Modelling the false alarm and missed detection rate for electronic watermarks

Jean-Paul Linnartz, Ton Kalker, Geert Depovere.

Philips Research Laboratories
Prof. Holstlaan 4, WY8, 5656 AA Eindhoven, The Netherlands
e-mail: (linnartz, kalker, depovere)@natlab.research.philips.com

Abstract. Theoretical modeling of watermarks allow prediction of the detector reliability and facilitates the development of more reliable systems. In particular, mathematical evaluation is relevant to estimate the rate at which "false alarms" occur. In this paper, the probability of incorrect detection (missed detection or false alarm) is expressed in terms of the watermark-energy-to-image-luminance-variance ratio. We present some counterintuitive results which show for instance that the reliability of detection significantly depends on spatial correlation in watermark. Moreover we find that a small but uncompensated random DC component in the watermark can have a significant effect on the reliability.

I Background

New multi-media networks and services facilitate the dissemination of audio and video content, but at the same time make illegal copying and copyright piracy simple. This has created a need to embed copyright data in the content in an indelible way. Particularly if watermark detection is part of an active copy control concept on Consumer Electronics (CE) and PC platforms, typical requirements include: [1], [2], [3], [4], [5], and [6]

1. Erasing or altering the watermark should be difficult.
2. The watermarking scheme should be robust against typical transmission and storage imperfections (such as lossy compression, noise addition, format conversion, bit errors) and signal processing artefacts (noise reduction, filtering), even if such operations are intended to erase the watermark.
3. It should be robust against typical attacks, e.g. those described in [5].
4. False alarms, i.e., positive responses for content that does not contain a watermark should not occur (orders of magnitude) more often than electronic or mechanical product failures. CE devices or PC-s should not fail to work due to an erroneously triggered watermark detector.
5. The watermark should be unobtrusive, and not be annoying to bona-fide users.

The "low false alarm" requirement appears to be too stringent to determine the error rates only by experiments. This has been our motivation to develop a mathematical model for the reliability of watermark detectors.

The aim of this paper is to contribute to the understanding and modelling of the reliability of watermark detectors. This involves the development of a mathematical framework and verification of critical assumptions.

The organization of the paper is as follows. Section II models the image (Section II.1) and the watermark (Section II.2- II.4). The analysis presented in this paper requires a detailed definition of the DC-component and the spatial correlation of the watermark, which we include in Section II.3 and II.4, respectively. Section III discusses the detector. The reliability of a generic correlator is derived, and special cases are dealt with. A few counterintuitive results are obtained, discussed and verified. Numerical results are plotted in Section IV, and verified by experiments. Section V concludes this paper.

II Formulation of the Model

Our model extends previous work, such as [1], [2], [3], and [6] to regard the image as noise or interference during the detection of a weak wanted signal (namely the watermark). However, we consider spatial correlation properties of the image to be known to a large extent and we address (pseudo-) randomness in the watermark generation.

II.1 Image Model

We consider a rectangular image of size $N = N_1 N_2$ pixels. The (gray level or) luminance of the pixel with coordinates $\mathbf{n} = (n_1, n_2)$, $(0 \leq n_1 \leq N_1 - 1, 0 \leq n_2 \leq N_2 - 1)$ is denoted as $p(\mathbf{n})$. We denote $\mathbf{0} = (0,0)$, $\mathbf{e}_1 = (1,0)$ and $\mathbf{e}_2 = (0,1)$, so $\mathbf{n} = n_1\mathbf{e}_1 + n_2\mathbf{e}_2$. The set of all pixel coordinates is denoted as A_N, where

$$A_N = \{\mathbf{n} : 0 \leq n_1 \leq N_1 - 1, 0 \leq n_2 \leq N_2 - 1\}.$$

In color pictures, $p(\mathbf{n})$ is a YUV or RGB vector, but for the sake of simplicity we restrict our discussion to the luminance of the image, in which $p(\mathbf{n})$ takes on real or integer values in a certain interval.

The k-th sample moment of the gray level of each pixel is denoted as $\mu_k = \mathrm{A}[p^k(\mathbf{n})] = \frac{1}{N}\sum_{\mathbf{n}\in A_N} p^k(\mathbf{n})$, where A is a spatial averaging operator over area A_N. In particular, μ_1 represents the average value or "DC-component" in a pixel and $\mu_2 = \mathrm{A}[p^2]$ represents the average power in a pixel and $E_p = N\mu_2$ is the total energy in an image. The sample variance is $\sigma^2 = \mathrm{A}[p(\mathbf{n}) - \mu_1]^2 = \mu_2 - \mu_1^2$.

We assume that the image has homogeneous statistical properties (wide-sense spatial stationarity), so the spatial correlation only depends on the difference vector $\boldsymbol{\Delta}$. We define

$$\Gamma_{p,p}(\boldsymbol{\Delta}) = \frac{1}{N}\sum_{\mathbf{n}\in A_N} p(\mathbf{n})p(\mathbf{n}+\boldsymbol{\Delta}),$$

In order to make the evaluation of our examples tractable, we simplify the image model assuming the first-order separable autocorrelation function (acf) [10], [7], [8]

$$\Gamma_{p,p}(\Delta) = \mu_1^2 + \sigma^2\alpha^{|\Delta|}$$

where we defined $|\Delta| = |\Delta_1| + |\Delta_2|$. Here α can be interpreted as a measure of the correlation between adjacent pixels in the image. Experiments, e.g. in [6] reveal that typically $\alpha \approx 0.9...0.99$. We denote $\tilde{p}(\mathbf{n})$ as the non-DC components of the image, i.e., $p(\mathbf{n}) = \mu_1 + \tilde{p}(\mathbf{n})$, so $\Gamma_{\tilde{p},\tilde{p}} = \sigma^2 \alpha^{|\Delta|}$.

