gamma_j t_j$ is a bias, which we interpret as a test image which is non-zero only in one pixel $\boldsymbol{n_j}$. For ease of notation we write t instead of $\gamma_j t_j$ and $\boldsymbol{n}$ instead of $\boldsymbol{n_j}$. Let $t(\boldsymbol{n}) = \alpha > 0$ and $\delta = \alpha k$. Then the effect Δy that t has on the decision variable is

$$\Delta y = R_w(t) = \alpha \cdot w(\boldsymbol{n}) \in \{-\alpha k, +\alpha k\} = \{-\delta, +\delta\}.$$

Since R_w is (at least in a first order approximation) a linear function we have that

$$R_w(q_i) \approx R_w(q_l + t) = R_w(q_l) + R_w(t) = y + \Delta y$$

and if $p(y)$ is a smooth (differentiable) function

$$p(R_w(q_i)) \approx p(y + \Delta y) \approx p(y) + \Delta y \cdot p'(y),$$

where $p'(y)$ denotes the derivative of function p evaluated in y. Thus the detector returns output $D_i = 1$ with probability $p(y) + \Delta y \cdot p'(y)$ for image q_i, that is

$$\begin{aligned} Pr(D_i = 1 | w(\boldsymbol{n}) = +k) &\approx p(y) + \delta \cdot p'(y), \\ Pr(D_i = 1 | w(\boldsymbol{n}) = -k) &\approx p(y) - \delta \cdot p'(y). \end{aligned} \qquad (2)$$

Henceforward, through many experiments with different q_i's the attacker is able to estimate the value of $p(y) + \Delta y \cdot p'(y)$. If this value is more than $p(y)$ the attacker concludes that $\Delta y = +\delta$ and $w(\boldsymbol{n}) = +k$. If it is less than $p(y)$ the attacker concludes that $\Delta y = -\delta$ and $w(\boldsymbol{n}) = -k$.

[1] Here q_l is the zero-order attempt, $\gamma_j t_j$'s describe first-order sensitivity measurements, and ϵ_i's describes second-order randomizations to obtain statistical averages.

The idea is that the attacker gathers information about the polarity of the watermark in the pixel $w(\boldsymbol{n})$ through the series of test images q_i. This results in a sequence of outputs of the detector $D_1, D_2, \ldots, D_i, \ldots$. In the remainder we denote $D_1, \ldots, D_i$ by D^i. The average amount of bits needed to describe a realization of D^i is measured by the uncertainty about D^i, denoted by the entropy function $H(D^i)$. The average amount of bits needed to describe D^i given the knowledge of $w(\boldsymbol{n})$ is measured by the conditional uncertainty about D^i given $w(\boldsymbol{n})$, denoted by the conditional entropy function $H(D^i|w(\boldsymbol{n}))$. The mutual information $I(D^i; w(\boldsymbol{n})) = H(D^i) - H(D^i|w(\boldsymbol{n}))$ between D^i and $w(\boldsymbol{n})$ measures the amount of information D^i and $w(\boldsymbol{n})$ have in common. Hence, $I(D^i; w(\boldsymbol{n}))$ measures the amount of information that the observation D^i reveals about the unknown $w(\boldsymbol{n})$. We notice that $w(\boldsymbol{n})$ takes on values $-k$ and $+k$ with equal probability, hence $H(w(\boldsymbol{n})) = 1$ bit. The entropy function h is defined as $h(x) = -x \log x - (1-x) \log(1-x)$ where the logarithm is of base 2. We notice that $h'(x) = \log((1-x)/x)$ and that the second derivative $h^{(2)}(x) = -1/(x(1-x) \ln 2)$. For a thorough treatment in information theory we refer to Cover and Thomas [8].

