On the design of a watermarking system: considerations and rationales

Jean-Paul Linnartz, Geert Depovere, and Ton Kalker

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

Abstract. This paper summarizes considerations and rationales for the design of a watermark detector. In particular, we relate watermark detection to the problem of signal detection in the presence of (structured) noise. The paper builds on the mathematical results from several previously published papers (by our own research group or by others) to verify and support our discussion. In an attempt to unify the theoretical analysis of watermarking schemes, we propose several extensions which clarify the interrelations between the various schemes.
New results include the matched filter with whitening, where we consider the effect of the image and watermark spectra and imperfect setting of filter coefficients. The paper reflects our practical experience in developing watermarking systems for DVD copy protection and broadcast monitoring. The aim of this paper is to further develop the insight in the performance of watermark detectors, to discuss appropriate models for their analysis and to provide an intuitive rationale for making design choices for a watermark detector.

1 Introduction

The understanding of reliable methods to detect embedded data or watermarks has progressed substantially over the past years. Many of the first proposals for watermarking emerged from the image processing community. It was also recognized that detection theory and spread-spectrum communication have several aspects in common with watermark detection, and results from these fields are now also exploited to improve the detection performance. This paper reviews the relation with detection theory and develops an intuitive understanding of the behavior of various approaches to watermark detection. It is not intended as a *how to* recipe, but rather as an attempt towards the development of a better understanding and a unified and more rigid theoretical modeling of watermark detection. Hitherto, several detection principles have been proposed and verified experimentally, but theoretical support often was meagre. Our discussion mostly refers to theoretical models, rather than to experiments. Nonetheless most of the models have been verified by experiments reported in previous publications. New results are obtained to further verify and illustrate detection performance. Our experiments have been conducted during the development of watermarking systems, both for a consumer electronics application, viz., the JAWS system [**?**] [**?**] proposed by the Millennium Group as a solution to DVD copy protection[**?**], and for a professional application, viz., the VIVA system for automated monitoring of advertisements and news

clips in television broadcasts [**?**]. In this paper, we address a large and important class of watermarks in which a pseudo-noise pattern is added to the luminance of the image pixels. This watermarking technique may involve adaptive embedding, based on perceptual masking models [**?**].

We are most concerned with the application of *embedded signaling*, i.e., of carrying additional data along with images or video. The three prime optimization criteria are low perceptibility, cost (intemately related to complexity), and robustness against common processing operations [**?**]. Robustness against intentional attempts to remove the watermark, and confidentiality in covert communication (hiding the fact that additional data is embedded) are of secondary importance. We do not specifically address the temporal aspect of motion pictures, so our results apply to images as well as to video. It has been shown that a watermark embedder can exploit its knowledge about the image. This leads to the modeling of a communication channel with side information at the transmitter [**?**]. Our paper (which primarily focusses on detection) ignores this aspect.

The outline of the paper is as follows. Section 2 formulates a watermarking model and defines parameters that we will use in our discussions in Section refsec:3. The subsections of Section 3 address specific watermark detectors or refinements of these. Our discussion progresses from very basic schemes, such as the correlator in Section 3.1, and develops further sophistications step by step. Section 3.2 discusses some important ingredients of the correlator concept, in particular the size of the watermark alphabet. Section 3.3 addresses non-stationarity in the image. It justifies Wiener filtering on theoretical grounds, but finds that under slightly different assumptions for the embedding process, another adaptive filtering is preferable. The model in Section 3.4 addresses spectral prefiltering, but finds shortcomings in some implementations, which can be resolved by the whitening filter of Section 3.5. Section 3.6 provides a frequency-domain interpretation of the whitened matched filter. It extends the classic discussion of whether one should mark the perceptually relevant or irrelevant areas of the image. This section also relates the effect of MPEG compression to the theory of quantization and dithering. Section 3.7 discusses phase-only matched filtering and relates this to the theoretical model of Section 3.3. Section 3.8 discusses the problem of threshold setting. Section **??** supports the discussion of Section 3 by a mathematical analysis. It provides a derivation of new results for watermark detecting with imperfect prefiltering. Section **??** concludes the paper.