Some of the above assumptions seem a crude approximation of the typical properties of images. From experiments such as those to be reported in section V, it will appear that estimates based on this simplification are nonetheless reasonably accurate for our purpose. The accuracy of the model will be verified by measuring μ, σ and α from images and using these parameters in a theoretical evaluation, which we compare with purely experimental results. These assumptions, however, exclude certain images, such as binary images, line art or computer-generated graphics with a limited number of grey levels.

II.2 Watermark Model

The watermark is represented by $w(\mathbf{n})$ which takes on real values in all pixels $\mathbf{n} \in A_N$. A watermark detector has to operate on the observed (marked or unmarked) image $q(\mathbf{n})$. We aim at detecting whether a particular watermark is present or not, based only on the knowledge of $w(\mathbf{n})$. In copy control applications, the unmarked original $p(\mathbf{n})$ is not available at the detector. Watermarked images have similar properties as unmarked images, except that perceptually invisible modifications have been made. We assume $q(\mathbf{n})$ to provide sufficiently reliable estimates of the properties (μ_k and $\Gamma_{p,p}$) of $p(\mathbf{n})$.

The watermark $w(\mathbf{n})$ is embedded in the image. Typically, $q(\mathbf{n}) = p(\mathbf{n}) + \theta(\mathbf{n})w(\mathbf{n})$, where we do not yet specify the embedding depth $\theta(\mathbf{n})$. In the analysis we will focus on detection, thus simplify the embedding process by taking $\theta(\mathbf{n}) \equiv 1$ for all $\mathbf{n} \in A_N$.

Our model implicitly assumes that no spatial transformation of the image (resizing, cropping, rotation, etc.) is conducted. Such transformation may require a search during detection, which is outside the scope of this analysis.

In the following, we will not consider a particular, fully described watermark but a class of watermarks having specific spatial properties. In practice, this could be the set of all watermarks that are generated by a certain algorithm, but with different seeds for the pseudo-random generator.

For two watermarks w_1 and w_2 out of such class, the (deterministic) spatial inner product is

$$\Gamma_{w_1,w_2}(\boldsymbol{\Delta}) = \frac{1}{N} \sum_{\mathbf{n} \in A_N} w_1(\mathbf{n}) w_2(\mathbf{n} + \boldsymbol{\Delta}),$$

where we assume for simplicity that $\mathbf{n} + \boldsymbol{\Delta}$ wraps around when it formally falls outside the set A_N. The total energy in the watermark equals

$$E_w = \sum_{\mathbf{n} \in A_N} w^2(\mathbf{n}) = N \Gamma_{w,w}(\mathbf{0}).$$

If we consider an ensemble of many watermarks generated by a particular watermark generation algorithm, the statistical correlation is defined as

$$R_{w_1,w_2}(\boldsymbol{\Delta}) = \mathrm{E}[w_1(\mathbf{n})w_2(\mathbf{n}+\boldsymbol{\Delta})]$$

For virtually all watermark generators that we are aware of, the expected value of a measured $\Gamma_{w,w}$ equals $R_{w,w}$ with negligible small deviations if the size of the watermark set A_N is sufficiently large.

Such a property does *not* hold for images, where due to the lack of ergodicity, it is unlikely that the spatial correlation $\Gamma_{p,p}$ of a particular image converges in mean-square to the statistical correlation $R_{p,p}$ over acollection of different images.

II.3 DC components

The DC content of the watermark is defined as $D_0 = \mathrm{A}[w(n)] = \frac{1}{N}\sum_{\mathbf{n}\in A_N} w(\mathbf{n})$. An individual watermark is DC-free iff $D_0 = 0$. This cases has been addressed extensively by Pitas [1] and others.

For an arbitrary value of D_0, we observe that

$$\begin{aligned} N^2 D_0^2 &= \textstyle\sum_{\mathbf{n}\in A_N}\sum_{\mathbf{k}\in A_N} w(\mathbf{n})w(\mathbf{k}) = E_w + \sum_{\mathbf{n}\in A_N}\sum_{\mathbf{k}\in A_N,\mathbf{k}\neq\mathbf{n}} w(\mathbf{n})w(\mathbf{k}) \\ &= E_w + \textstyle\sum_{\mathbf{n}\in A_N}\sum_{\boldsymbol{\Delta}\neq\mathbf{0}} w(\mathbf{n})w(\mathbf{n}+\boldsymbol{\Delta}) = E_w + N\sum_{\boldsymbol{\Delta}\neq\mathbf{0}} \Gamma_{w,w}(\boldsymbol{\Delta}) \end{aligned} \tag{1}$$

This implies that a designer who desires to choose D_0 and E_w must face restrictions on the spatial correlation $\Gamma_{w,w}$.

In practice, a watermark can for instance be created by randomly generating a $+k$ or $-k$ pixel value independently for each pixel $\mathbf{n}$. Then, D_0 is a random variable with zero-mean ($\mathrm{E}D_0 = 0$) and positive variance. Thus, each individual watermark does not necessarily have a zero DC component. We call a *watermark generation process* or a set of watermarks to be "statistically DC-free" or "DC-free in the mean" iff $\mathrm{E}D_0 = 0$. This is a necessary, but not a sufficient condition for all individual watermarks to be DC-free. When the generation process is DC-free with probability one, we call it absolutely DC-free, or adopting a term used in probability theory [15], we write "(almost) surely" (a.s.) DC-free.