To defend the confidentiality of the watermark, the system designer of the copy protection scheme keeps the information that D^i reveals about the watermark as small as possible. He designs function p such that $I(D^i; w(\boldsymbol{n}))$ is small enough. Let us analyse this mutual information. Let us consider the special case $i = 1$. From approximations (2) and the definition of entropy we infer that

$$\begin{aligned}
H(D_1|w(\boldsymbol{n}) = +k) &\approx h(p(y) + \delta p'(y)),\\
H(D_1|w(\boldsymbol{n}) = -k) &\approx h(p(y) - \delta p'(y)),\\
H(D_1|w(\boldsymbol{n})) &\approx (h(p(y) + \delta p'(y)) + h(p(y) - \delta p'(y)))/2,
\end{aligned}$$

and $Pr(D_1 = 1) = ((p(y) + \delta p'(y)) + (p(y) - \delta p'(y)))/2 = p(y)$, hence

$$H(D_1) = h(p(y)).$$

We conclude that

$$\begin{aligned}
I(D^1; w(\boldsymbol{n})) &= H(D_1) - H(D_1|w(\boldsymbol{n}))\\
&\approx h(p(y)) - \frac{1}{2}[h(p(y) + \delta p'(y)) + h(p(y) - \delta p'(y))] \qquad (3)\\
&\approx -\frac{(\delta p'(y))^2}{2} h^{(2)}(p(y))\\
&= \frac{(\delta p'(y))^2}{2} \frac{1}{p(y)(1 - p(y)) \ln 2}. \qquad (4)
\end{aligned}$$

Let us consider the more general case $i \geq 1$. Let

$$\begin{aligned}
F_s(x) &= (p(y) + x)^s (1 - p(y) - x)^{i-s},\\
B(x) &= -\sum_{0 \leq s \leq i} \binom{i}{s} F_s(x) \log\left(1 + \frac{F_s(-x)}{F_s(x)}\right),
\end{aligned}$$

and let its Taylor sequence be

$$B(x) = \sum_{j \geq 0} B^{(j)}(0) \frac{x^j}{j!}.$$

Then the following theorem holds. For its proof we refer to the appendix.

Theorem 1. *Assuming that equalities hold in (2) we have*

$$I(D^i; w(\boldsymbol{n})) = \sum_{j \geq 1} B^{(2j)}(0) \frac{(\delta p'(y))^{2j}}{(2j)!},$$

where

$$B^{(2)}(0) = \frac{i}{p(y)(1-p(y)) \ln 2}.$$

Hence,

$$I(D^i; w(\boldsymbol{n})) = i \cdot I(D^1; w(\boldsymbol{n})) + \sum_{j \geq 2} B^{(2j)}(0) \frac{(\delta p'(y))^j}{(2j)!}.$$

The system designer of the copy protection scheme wants to design $p(y)$ such that $I(D^i; w(\boldsymbol{n}))$ is as small as possible given that $p(y) = 0$ for $y \leq y_1$ and $p(y) = 1$ for $y \geq y_2$. We notice that the size of interval $[y_1, y_2]$ is related to the reliability of the detector, the smaller the interval the better the reliability. So, in practise the system designer chooses firstly the size of this interval such that the reliability of the detector will be in a reasonable range. Secondly, the system designer constructs an optimal function $p(y)$ (optimal in the sense that $I(D^i; w(\boldsymbol{n}))$ is as small as possible).

We notice that for a fixed function $p(.)$, $I(D^i; w(\boldsymbol{n}))$ solely depends on the value y. Therefore we define

$$I_i(y) = I(D^i; w(\boldsymbol{n}))$$

and we infer from Theorem 1 and (4) that a first order approximation gives

$$I_i(y) \approx \frac{2i\delta^2}{\ln 2} \left\{ \frac{(p'(y))^2}{1 - (2p(y) - 1)^2} \right\}.$$