2 Preliminaries

We consider two stochastic processes: $\mathcal{W}$ generates watermarks W and $\mathcal{P}$ generates images P. Processed images, derived from $\mathcal{P}$ will be denoted as $\mathcal{Q}$ and $\mathcal{R}$. The watermark is seen as a random process because it is created from a pseudo-random sequence generator, which is fed by a random *seed*. We want our system performance to be sufficiently independent of the choice of this seed. Earlier analysis has shown that this can be ensured if certain restrictions are imposed on the sequence generation process. DC-freeness is one such requirement [**?**].

The image and its watermark have a size of N_1 by N_2 pixels with a total of $N = N_1 N_2$ pixels. The intensity level (called *luminance*) of the pixel with coordinates $\vec{n} =$

(n_1, n_2), $(0 \leq n_1 \leq N_1 - 1, 0 \leq n_2 \leq N_2 - 1)$ for image P (upper case!) is denoted as $p(\vec{n})$ (lower case!). The set of all pixel coordinates is denoted as A. We restrict our discussion to gray scale images in which $p(\vec{n})$ takes on real or integer values in a certain interval. Whenever convenient we will represent $p(\vec{n})$ as a z-expression $p(\vec{z})$ defined by

$$p(\vec{z}) = \sum_{\vec{n} \in A} p(\vec{n}) z^{-\vec{n}} = \sum_{\vec{n} \in A} p(\vec{n}) z_1^{-n_1} z_2^{-n_2}. \tag{1}$$

2.1 Image Model

In some (but not all) analyses, we will make the simplification that the stochastic processes W and P are wide-sense stationary (WSS) and ergodic [**?**]. By ergodicity we are allowed to approximate the statistical k-th moment $\mu_k(p)$ by the spatial k-th moment $m_k(p)$, viz.,

$$\mu_k(p) = \mathrm{E}[p^k(\vec{n})] = m_k(p) = \frac{1}{N} \sum_{\vec{n} \in A} p^k(\vec{n}). \tag{2}$$

WSS means that the statistical autocorrelation function $\Gamma_{p,p}(\vec{n}, \vec{m})$ only depends on the difference vector $\vec{\Delta} = (\Delta_1, \Delta_2) = (\vec{n} - \vec{m})$. That is,

$$\Gamma_{p,p}(n, m) = \mathrm{E}[p(\vec{n}) p(\vec{m})] = \mathrm{C}_{p,p}(\Delta) = \frac{1}{N} \sum_{\vec{n}} p(\vec{n}) p(\vec{n} + \vec{\Delta}). \tag{3}$$

A simple model for images is the first-order separable autocorrelation function [**?**]

$$\mathrm{C}_{p,p}(\vec{\Delta}) = m_1^2(p) + \sigma_p^2 \alpha^{|\vec{\Delta}|}, \tag{4}$$

where $|\vec{\Delta}| = |\Delta_1| + |\Delta_2|$. The standard deviation s_p is defined as $\sigma_p^2 = m_2(p) - m_1^2(p)$. The quantities $m_1(p)$ and $m_2(p)$ are referred to as the *DC-component* and the *power* of the image p, respectively. The value α reflects the correlation between adjacent pixels in the image. In other parts of the discussion we refine the WSS, ergodicity and autocorrelation models by assuming that these properties only apply locally.

We denote $\tilde{p}(\vec{n})$ as the DC-free component of the image, that is $\tilde{p}(\vec{n}) = p(\vec{n}) - m_1(p)$, so

$$\mathrm{C}_{\tilde{p}\tilde{p}}(\vec{\Delta}) = \sigma_{\tilde{p}}^2 \alpha^{|\vec{\Delta}|}. \tag{5}$$

To avoid problems discussed elsewhere [**?**], we will assume that in the watermark detector all signals have been processed by subtracting the DC-component such that $\tilde{p} = p$, or, equivalently $m_1(p) = 0$.