II.4 Watermark Spectrum

Quasi-white watermarks: For a watermark with a given D_0, one can consider the class of (quasi-) white watermarks, which satisfy $\Gamma_{w,w}(\boldsymbol{\Delta_1}) = \Gamma_{w,w}(\boldsymbol{\Delta_2}) = \eta$ for $\boldsymbol{\Delta_1}, \boldsymbol{\Delta_2} \neq \mathbf{0}$, where η is some constant ($|\eta| << E_w/N$). In such case the spatial autocorrelation resembles a δ-function with a peak of amplitude E_w/N at $\boldsymbol{\Delta} = \mathbf{0}$. For $\boldsymbol{\Delta} \neq \mathbf{0}$,

$$\Gamma_{w,w}(\boldsymbol{\Delta}) = \eta = (N^2 D_0^2 - E_w)/(N(N-1)) < 0.$$

It can be shown that the corresponding spectral energy density is flat, except for a DC-component.

In particular, we see that for a DC-free watermark ($D_0 = 0$) with non-zero energy ($E_w > 0$), the watermark values $w(\mathbf{n_i})$ and $w(\mathbf{n_j}), \mathbf{n_i} \neq \mathbf{n_j}$ cannot be statistically uncorrelated, because $\eta < 0$. *

We will call a watermark generation process "quasi white and DC-free a.s." if $D_0 = 0$ a.s. and its autocorrelation function is

$$\Gamma_{w,w}(\boldsymbol{\Delta}) = \begin{cases} E_w/N & \text{if } \boldsymbol{\Delta} = \mathbf{0} \\ \frac{N^2 D_0^2 - E_w}{N(N-1)} & \text{if } \boldsymbol{\Delta} \neq \mathbf{0} \end{cases}$$

In the performance analysis, we will mainly use the statistical correlation $R_{w,w}$ over the ensemble of watermarks rather than the deterministic $\Gamma_{w,w}$. The behavior of R can be shown to be similar to that of Γ.

Let's consider some pixel $\mathbf{n_0}$ with a non-zero watermark value $w(\mathbf{n_0}) = k_0$. This implies that the $N-1$ other pixels in the image must compensate for this through

$$\sum_{n_i \in A_N \setminus \mathbf{n_0}} w(\mathbf{n_i}) = ND_0 - k_0$$

For a quasi-white watermark generation process, we define $R_{w,w}(\boldsymbol{\Delta_1}) = R_{w,w}(\boldsymbol{\Delta_2}) = \eta_R$ for $\boldsymbol{\Delta_1}, \boldsymbol{\Delta_2} \neq \mathbf{0}$. We find

$$\mathrm{E}[w(\mathbf{n_i})|w(\mathbf{n_0}) = k_0] = (ND_0 - k_0)/(N-1),$$

so, for $\boldsymbol{\Delta} \neq \mathbf{0}$,

$$R_{w,w}(\boldsymbol{\Delta}) = \mathrm{E}[\mathrm{E}[w(\mathbf{n_0})w(\mathbf{n_0} + \boldsymbol{\Delta})|[w(\mathbf{n_0})]] = \mathrm{E}\left[w(\mathbf{n_0})[\frac{ND_0 - w(\mathbf{n_0})}{N-1}]\right]$$

We get

$$R_{w,w}(\boldsymbol{\Delta}) = \begin{cases} \frac{E_w}{N} & \text{if } \boldsymbol{\Delta} = \mathbf{0} \\ N\frac{D_0^2}{N-1} - \frac{E_w}{N(N-1)} & \text{if } \boldsymbol{\Delta} \neq \mathbf{0} \end{cases}.$$

A "purely white" watermark generation process requires that the correlation equals exactly zero except at $\boldsymbol{\Delta} = \mathbf{0}$, where $R_{w,w}(\mathbf{0}) = E_w/N$. We have seen that purely white watermarks cannot be absolutely DC free, but have $D_0 = \sqrt{E_w}/N$

Low pass watermark As an other example, we will treat the case that the watermark has a low-pass spatial spectrum. This method has been advocated for instance by Cox et al. [4]. In such situation, a potential attacker can less easily tamper with the watermark by low-pass filtering. Moreover, JPEG compression typically removes or distorts high-frequency components. A low-pass watermark can be generated by spatially filtering a (quasi-) white watermark. Perceptually

* A similar small negative correlation outside the origin ($\boldsymbol{\Delta} \neq \mathbf{0}$) is often ascribed to a peculiarity of maximum-length pseudo-random sequences, as generated by a Linear Feedback Shift Registers (LFSR). However, the above argument reveals that it is fundamentally related to the requirement of the DC value. See also [13] and [14].

this appears as a smoothing. A first-order two dimensional IIR spatial smoothing filter computes [11]

$$w_2(\mathbf{n}) = (1-\beta^2)^2 \left[w_1(\mathbf{n}) + \beta w_2(\mathbf{n}-\mathbf{e_1}) + \beta w_2(\mathbf{n}-\mathbf{e_2}) - \beta^2 w_2(\mathbf{n}-\mathbf{e_2}-\mathbf{e_2}) \right]$$

from an original w_1. It can be shown that in case of a purely white watermark w_1 at the input, such a first-order filter generates a new watermark w_2 with correlation function [11]

$$R_{w_2,w_2} = \frac{E_w}{N} \beta^{|\Delta|}$$

Another method of generating a low pass watermark is to use a pseudo random $\{-k, k\}$ generator which gives a statistically dependent output for neighboring pixels.

III Correlator detector

Correlation detectors are interesting to study, for several reasons. They are a mathematical generalization of the simple but nonetheless important scheme in which $w \in \{-1, 0, +1\}$. To discuss this subclass of correlators first, let's denote $A_- = \{\mathbf{n} : w(\mathbf{n}) = -1\}$ and $A_+ = \{\mathbf{n} : w(\mathbf{n}) = +1\}$. Watermarks are detected by computing the sum of all pixel luminances at locations where the watermark is negative, i.e., $s_- = \sum_{\mathbf{n}\in A_-} q(\mathbf{n})$ and the sum of all luminances where the watermark is positive, i.e., $s_+ = \sum_{\mathbf{n}\in A_+} q(\mathbf{n})$. Then, an expression such as $y = (s_+ - s_-)/N$ is used as a decision variable. See for instance [1], [2] for schemes that are related or can be shown to be equivalent. From our more general results to follow it can be concluded that that two aspects have a significant effect on performance.