By substituting

$$p(y) = \frac{1}{2} - \frac{1}{2} \cos(r(y)) \tag{5}$$

with $r(y) = 0$ for $y \leq y_1$ and $r(y) = \pi$ for $y \geq y_2$ we obtain

$$I_i(y) \approx \frac{2i\delta^2}{\ln 2} \frac{(r'(y))^2}{4}.$$

Hence,

$$|r'(y)| \approx \frac{\sqrt{2 \ln 2 I_i(y)/i}}{\delta}, \tag{6}$$

where $I_i(y)/i$ is the information leakage expressed in watermark bits per experiment. The system designer wants to have

$$\sup_y I_i(y)/i$$

as small as possible. The requirement that $\sup_y I_i(y)/i$ is as small as possible is equivalent to the requirement that $\sup_y |r'(y)|$ is as small as possible, see (6). We conclude that $r(y)$ linearly increases in the interval $[y_1, y_2]$. Thus

$$r(y) = \pi \frac{y - y_1}{y_2 - y_1} \tag{7}$$

and $\pi/(y_2 - y_1) \approx (\sqrt{2 \ln 2 I_i(y)/i})/\delta$, that is

$$I_i(y)/i \approx \frac{\pi^2}{2 \ln 2} \left\{ \frac{\delta}{y_2 - y_1} \right\}^2 \tag{8}$$

is the information leakage expressed in watermark bits per experiment. We have constructed the optimal shape of $p(y)$ and we conclude that the information leakage, expressed in watermark bits per experiment, decreases quadratically in the size of the decision interval. We notice that the reliability of the watermarking scheme gets worse (higher probabilities of missed detection and false alarm) if the size of the decision interval increases.

We have analysed a first order approximation of an optimal shape for $p(y)$. This means that (8) gives a first order approximation of the information leakage. A better approximation (actually an upper bound) is given by the next theorem. Its proof is presented in the appendix.

Theorem 2. *Assuming that equalities hold in (2) and that $p(y)$ is defined by equations (5) and (7) we have that*

$$I(D^i; w(\boldsymbol{n})) \leq i \cdot I$$

with

$$I = 1 - h\left(\frac{1}{2} - \frac{\delta\pi}{2(y_2 - y_1)}\right) \approx \frac{\pi^2}{2 \ln 2} \left\{ \frac{\delta}{y_2 - y_1} \right\}^2 \approx I(D^1; w(\boldsymbol{n}))$$

if $\delta/(y_2 - y_1) < 1/\pi$, and $I = 1$ if $\delta/(y_2 - y_1) \geq 1/\pi$. Here $y_2 - y_1$ is the transition width of the decision interval and $\Delta y \in \{+\delta, -\delta\}$ is the effect that modifying one pixel has on the decision variable. Parameter I expresses the information leakage in watermark bits per experiment. The reliability of the watermarking scheme gets worse if the size of the decision interval increases.

6 Discussion

Example 1. Let us consider a digitized representation of a television frame in the NTSC standard, having $N = N_1 \times N_2 = 480$ by 720 pixels, with $w(\boldsymbol{n}) = \pm 1$.

Then $R_w(w) = 345600$. A useful choice of detection thresholds can be $y_1 = 115200$ and $y_2 = 230400$. If the luminance is quantized into 8 bits $(0, \ldots, 255)$ one pixel test t can influence the decision variable y by at most $\delta = 255$ but a more realistic value is $\delta \approx 100$ relative to mid grey. In such case, $I = 5.4 \cdot 10^{-6}$ bits per test. So recovering the full watermark is 186000 times more difficult than without the randomized decision threshold.

In an attempt to increase δ the attacker may use a different base $\{t_0, t_1, \ldots, t_{N-1}\}$ (in the previous example and in Section 5 $t_j(\boldsymbol{n_j}) = \alpha$ and $t_j(\boldsymbol{n_m}) = 0$ for $m \neq j$). The effect of t_j on the decision variable is $R_w(t_j) = \sum_{\boldsymbol{n} \in A} w(\boldsymbol{n}) t_j(\boldsymbol{n})$. Notice that $w = \{w(\boldsymbol{n})\}_{\boldsymbol{n} \in A}$ is a random variable to the attacker and that the expected effect of t_j on the decision variable is $\mathrm{E}[R_w(t_j)] = 0$. For a spectrally white watermark, i.e., if $\mathrm{E}[w(\boldsymbol{n}) w(\boldsymbol{n} + \Delta)] = 0$ for $\Delta \neq 0$, we find the second moment