2.2 Watermark Model

A watermark $w(\vec{n})$ is modeled as a sample drawn from the stochastic process W. The *energy* in a watermark w equals $N\mathrm{C}_{w,w}(0) = Nm_2(w)$ and is denoted as E_w. Similarly as in the case of images we assume that w is DC-free, i.e. $\tilde{w} = w$. *White watermarks* have a spatial autocorrelation function which approaches the discrete Dirac distribution when the image size is large enough: $\mathrm{C}_{w,w}(\vec{\Delta}) = N^{-1}E_w\delta(\vec{\Delta})$.

Our method of creating a *low-pass watermark* is by spatially filtering a white watermark source W with a first-order two dimensional spatial smoothing IIR filter $S_\beta(\vec{n})$,

$$S_\beta(\vec{z}) = \frac{1-\beta^2}{(1-\beta z_1^{-1})(1-\beta z_2^{-1})}. \tag{6}$$

In this case the autocorrelation becomes:

$$\mathrm{C}_{ww}(\Delta) = \frac{E_w}{N}\beta^{|\Delta|}. \tag{7}$$

The watermark is embedded according to

$$r(\vec{n}) = p(\vec{n}) + \phi(\vec{n})w(\vec{n}), \tag{8}$$

where $\phi(\vec{n})$, $(\phi(\vec{n}) > 0)$ denotes a local embedding depth, which adaptively depends on a local neighborhood of $\vec{n}$. Mostly, a global embedding depth condition guarantees that $m_2(\phi w) \approx m_2(\phi)m_2(w)$ equals E_w/N, thus $m_2(\phi) = 1$.

3 Discussion

Several early papers [**?**] [**?**] [**?**] propose a watermarking system which is equivalent to increasing the luminance of one set of pixels in the image by one quantization step, and decreasing it by one quantization step in a second set of pixels. The number of elements in both sets was taken equal. Thus, $w \in \{-1, 0, +1\}$. We denote $A_- = \{\vec{n} : w(\vec{n}) = -1\}$ and $A_+ = \{\vec{n} : w(\vec{n}) = +1\}$. Here, $A_+ \cap A_- = \emptyset$ and $A_+ \cup A_- \subseteq A$. Watermarks are detected by computing the sum of all pixel luminance values at locations where the watermark is negative, i.e., $s_- = \sum_{\vec{n}\in A_-} r(\vec{n})$ and the sum of all luminance values where the watermark is positive, i.e., $s_+ = \sum_{\vec{n}\in A_+} r(\vec{n})$. Then, an expression such as $d = (s_+ - s_-)/N$ is used as a decision variable. This scheme was later improved and the underlying model generalized to include

- adaptive embedding, to exploit masking properties of the image,
- real-valued watermarks w,
- embedding in different domain such as the DCT transform domain
- methods to exploit correlation in image pixels to improve the detector performance

3.1 Generalization to correlation

The detector of the previous subsection is a special case of a *correlator detector* or *matched filter* [**?**]. In a correlator detector, a decision variable d is extracted from the suspect image $R = \{r(\vec{n})\} = P + W$ according to correlation with a locally stored copy of a (not necessarily equal) watermark $\hat{w}(\vec{n})$, so

$$d = C_{\hat{w},r}(0) = \frac{1}{N} \sum_{\vec{n}} \hat{w}(\vec{n}) r(\vec{n}) = d_p + d_w. \tag{9}$$

Here the watermark contribution d_w equals $d_w = C_{\hat{w},w}(0)$ if the watermark is present and $d_w = 0$ otherwise. The image contribution $d_p = C_{\hat{w},p}(0)$ is a zero-mean *projection* of the image p on the watermark $\hat{w}$. Its variance determines the amount of noise or interference to the watermark.