[Spatial Correlation] Spatial correlation occurs if the probability that $\mathbf{n_i} \in A_-$ statistically depends on whether $\mathbf{n_j} \in A_-$ for some pair of differing pixel locations $\mathbf{n_i} \neq \mathbf{n_j}$. We will see that high spatial correlation substantially reduces reliability.

[Compensation of DC components in the watermark] If the number of pixels in sets A_- and A_+ are generated as binomial random variables such that the *expected value* of the number of elements in both sets is identical ($\mathrm{E}D_0 = 0$, but the variance of $D_0 = O(N^{-1})$), the performance is significantly worse than when the number of elements is always precisely the same ($D_0 = 0$ a.s.). This is in contrast to our intuition that if pixels are put in A_- and A_+ with probability $1/2$ and independently of each other, the statistical effect of a differing number of elements in each class would become negligibly small for increasing image sizes. Our theoretical and experimental results refute this belief.

Another reason to address correlators (sometimes called "matched filters" [12]) is that these are known to be the *optimum* detector for a typical situation often encountered in radio communications, namely the Linear Time-Invariant (LTI), frequency non-dispersive, Additive Gaussian Noise (AWGN)

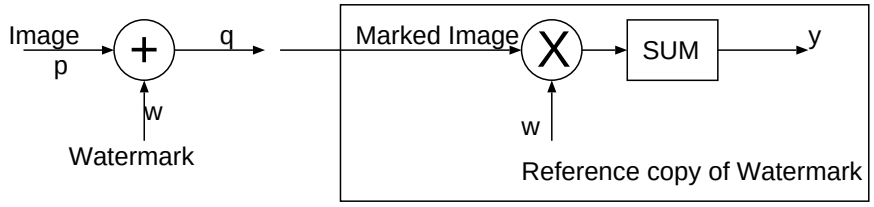

Fig. 1. Watermark Embedder and Correlation Detector

channel, when the receiver has full knowledge about the alphabet of waveforms used to transmit a message. Less ideal situations often are addressed with appropriate modifications of the matched filter.

In a correlator detector, a decision variable y is extracted from the suspect image $q(\mathbf{n})$ according to correlation with a locally stored copy of the watermark $\hat{w}(\mathbf{n})$ typically with $\hat{w}(\mathbf{n}) = C_d w(\mathbf{n})$, where w.l.o.g. we assume the constant C_d to be unity. The decision variable is $y = \Gamma_{\hat{w},q}(\mathbf{0})$, with

$$\Gamma_{\hat{w},q}(\boldsymbol{\Delta}) = \frac{1}{N} \sum_{\mathbf{n} \in A_N} \hat{w}(\mathbf{n}) q(\mathbf{n} + \boldsymbol{\Delta})$$

Figure 1 illustrates this correlation detector. The model covers all detectors in which the decision variable is a linear combination of pixel luminance values in the image. Hence, it is a generalization of many detectors proposed previously. It covers a broader class of watermarks then the binary ($w(\mathbf{n}) \in \{-k, k\}$) or ternary ($w(\mathbf{n}) \in \{-k, 0, k\}$) watermarks. In particular, it also includes methods in which the detection in conducted by correlation in the domain of DCT coefficients.

For our analysis, we separate y into a deterministic contribution y_w from the watermark, plus filtered "noise" from the image y_p.

$$y_w = \frac{1}{N} \sum_{\mathbf{n} \in A_N} \hat{w}(\mathbf{n}) \theta(\mathbf{n}) w(\mathbf{n})$$

Taking a uniform embedding depth $\theta(n) \equiv 1$, we get $y_w = \Gamma_{w,w}(\mathbf{0}) = \frac{E_w}{N}$ Moreover,

$$y_p = \frac{1}{N} \sum_{\mathbf{n} \in A_N} \hat{w}(\mathbf{n}) p(\mathbf{n})$$

Regarding y_p, the mean value is found as the product of the DC component in the watermark and the image, namely

$$\mathrm{E} y_p = \frac{1}{N} \mathrm{E} \sum_{\mathbf{n} \in A_N} \hat{w}(\mathbf{n}) p(\mathbf{n}) = \frac{1}{N} \left[\mathrm{E}\hat{w}(\mathbf{n})\right] \sum_{\mathbf{n} \in A_N} p(\mathbf{n}) = \mu_1 \mathrm{E}\hat{D}_0$$

Note that for a particular watermark with known DC-component, $\mathrm{E} y_p | D_0 = \mu_1 \hat{D}_0$. Up to this point, results are irrespective of the correlation in pixels.

The second moment is found as

$$\mathrm{E}y_p^2 = \mathrm{E}\left[\frac{1}{N}\sum_{\mathbf{n}\in A_N}\hat{w}(\mathbf{n})p(\mathbf{n})\right]^2 = \tag{2}$$

$$\frac{1}{N^2}\mathrm{E}\left[\sum_{\mathbf{n_i}\in A_N}\sum_{\mathbf{n_j}\in A_N}\hat{w}(\mathbf{n_i})p(\mathbf{n_i})\hat{w}(\mathbf{n_j})p(\mathbf{n_j})\right]$$

In the above expression it is tempting to assume that cross terms with $\mathbf{n_i} \neq \mathbf{n_j}$ all become zero or negligibly small for sufficiently large images. However, a DC component may be present. Furthermore, in the following sections we will show that for correlated pixels ($\alpha > 0$) and spectrally non-white watermarks (e.g., $\beta > 0$) , non-zero cross terms substantially affect the results, even if $D_0 = 0$. Therefore we will not make this assumption here.