$$\begin{aligned} \mathrm{E}[R_w(t_j)^2] &= \sum_{\boldsymbol{n} \in A} \sum_{\boldsymbol{n} + \Delta \in A} \mathrm{E}[w(\boldsymbol{n}) w(\boldsymbol{n} + \Delta)] t_j(\boldsymbol{n}) t_j(\boldsymbol{n} + \Delta) \\ &= k^2 \sum_{\boldsymbol{n} \in A} t_j(\boldsymbol{n})^2 = k^2 E_{t_j}, \end{aligned} \tag{9}$$

where $E_{t_j} = \sum_{\boldsymbol{n} \in A} t_j(\boldsymbol{n})^2$ is the energy in the test image t_j. Experiments with test image t_j reveal information about the value of $R_w(t_j)$ which gives us a linear relationship between the values of $w(\boldsymbol{n})$, $\boldsymbol{n} \in A$. We define the expected information leakage $I_i(y)$ in i experiments expressed in watermark bits by $I_i(y) = I(D^i; w) \approx i \cdot I(D^1; w) = i \cdot (H(D_1) - H(D_1|w))$ (notice that $I(D^i; w) = I(D^i; w(\boldsymbol{n_j}))$ if t_j is non-zero only in one pixel $\boldsymbol{n_j}$). See (2),

$$p(R_w(q_i)) \approx p(y) + R_w(t_j) p'(y). \tag{10}$$

Hence, $H(D_1) = h(\mathrm{E}[p(y) + R_w(t_j) p'(y)]) = h(p(y))$ and

$$\begin{aligned} H(D_1|w) &= \sum_{\hat{w}} Pr(w = \hat{w}) H(D_1 | w = \hat{w}) = \sum_{\hat{w}} Pr(w = \hat{w}) h(p(y) + R_w(t_j) p'(y)) \\ &= \mathrm{E}[h(p(y) + R_w(t_j) p'(y))] \\ &= (\mathrm{E}[h(p(y) + R_w(t_j) p'(y))] + \mathrm{E}[h(p(y) - R_w(t_j) p'(y))])/2. \end{aligned}$$

We obtain that the information leakage expressed in watermark bits per experiment equals

$$\begin{aligned} I_i(y)/i &\approx \mathrm{E}[h(p(y)) - (h(p(y) + R_w(t_j) p'(y)) + h(p(y) - R_w(t_j) p'(y)))/2] \\ &\approx \mathrm{E}[\pi^2 R_w(t_j)^2 / (2 \ln 2 (y_2 - y_1)^2)], \text{ see (4)}, \\ &= \frac{\pi^2}{2 \ln 2} \left\{ \frac{k \sqrt{E_{t_j}}}{y_2 - y_1} \right\}^2, \text{ see (9)}. \end{aligned}$$

We have generalized (8) towards this new setting. Notice that for large E_{t_j} approximation (10), and hence the generalized formula, is not accurate anymore.

We conclude that for large E_{t_j} the attacker gains substantial information. However, large E_{t_j} is not suitable for watermark detection methods where $S_=$ is not a perfect hyperplane. Then $q_l + t_j$ would be influenced too much by t_j because of its large energy E_{t_j}.

7 Concluding remarks

Electronic watermarks are a useful technical mechanism to protect Intellectual Property Rights. The use of watermarks in copy control for consumer electronic products, however, is not yet fully understood. We have investigated the sensitivity attack. The proposed countermeasure increases the workload by orders of magnitude, but the workload remains linear in the number of pixels.

In [1] a sensitivity attack is described that shows that if a watermark detection algorithm could be placed in a perfectly tamperproof box, this does not necessarily imply that the attacker cannot find a method to remove the watermark. This result questions the possibility to build perfect "public" watermarking schemes in which that attacker knows how to detect a watermark, but despite this knowledge he cannot remove or alter the watermark. A necessary condition for such system to be secure is that it should withstand the attack described here. Knowledge of the detection algorithm implies that the attacker can use the detector as an oracle to gain information about the watermark. As the attack proves, this is often sufficient to remove the watermark pixel by pixel. If the attack, or a more sophisticated elaboration of it, is successful against a black-box watermark detector, it would certainly be able to remove a watermark for which the attacker has the full details of the detection algorithm. All watermarking methods known to the authors are of the secret-key type, i.e., the watermark detector contains secret information, which could be exploited by an attacker to remove the watermark.