Ignoring some subtleties, the matched filter theorem [**?**] can be summarized as the statement that $\hat{w}(\vec{n}) = w(\vec{n})$ is the optimum choice for the local copy $\hat{w}$. Important assumptions are that the watermark signal (and its location or phase) are known and that the noise is additive, white and Gaussian. Under these conditions the decision variable d has the best achievable signal to noise ratio SNR, which is defined as $g = m_2(d_w)/m_2(d_p)$. Also, once d is known, no other properties can be extracted from R that would further improve the detector reliability.

The white noise assumption is equivalent to assuming that pixels in an image have random luminance values, independent from pixel to pixel. This may not model real-world images very well, but the matched filter theory also provides a foundation for further improvement, in casu the *whitened matched filter* which we will address later.

The Gaussian assumption may also lack realism, but up to now we have not found any paper in open literature which describes how to exploit the precise probability distribution of image luminance values to enhance detector reliability. Experiments, e.g. [**?**][**?**], confirmed that after accumulation of many pixels, $d_p(w)$ has a Gaussian distribution if N is sufficiently large and if the contributions to the summing are sufficiently independent. In [**?**] the model for the tails of the distribution is refined, leading to the conclusion that the Gaussian assumption for correlation values leads to pessimistic predictions for false positives.

3.2 Corollary

The concept of the matched filter directly suggests how to handle watermarks which draw w-values from a larger alphabet than $\{-1, 0, +1\}$. The matched filter detector multiplies every pixel of the suspect image with the luminance value of the reference watermark, $\hat{w} = w$. Thus, the detector should weigh most heavily the pixels (or frequencies) in which most watermark energy has been put, in fact, the best weighing is proportionally to the strength of the watermark. The use of a multi-valued watermark has several advantages.

- It is stronger against specific attacks, such as collusion attacks [**?**] or the histogram attack [**?**].

– Moreover, it is useful to have real-valued watermarks (or a discrete alphabet of sufficient size) when adaptive embedding is used. If the pixel modification ϕw is quantized and if w only has binary values (+1, -1), undesired discontinuities may occur. In such case, the effect of a carefully calculated ϕ is reduced to a crude switching of the watermark level. Boundaries may be particularly disturbing to the human eye.
– Real-valued watermarks occur naturally if the watermark is defined in one domain (e.g. spatial domain) but detected in another domain (e.g. JPEG or MPEG DCT coefficients).

Another observation is that detection based on correlation is equivalent to extracting a decision variable which is a linear combination of pixel luminance values. Hence correlation can be performed in any transform domain for which energy preservation is guaranteed. It can be calculated for instance in pixel domain, in an image-wide DCT, block-based DCT or FFT. While several embedding methods have been based on modifying MPEG or JPEG DCT coefficients, such watermarks can also be detected by correlation in the spatial domain, or vice versa. Domain transforms can further be used to speed up the correlation calculation or as a computationally efficient manner to search for watermarks in altered (e.g. shifted) images [**?**].

An aspect relevant to the complexity of the detector, is the ability to use *tiling* [**?**]. This is a method of spatially repeating a watermark pattern of size M_1 by M_2 with $M_1 < N_1$, $M_2 < N_2$ in the image, according to

$$r(\vec{n}) = p(\vec{n}) + \phi(\vec{n})w(n_1 \mathrm{mod} M_1, n_2 \mathrm{mod} M_2). \tag{10}$$

Since correlation is linear, i.e., $C_{p+q,w} = C_{p,w} + C_{q,w}$, the computational speed detection can be improved by a factor $N_1N_2/(M_1M_2)$ by cyclically wrapping the suspect image to one of size M_1M_2, on which correlation is then performed [**?**].