Because of the Central Limit Theorem, y_p has a Gaussian distribution if N is sufficiently large and if the contributions in the sums are sufficiently independent. The Gaussian behaviour will be verified experimentally in Section IV. If we apply a threshold y_{thr} to decide that the watermark is present if $y > y_{thr}$, the probability of a *missed detection* (the watermark is present in $q(\mathbf{n})$, but the detector thinks it is not; false negative) is

$$P_{md} = \frac{1}{2}\mathrm{erfc}\frac{y_w - y_{thr} + \mathrm{E}y_p}{\sqrt{2}\sigma_{y_p}}$$

where σ_{y_p} is the standard deviation of y_p, with $\sigma_{y_p}^2 = \mathrm{E}y_p^2 - [\mathrm{E}y_p]^2$. We find

$$P_{md} = \frac{1}{2}\mathrm{erfc}\frac{E_w + \mu_1 N\mathrm{E}\hat{D}_0 - N_1 N_2 y_{thr}}{\sqrt{2}N\sigma_{y_p}}$$

The presence of D_0 and μ_1 in this expression suggest that either the DC-terms must be appropriately compensated in selecting y_{thr} or that the suspect image $q(\mathbf{n})$ must be preprocessed to subtract the DC-term μ_1.

Given that no watermark is embedded, a *false alarm* (false positive) occurs with probability

$$P_{fa} = \frac{1}{2}\mathrm{erfc}\frac{y_{thr} - \mathrm{E}y_p}{\sqrt{2}\sigma_{y_p}}$$

III.1 Example 1: (Quasi-) White and DC-free watermark

The quasi-white and DC-free watermark reasonably models most of the early proposals for increasing and decreasing the pixel luminance according to a pseudo random process. Using $p(\mathbf{n}) = \mu_1 + \tilde{p}(\mathbf{n})$, one can write

$$\mathrm{E}y_p^2 = \mu_1^2 D_0^2 + \tfrac{1}{N^2}\textstyle\sum_{\mathbf{n}\in A_N}\sum_{\boldsymbol{\Delta}:\mathbf{n}+\boldsymbol{\Delta}\in A_N}\mathrm{E}\left[\hat{w}(\mathbf{n})\hat{w}(\mathbf{n}+\boldsymbol{\Delta})\tilde{p}(\mathbf{n})\tilde{p}(\mathbf{n}+\boldsymbol{\Delta})\right]$$

For an a.s. DC-free watermark, the first term is zero. In the forthcoming evaluation, the image size N is considered to be large enough and α is assumed to be sufficiently smaller than unity to justify ignoring of boundary effects. To be more precise, we consider $R_{w,w}(\boldsymbol{\Delta})\Gamma_{\bar{p},\bar{p}}(\boldsymbol{\Delta})$ to vanish rapidly enough with increasing $\boldsymbol{\Delta}$ to allow the following approximation: we consider the sum over $\boldsymbol{\Delta}$ to cover the entire plane R^2 even though the size of the image is finite. This allows us to write

$$\sigma_{y_p}^2 = \mathrm{E}y_p^2 = \frac{1}{N}\sum_{\boldsymbol{\Delta}\in R^2} R_{w,w}(\boldsymbol{\Delta})\Gamma_{\bar{p},\bar{p}}(\boldsymbol{\Delta})$$

Thus

$$\mathrm{E}y_p^2 = \frac{E_w\sigma^2}{N^2} - \sum_{\boldsymbol{\Delta}\neq\mathbf{0}}\left[\frac{D_0^2}{N-1} - \frac{E_w}{N^2(N-1)}\right][\sigma^2\alpha^{|\Delta|}]$$

We use $\sum_{\boldsymbol{\Delta}\neq\mathbf{0}}\alpha^{|\Delta|} = [(1+\alpha)/(1-\alpha)]^2 - 1$ to express

$$\mathrm{E}y_p^2 = \frac{E_w\sigma^2}{N(N-1)} + \sigma^2\left[\frac{N^2D_0^2 - E_w}{N^2(N-1)}\right]\left[\frac{1+\alpha}{1-\alpha}\right]^2$$

The second term becomes negligible for large N. We see that the effect of pixel correlations is significant only if α is close to unity (little luminance changes) in a small-size image. If the image is large enough, that is, if $N >> [(1+\alpha)/(1-\alpha)]^2$, we may approximate

$$\mathrm{E}y_p^2 \approx \frac{E_w\sigma^2}{N^2}$$

Experiments of Section IV confirm that in practical situations this is a reasonable approximation, provided that the watermark is white. Inserting the value obtained for the standard deviation σ_{y_p}, the error probability becomes

$$P_{fa} = \frac{1}{2}\mathrm{erfc}\frac{Ny_{thr}}{\sqrt{2E_w}\sigma}$$

In Figure 2, we consider a DC-free watermark and $y_{thr} = \frac{E_w}{2N}$ which provides $P_{fa} = P_{md}$. In practice one would presumably like to improve P_{fa} at the cost of P_{md}, but this corresponds to a simple horizontal shift of the curve. We plot

$$P_{fa} = P_{md} = \frac{1}{2}\mathrm{erfc}\sqrt{\frac{E_w}{8\sigma^2}}$$

versus the watermark-energy-to-image-luminance-variance E_w/σ^2, expressed in dB i.e., $10\log_{10}(E_w/\sigma^2)$.

Defining the watermark-energy-to-image-luminance-variance as E_w/σ^2, thus as a signal-to-noise ratio, ensures that the curves and mathematical expressions become independent of the image size, its average luminance and sample variance. Moreover it matches common practice in statistical communication theory where one uses E_b/N_0, where E_b is the average energy per bit, and N_0 is the spectral power density of the noise.