A Proofs

A.1 Proof of Theorem 1

Assuming that equalities hold in (2) we will prove

$$I(D^i; w(\boldsymbol{n})) = \sum_{j\geq 1} B^{(2j)}(0)\frac{(\delta p'(y))^j}{(2j)!}.$$

For realizations $d^i = (d_1, \ldots, d_i)$ with $s(d^i) = |\{l : d_l = 1\}|$ we have that

$$\begin{aligned}
Pr(D^i = d^i | w(\boldsymbol{n}) = +k) &= F_{s(d^i)}(+\delta p'(y)),\\
Pr(D^i = d^i | w(\boldsymbol{n}) = -k) &= F_{s(d^i)}(-\delta p'(y)),\\
Pr(D^i = d^i) &= (F_{s(d^i)}(+\delta p'(y)) + F_{s(d^i)}(-\delta p'(y)))/2.
\end{aligned}$$

By definition of entropy and conditional entropy

$$
\begin{aligned}
H(D^i) &= -\sum_{d^i} Pr(D^i = d^i) \log Pr(D^i = d^i), \\
H(D^i|w(\boldsymbol{n})) &= (H(D^i|w(\boldsymbol{n}) = +k) + H(D^i|w(\boldsymbol{n}) = -k)/2 \\
&= -\frac{1}{2}\sum_{d^i} Pr(D^i = d^i|w(\boldsymbol{n}) = +k) \log Pr(D^i = d^i|w(\boldsymbol{n}) = +k) + \\
&\quad -\frac{1}{2}\sum_{d^i} Pr(D^i = d^i|w(\boldsymbol{n}) = -k) \log Pr(D^i = d^i|w(\boldsymbol{n}) = -k).
\end{aligned}
$$

By combining all equations and noticing that $B(0) = -1$ we obtain

$$
\begin{aligned}
I(D^i; w(\boldsymbol{n})) &= -\frac{1}{2}\sum_{d^i} F_{s(d^i)}(+\delta p'(y)) \log \frac{1}{2}\left(1 + \frac{F_{s(d^i)}(-\delta p'(y))}{F_{s(d^i)}(+\delta p'(y))}\right) + \\
&\quad -\frac{1}{2}\sum_{d^i} F_{s(d^i)}(-\delta p'(y)) \log \frac{1}{2}\left(1 + \frac{F_{s(d^i)}(+\delta p'(y))}{F_{s(d^i)}(-\delta p'(y))}\right) \\
&= -[B(0) - \frac{1}{2}\{B(+\delta p'(y)) + B(-\delta p'(y))\}] \\
&= \sum_{j \geq 1} B^{(2j)}(0) \frac{(\delta p'(y))^{2j}}{(2j)!}.
\end{aligned}
$$

Straightforward, but lengthy, computations give the desired expression for $B^{(2)}(0)$.

A.2 Proof of Theorem 2

We notice that $D_l \leftarrow w(\boldsymbol{n}) \rightarrow D^{l-1}$ is a Markov sequence since D_l and D^{l-1} only depend on each other because of their relation towards $w(\boldsymbol{n})$. Therefore we may conclude that

$$
\begin{aligned}
I(D_l; w(\boldsymbol{n})|D^{l-1}) &= H(D_l|D^{l-1}) - H(D_l|w(\boldsymbol{n})D^{l-1}) \\
&= H(D_l|D^{l-1}) - H(D_l|w(\boldsymbol{n})) \\
&\leq H(D_l) - H(D_l|w(\boldsymbol{n})) = I(D_l; w(\boldsymbol{n})).
\end{aligned}
$$