3.3 Exploiting non-stationarity

Most watermarks are embedded with an adaptive depth ϕ which depends on the masking properties of the image. Ideally the detector should take this into account. This section presents an optimum method for this. Let's assume that the image can be partitioned into I sub-images A_0, A_1, .., A_i, .. A_{I-1}, with N_0, N_1, .. ,N_i, .., N_{I-1} pixels, respectively, with $\sum_I N_i = N$. Each sub-image A_i has its own variance $\sigma^2_{p,i}$, autocorrelation α_i, and masking properties ϕ_i, which are constant within A_i. In practice, feature extraction algorithms may be used to partition A [1]. The problem of optimal detection relates to *diversity* [**?**] radio reception, a system which combines signals received by multiple antennas. Borrowing from this theory, one can extract I decision variables $d_0, d_1, ..d_{I-1}$, defined as

[1] We will use later that this partitioning is not particular to the spatial domain, but may also occur in frequency domain.

$$d_i = \frac{1}{N_i} \sum_{\vec{n} \in A_i} r(\vec{n}) \hat{w}(\vec{n}) = d_{i,w} + d_{i,p}. \tag{11}$$

Here $d_{i,w} = N_i^{-1} \sum_{\vec{n} \in A_i} \phi(\vec{n}) w(\vec{n}) \hat{w}(\vec{n}) = \phi_i N_i^{-1} \sum_{\vec{n} \in A_i} w(\vec{n}) \hat{w}(\vec{n})$ and $d_{i,p} = N_i^{-1} \sum_{\vec{n} \in A_i} p(\vec{n}) \hat{w}(\vec{n})$. We define $E_{i,w} = \sum w^2(\vec{n})$, so if $\hat{w} = w$, $d_{i,w} = \phi_i N_i^{-1} E_{i,w}$. Section 3.8 describes how to estimate the variance $\sigma_{i,p}^2 = m_2(d_{i,p})$ in a practical detector. In the case that the local copy $\hat{w}$ is white Gaussian noise, it not hard to show that $sigma_{i,p}^2$ is proportional to the variance of the pixel data $m_2(p)$.

A single decision variable can be combined from

$$d = \sum_{i=0}^{I-1} d_i f_i. \tag{12}$$

Using Cauchy's inequality, the selection of the weigh factor f_i can be optimized for optimal SNR as follows,

$$\begin{aligned} g &= \frac{\{\sum_{i=0}^{I-1} d_{i,w} f_i\}^2}{\sum_{i=0}^{I-1} \sigma_{i,p}^2 f_i^2} \leq \frac{\sum_{i=0}^{I-1} d_{i,w}^2 \sigma_{i,p}^{-2} \sum_{i=0}^{I-1} \sigma_{i,p}^2 f_i^2}{\sum_{i=0}^{I-1} \sigma_{i,p}^2 f_i^2} \\ &= \sum_{i=0}^{I-1} \frac{d_{i,w}^2}{\sigma_{i,p}^2} = \sum_{i=0}^{I-1} \frac{\phi_i^2 E_{i,w}^2}{N_i^2 \sigma_{i,p}^2} = \sum_{i=0}^{I-1} \frac{\phi_i^2}{\sigma_{i,p}^2}, \end{aligned} \tag{13}$$

where we assume (without loss of generality) that $E_{i,w}/N_i = 1$. Equality holds if

$$f_i = \frac{d_{i,w}}{\sigma_{i,p}^2} = \frac{\phi_i E_{i,w}}{N_i \sigma_{i,p}^2} = \frac{\phi}{\sigma_{i,p}^2}. \tag{14}$$

Next we will consider two special cases of this result. These special cases can be shown to be approximately equivalent to a previously proposed detection methods.

The first case is that of an embedder using white Gaussian noise as a watermark pattern, with $\phi_i \approx \sigma_{i,p}$. This approximates a typical embedder which assumes that the perceptivity of a watermark is related to the standard deviation $\sigma_{i,p}$ in the pixel luminance. Inserting this into the formula for optimum weighing of the subimage decision variables f_i, we find $f_i = E_{i,w} \sigma_{i,p}^{-1}$ as the optimal choice. We will use this in our discussion of phase-only matched filtering in Section 3.7.