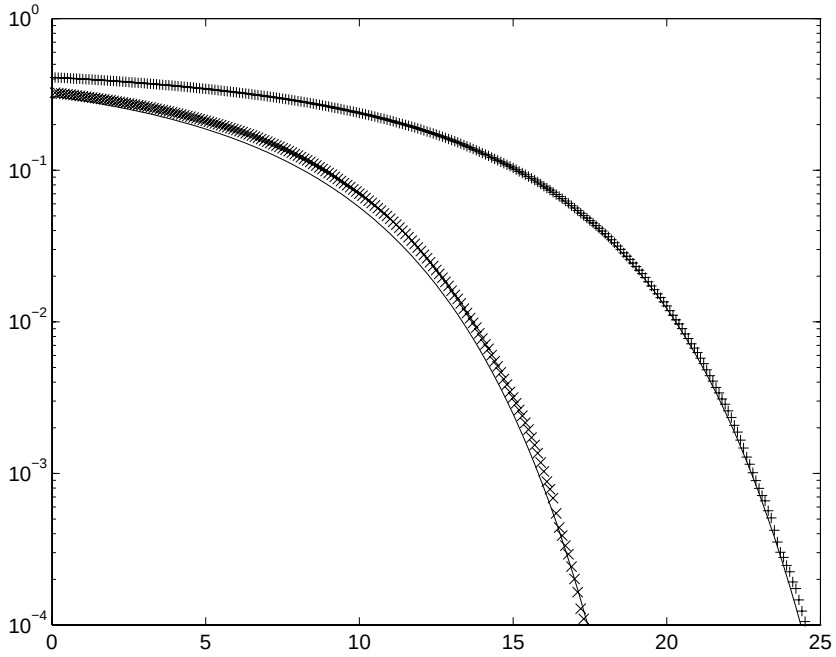

Fig. 2. Watermark detection error rates P_{fa} and P_{md} versus signal-to-noise ratio E_w/σ^2 for correlation detector. Experiments on "Lenna": "×": $D_0 = 0$ "+": random watermark, independent pixels, DC-free i.m. Solid lines: corresponding theoretical curves.

For reliable detection E_w/σ^2 is much larger than unity, but this does not imply that the watermark $w(\mathbf{n_i})$ in a particular pixel exceeds the luminance variance. This is similar to spread-spectrum radio where $E_b/N_0 >> 1$ despite the fact that spectrally spoken the signal power is below the broadband noise [12].

It appears to be less relevant over how many pixels the watermark energy is spread as long as the total energy E_w and the image properties (σ in particular) are fixed. However, example III.2.2 will show that if the watermark energy is not embedded by a spectrally white watermark, as assumed here, the results are different and do depend on the watermark "waveform" used. This is in contrast to spread-spectrum radio over AWGN channels.

III.2 Example 2: The effect of a non-zero DC component

This section evaluates the method mentioned in section II.3, II.4 and III to generate a watermark by randomly choosing $w(\mathbf{n}) \in \{-k, +k\}$ with probability $1/2$ independently for every pixel. The resulting w may or may not be filtered by a two-dimensional first-order filter to created suitable low-pass spatial properties.

Before any such filtering the number of elements in A_+ equals $N(1+D_0)/2$, which is a binomial random variable with mean $N/2$ and variance $N/4$. Thus D_0 is zero-mean with variance $1/N$.

Intuitively one may believe that if N is large enough, setting a fixed detection threshold but without specifically compensating for a small random D_0 will not affect the performance of the detector significantly. Here we will prove different and show that the standard deviation of D_0 remains significant for large N and does not vanish rapidly enough to be negligible. **

We address the ensemble-mean behavior. Averaging over all possible watermarks generated this way, we get

$$\mathrm{E}y_p^2 = \frac{1}{N^2} \sum_{\mathbf{n}\in A_N} \sum_{\boldsymbol{\Delta}:\mathbf{n}+\boldsymbol{\Delta}\in A_N} [\mathrm{E}w(\mathbf{n})w(\mathbf{n}+\boldsymbol{\Delta})p(\mathbf{n})p(\mathbf{n}+\boldsymbol{\Delta})]$$

If we ignore boundary effects, i.e., summing over $\boldsymbol{\Delta} \in R^2$, we get

$$\mathrm{E}y_p^2 = \frac{1}{N_1 N_2} \sum_{\boldsymbol{\Delta}\in R^2} R_{w,w}(\boldsymbol{\Delta}) \Gamma_{p,p}(\boldsymbol{\Delta})$$

Inserting the previously discussed correlation functions of a low-pass image and a low-pass watermark, one sees that

$$\mathrm{E}y_p^2 = \sum_{\boldsymbol{\Delta}\in R^2} \frac{E_w}{N^2} \beta^{|\Delta|} [\mu_1^2 + \sigma^2 \alpha^{|\Delta|}]$$

$$\mathrm{E}y_p^2 = \frac{E_w \sigma^2}{N^2} \left[\frac{1+\alpha\beta}{1-\alpha\beta}\right]^2 + \frac{E_w \mu_1^2}{N^2} \left[\frac{1+\beta}{1-\beta}\right]^2$$

and, as we saw before $\mathrm{E}y_p = \mu_1 \mathrm{E}D_0$, so

$$\mathrm{E}\sigma_{y_p}^2 = \frac{E_w \sigma^2}{N^2} \left[\frac{1+\alpha\beta}{1-\alpha\beta}\right]^2 + \frac{E_w \mu_1^2}{N^2} \left[\frac{1+\beta}{1-\beta}\right]^2 - \mu_1^2 \left[\mathrm{E}D_0\right]^2$$

The two first terms are order $O(N^{-2})$. For large N, the second term (which accounts for variations in the DC offset) does *not* vanish compared to the first term. This result is somewhat counterintuitive as it shows that the effect of statistical fluctuations in D_0 does *not* vanish fast enough if the watermark is laid over more pixels.