Hence, by using (3)

$$
I(D^i; w(\boldsymbol{n})) = \sum_{l=1}^{i} I(D_l; w(\boldsymbol{n})|D^{l-1}) \leq \sum_{l=1}^{i} I(D_l; w(\boldsymbol{n})) = ig(p(y), \delta p'(y)),
$$

where for $0 \leq z \leq x \leq 1 - z \leq 1$

$$
g(x, z) = h(x) - \frac{1}{2}[h(x + z) + h(x - z)].
$$

Let us do some function research for $g(x,z)$ seen as function of x in the by us considered interval $[z, 1-z]$. We notice that

$$\frac{d}{dx}g(x,z) = h'(x) - \frac{1}{2}[h'(x+z) + h'(x-z)],$$

thus $\frac{d}{dx}g(x,z)|_{x=1/2} = 0$. Further $\frac{d}{dx}g(x,z)|_{x=z} = h'(z) - h'(2z)/2 = \log((1-z)/z) - \log((1-2z)/2z) = \log((2-2z)/(1-2z)) > 0$ and similarly $\frac{d}{dx}g(x,z)|_{x=1-z} < 0$. For function $\frac{d}{dx}g(x,z)$ we compute

$$\begin{aligned}\ln 2 \frac{d^2}{dx^2}g(x,z) &= \frac{-1}{x(1-x)} + \frac{1}{2}\left[\frac{1}{(x+z)(1-x-z)} + \frac{1}{(x-z)(1-x+z)}\right] \\ &= \frac{-1}{x(1-x)} + \frac{x(1-x)+z^2}{(x^2-z^2)((1-x)^2-z^2)}.\end{aligned}$$

This appears to be ≤ 0 iff $1/2 - \sqrt{z^2 - 3/4} \leq x \leq 1/2 + \sqrt{z^2 - 3/4}$. We conclude that $\frac{d}{dx}g(x,z) > 0$ if $x < 1/2$, $\frac{d}{dx}g(x,z) = 0$ if $x = 1/2$, and $\frac{d}{dx}g(x,z) > 0$ if $x > 1/2$. Hence, we have that $g(x,z)$ is maximal for $x = 1/2$.

We notice that

$$\frac{d}{dz}g(1/2,z) = \log\frac{1/2+z}{1/2-z} > 0.$$

Hence, $g(1/2,z)$ is increasing in z. We have that $p((y_2-y_1)/2) = 1/2$ and $\delta p'(y) \leq \delta p'((y_2-y_1)/2) = \delta\pi/(2(y_2-y_1))$. Hence, $g(p(y), \delta p'(y))$ is maximal for $y = (y_2-y_1)/2$ and we have that

$$g(p(y), \delta p'(y)) \leq 1 - h(1/2 - \delta\pi/(2(y_2-y_1))).$$

References

1. I.J. Cox and J.M.P.G. Linnartz. "Public watermarks and resistance to tampering". ICIP 97.
2. W. Bender, D. Gruhl, N. Morimoto, and A. Lu. "Techniques for data hiding". *IBM Systems Journal*, Vol. 35.(3/4), 1996.
3. I.J. Cox, J. Kilian, F.T. Leighton and T. Shamoon. "A secure, robust watermark for multimedia". In *Information Hiding: First Int. Workshop Proc., Lecture Notes in Computer Science*, volume 1174, R. Anderson, ed., Springer-Verlag, pages 185–206, 1996.
4. J.P.M.G. Linnartz, A.C.C. Kalker, G.F. Depovere, and R. Beuker. "A reliability model for detection of electronic watermarks in digital images". In *Proc. Benelux Symposium on Communication Theory, Enschede, October*, pages 202–208, 1997.
5. I. Pitas and T.H. Kaskalis. "Signature casting on digital images". In *Proc. IEEE Workshop on Nonlinear Signal and Image Processing, Neos Marmaras, June*, 1995.
6. J.R. Smith and B.O. Comiskey. "Modulation and information hiding in images". In *Information Hiding: First Int. Workshop Proc., Lecture Notes in Computer Science*, volume 1174, R. Anderson, ed., Springer-Verlag, pages 207–226, 1996.
7. A.B. Watson. "*Digital Images and Human Vision*". The MIT Press, 1993.
8. T.M. Cover and J.A. Thomas. "*Elements of Information Theory*". John Wiley and Sons, Inc., 1991.