The second case is fixed-depth embedding, i.e., $\phi_i =$ constant. Our generic formula Equation 14 proposes that a detector should weigh $f_i = \sigma_{i,p}^{-2}$. Intuitively we explain this as follows: one division by $\sigma_{i,p}$ makes the noise power identical for all pixels, the second division weighs pixels proportionally to their strength. Radio engineers call the latter weighing *maximum ratio combining* [**?**]. This special case relates to extracting a Wiener filtered copy from the suspect image $R = Q - Wiener(Q)$, proposed in [**?**], which corresponds to

$$f_i = 1 - \frac{N_i \sigma_{i,p}^2}{N_i \sigma_{i,p}^2 + E_{i,w}} \tag{15}$$

Since one can approximate this as

$$f_i \approx \frac{E_{i,w}}{N_i \sigma_{i,p}^2}, \tag{16}$$

we can now provide a justification for the use of wiener filtering on theorectical grounds. We refer to [**?**] for a quantification of the improvement gains. Also, [**?**] develops a model for the impact of adaptive filtering, though it was used as an attack to minimize $C_{w,r}$ by $R = Wiener(Q)$.

3.4 Prefiltering

In this section we consider watermark detection when correlation is preceded by filtering of the image. In the following sections we will also consider prefiltering of the watermark. When an image is linearly filtered, a new image R is created in which each pixel luminance is a combination of pixel luminance values in the original image Q. Most filters operate locally, thus combining pixels in the neighborhood (small $\vec{\Delta}$) of the pixel that is created in the new image, according to the convolution

$$R = \sum_{\vec{\Delta}} h(\vec{\Delta}) Q(\vec{n} + \vec{\Delta}), \tag{17}$$

Here $h(\vec{\Delta})$ are the filter coefficients of a *filter* H. Reffering to Figure 1, this role is conducted by filter H_1, $R = H_1(Q)$, for the image and by filter H_2 for the watermark, $\hat{W} = H_2(W)$.

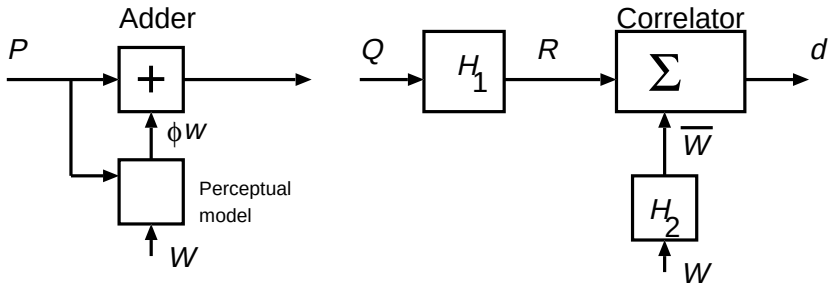

Fig. 1. Embedder and Correlator detector with prefiltering

When a correlation detector is preceded by filtering, the SNR in the decision variable differs from the result for an unfiltered image [**?**] [**?**] [**?**] [**?**]. For linear prefiltering

it is not difficult to show that the situation of Figure 1 is equivalent to correlation with W on the image $(H_2^* H_1)(R)$, where H_2^* denotes the time-inverse of the filter H_2.

An edge-enhancement filter [**?**] or median filter [**?**] filter can be used to predict the image in pixel n from the neighbouring pixels. This prediction is extracted from the actual luminance, according to $R = H_2(Q) = Q - Pre(Q)$, where $Pre()$ denotes prefiltering. This exploits the redundancy in the pixels of the video. These filters reduce $m_2(d_p)$, the variance of the noise, and were shown to give a performance improvement. However these do not necessarily also maximize E$[d_w]$ or the signal to noise ratio. Optimization of the SNR g leads to the *whitened matched filter*.