In the special case of white watermarks, i.e., for $\beta \to 0$, one would expect $\sigma_{y_p}^2 = E_w \sigma^2 / N^2$. However, it tends to $E_w \mu_2 / N^2$ where $\mu_2 = \sigma^2 + \mu_1^2$. To illustrate the consequences of this result, we discuss two different watermark detectors.

** A similar (but less significant) effect of random DC components occurs if a watermark is built by spatially repeating the same basic small pattern in a large size image and cutting the watermark near the image boundaries.

III.2.1 The first system is designed around the observation that E(y_p|watermark) $= E_w/N$ and E(y_p|no watermark) $= 0$, where the expectation includes all watermarks in the class. For a threshold half-way, i.e., $y_{thr} = E_w/(2N)$, the error rate goes into

$$P_{fa} = P_{md} = \frac{1}{2}\text{erfc}\sqrt{\frac{E_w}{8\left[\sigma^2\left[\frac{1+\alpha\beta}{1-\alpha\beta}\right]^2 + \mu_1^2\left[\frac{1+\beta}{1-\beta}\right]^2\right]}}$$

III.2.2 Alternatively, in the second system, the detection threshold is based on precise knowledge of the watermark including it DC component D_0. E(y_p|watermark, D_0) $= E_w/N + \mu_1 D_0$ and $\mu_1 D_0$ otherwise. That is, the threshold is $y_{thr} = \mu_1 D_0 + E_w/(2N)$.

$$P_{fa} = P_{md} = \frac{1}{2}\text{erfc}\sqrt{\frac{E_w}{8\sigma^2}\left[\frac{1-\alpha\beta}{1+\alpha\beta}\right]^2}$$

The same performance can be achieved by $\tilde{q} = p - \mu_1$ instead of with q as input to the detector. This result reduces to the case of Section III.1 for white watermarks ($\beta = 0$).

We see that the second system outperforms the first one, In experiments, we found that typically μ_2 is about four times larger than σ^2. Hence, performance is better by about 5 to 10 dB. In Figure 2 we took $\beta = 0$. For low-pass watermarks, the differences would be more pronounced.

IV Computational and Experimental Results

The use of randomly generated sequences provides watermarks that *on the average* may have the desired properties, but without a guarantee that individual watermarks also accurately possess the desired properties. This would lead to different results in our analysis and our experiments.

Therefore, in our experiments, we created quasi white and absolutely DC-free watermarks through appropriate pseudo-random sequences. An appropriate choice appeared to be binary watermarks, $w(\mathbf{n}) \in \{-k, k\}$ with $\beta = 0$ generated by a 2-dimensional LFSR maximal length sequence [13] [14] of length $2^{14} - 1 = (2^7 - 1)(2^7 + 1) = 127 \cdot 129$, with 127 and 129 being relatively prime. Such sequences have a negligibly small DC component: since $\sum_{\mathbf{n}} w(\mathbf{n}) = -1$, we get $D_0 = 1/(2^{14} - 1)$. Their spatial correlation function has the appropriate δ-function shape. Repetition of the 127 by 129 basic pattern leads to a periodic correlation function, but maintains virtually zero correlation outside the peaks. Experiments revealed that effects of cutting off this pattern at non-integer repetition numbers had negligible effect for the large images that we used $N_1 = 720, N_2 = 480$.

Figure 2 compares the above theoretical results with measurements of the "Lenna" image. In the figure, we combined the results from one image and

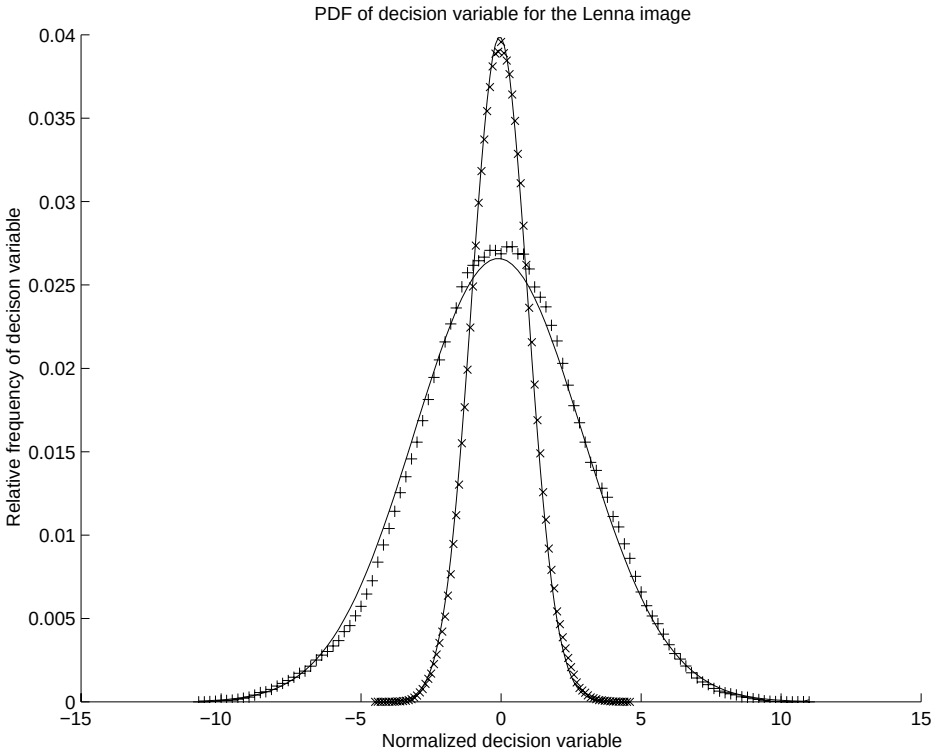

Fig. 3. Probability density of decision variable y_p for white and low-pass watermark on "Lenna". Theory: solid lines. Experiments: "$\times$" white ($\beta = 0$) and "+" low-pass ($\beta = 0.5$) watermark.

many watermarks. We computed the components of the decision variable and estimated which signal-to-noise ratio would be needed to achieve detection just above the threshold. Any particular image with a particular watermark gives a step-wise transition in the performance plot, at $y_w = 2y_p$. Combination of this step for many (more than 10^4) watermarks created the smooth curve.