3.5 Whitening Prefilter

For non-white noise, one can first prefilter the suspect image Q into R, such that its frequency spectrum is sufficiently white (see Figure 2. Subsequently, a matched filter Σ_1 is used for $R = H_1(Q)$. Readers who are familiar with information theory (e.g. [**?**]) may wish to skip the next paragraph, which sketches the proof by contradiction to see that this detector is optimum.

If the prefilter H_1 is invertible, $R = H_1(Q)$ and Q intrinsically carry the same information, so optimum detectors for R and Q must have the same reliability. Let $\Sigma_2(Q)$ be a fictitious detector which detects watermarks in Q more reliably than $H_1(\Sigma_1(Q))$, i.e., the concatenation of Σ_1 and H_1 executed on Q. Then, for images R (with white noise), the detector $\Sigma_2(H_1^{-1}(R))$ would outperform $\Sigma_1(R)$. This is at odds with the the matched filter theorem that $\Sigma_1(R)$ is optimum for R. So, $H_1^{-1}(\Sigma_2))$ cannot outperform Σ_1. This implies that $\Sigma_2(Q) = \Sigma_2(H_1^{-1}(H_1((Q)))$ cannot perform better than $\Sigma_1(H_1(Q))$. Thus, $\Sigma_1(H_1(Q))$ must be optimum

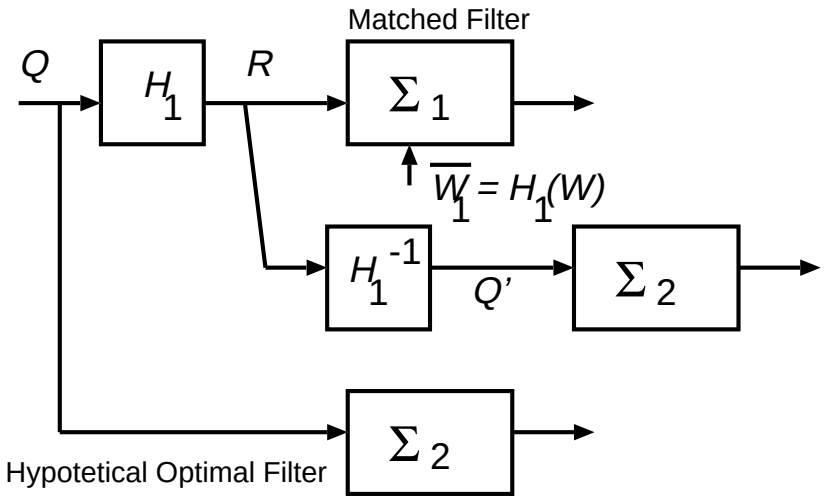

Fig. 2. Whitening Matched Filter H_1 and Σ_1, and fictitious better detector Σ_2

Whitening can be interpreted as a form of (high-pass) prefiltering before correlation. The whitened matched filter differs from the prefilter concept of he previous section in the sense that not only the suspect image but also the locally stored reference watermark is filtered ($\hat{W} = H_1(W)$). In fact in Figure 2, Σ_1 is a matched filter for R, where the watermark component in R is $H_1(W)$.

Implementation-wise we observe that the correlator, the correlator with prefiltering, and the whitened matched filter all create a decision variable which is a *linear* combination of the pixel luminance values. Thus, any of these can be implemented just as correlator, without any prefiltering. In other words, using the implementation of Figure 1, the detector does not have to execute filtering operartions H_1 and H_2 in real-time. Both H_1 and H_2 can be incorporated in a precomputed $\hat{W}$. We refer to Section **??** for an analysis of the perforamance and the sensitivity to the accuracy of the filter setting. Experiments with whitening are reported in [**?**].