We computed y_p by correlating with a normalized reference copy of the watermark $\hat{w} \in \{-1, +1\}$. In such case $y_w = k$, where k is the embedding depth of the watermark. We measured y_p from the image. To ensure correct detection for a particular watermark and image, the embedder must choose $k \geq 2y_p$, so $E_w = k^2 N_1 N_2 \geq 4y_p^2 N_1 N_2$. The shape of the curve confirms the (approximately) Gaussian behaviour of y_p. Other images produced the same curve.

To get consistent results, we had to generate a large set of watermarks fully independently. If one simply shifts the same watermark to create the set of test watermarks, correlation of pixels in the image leads to correlated decision variables. Moreover, shifting a single watermark can not simulate the effect of

a random DC component. This would lead to a significant deviation from the theoretical curve and to a more stepwise (non-Gaussian) decrease of errors rates with increasing SNR. We noted that images with higher correlation are more sensitive to correlations among different watermarks during the experiments.

Figure 3 compares white and low-pass watermarks. Both watermarks have been generated using a pseudo-random generator to create a white watermark. For the low-pass watermark, this was then filtered. The experiments confirm the Gaussian distribution. This result clearly shows that if the watermark is confined to low-pass components of the image, this significantly affects the reliability of detection. The standard deviation of the decision variable is larger. In this case, the random +/- terms in y_p, which are due to multiplying the image p with the locally stored copy of the watermark $\hat{w}$, do not cancel as rapidly as these would vanish for a white watermark. If the watermark contains relatively strong low-frequency components (large β), the variance of y_p is stronger and the error rate is larger.

If the watermark contains relatively strong high-frequency components $\beta \approx 0$, the variance is weaker, so the watermark sees less interference from the image itself. However, such high-frequency watermark is more vulnerable to erasure by image processing, such as low-pass filtering (smoothing).

V Conclusions

In this paper, we presented a mathematical framework to model the detection of electronic watermarks embedded in digital images, in particular for correlator detectors or matched filters. Several essential differences appear with the theory of (radio) transmission over a linear time-invariant channel with AWGN. Our model predicts reliability performance (missed detection and false alarms). In some special cases, particularly that of a white watermark, the signal-to-noise ratio (watermark-to-content-energy) appears the only factor to influence the reliability of detection. This leads to expressions for error probabilities similar to those experienced in radio communication (e.g. error function of square root of signal-to-noise ratio) However, the spectral properties of the watermark have a significant influence.

If a watermark detector is part of a standardized active copy control system, false alarms are a critical performance parameter. We believe that the analysis of this paper has provided enhanced insight in the rate at which these errors occur.

VI Acknowledgements

The authors greatly appreciated fruitful discussions with Joop Talstra and Jaap Haitsma.

References

1. I. Pitas, T. Kaskalis : "Signature Casting on Digital Images", Proceedings IEEE Workshop on Nonlinear Signal and Image Processing, Neos Marmaras, June 1995
2. W. Bender, D. Gruhl, N. Morimoto and A. Lu, "Techniques for data hiding", IBM Systems Journal, Vol. 35. No. 3/4 1996
3. J.R. Smith, B. O. Comiskey, "Modulation and Information Hiding in Images", in Proc. Workshop on Information Hiding, Univ. of Cambridge, U.K., May 30 - June 1, 1996, pp. 191 - 201
4. I. Cox, J. Kilian, T. Leighton and T. Shamoon, "A secure, robust watermark for multimedia", in Proc. Workshop on Information Hiding, Univ. of Cambridge, U.K., May 30 - June 1, 1996, pp. 175-190
5. I.J. Cox and J.P.M.G. Linnartz, "Public Watermarks and resistance to tampering", Presented at ICIP 97, Santa Barbara, CA, October 1997.
6. J.P.M.G. Linnartz, A.C.C. Kalker, G.F.G. Depovere and R. Beuker, "A reliability model for detection of electronic watermarks in digital images", Benelux Symposium on Communication Theory, Enschede, October 1997, pp. 202-208
7. Ton Kalker, "Watermark Estimation Through Detector Observations", in Proc.of the IEEE Benelux Signal Processing Symposium",1998, "Leuven, Belgium", pp. 119–122.
8. Ton Kalker, Jean-Paul Linnartz and Geert Depovere, "On the Reliability of detecting Electronic Watermarks in Digital Images, acc. at Eusipco-98
9. I. J. Cox, M. L. Miller, "A review of watermarking and the importance of perceptual modeling", Proc. of Electronic Imaging 97, Feb. 1997.
10. N.S. Jayant and P. Noll., "Digital Coding of waveforms" Prentice Hall, 1984.
11. Ch. W. Therrien, "Discrete Random Signals and Statistical Signal Processing" Prentice Hall, 1992.
12. "Wireless Communication, The Interactive MultiMedia CD ROM", Baltzer Science Publishers, Amsterdam, 2nd Edition, 1997, http://www.baltzer.nl/wirelesscd
13. F.J. McWilliams and N.J.A. Sloane, "Pseudo-Random Sequences and arrays", Proc . of IEEE, Vol. 64, No.12, Dec. 1976, pp. 1715-1729
14. D. Lin and M. Liu, "Structure and Properies of Linear Recurring m-arrays", IEEE Tr. on Inf. Th., Vol. IT-39, No. 5, Sep. 1993, pp. 1758-1762
15. G.R. Grimmet and D.R. Stirzaker, "Probability and Random Processes", chapter on convergence of random variables, Oxford Science Publishers, 2nd Edition, 1992.

This article was processed using the LaTeX macro package with LLNCS style