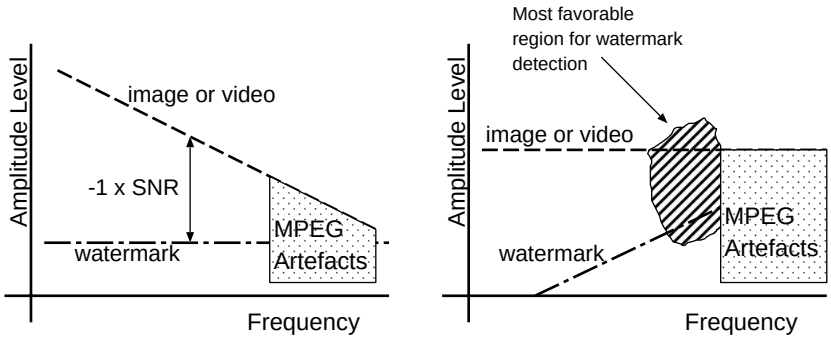

Fig. 3. Frequency components in image and watermark, without and with whitening

3.6 Frequency Components

The whitening concept throws new light on the discussion whether the watermark should be located in perceptually relevant or irrelevant areas (or frequency components) of the image. Whitening enhances high frequency components, which are relatively weak in a typical image. Whitening weakens the low frequency components, which are strong in a typical image. Thus these prefiltering methods tend to weigh high frequency components more heavily in the decision variable than low frequency components. The intuition behind it is that at low frequencies, the image itself causes stronger interference to the watermark than at higher frequencies.

A limitation of these models is that they not model typical distortion by MPEG or JPEG very well. In a first order approximation, these techniques can be interpreted to crudely quantize the medium to high-frequency components, and remove (i.e., quantize to zero) the upper higher frequencies. For the medium to upper frequencies, the watermark may not be affected so dramatically, because typically the image *dithers* [**?**] [**?**] the watermark. Dithering is illustrated in Figure 4.

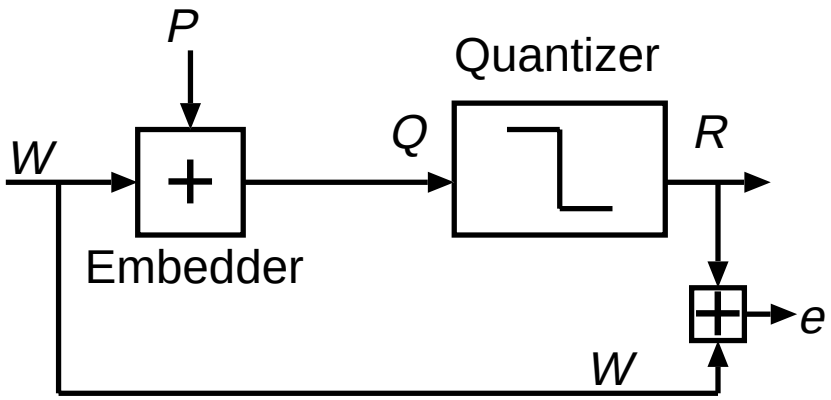

Fig. 4. Watermark embedding and lossy compression interpreted as a dithering system on watermarks.

Dithering can ensure that the error $E = R - W$ is statistically uncorrelated with the watermark W, i.e., $C_{w,e} = 0$ [**?**]. It follows the that the expected decision variable of the watermark detector becomes, for $\hat{W} = W$,

$$\mathrm{E}[d] = C_{w,R} = C_{w,w+e} = C_{w,w} + C_{w,e} = C_{w,w} = \mathrm{E}[d_w] \tag{18}$$

That is, after *dithering by the image*, compression or quantization does not affect the correlation performance compared to a system without the quantizer.

On the other hand, for the upper highest frequencies the dithering effect is lost[2] and the watermark components in these frequencies do not contribute to d_w [**?**]. As these frequencies may nonetheless contain noise, the detector must avoid to excessively weigh upper high frequencies in the correlation. This requires a modification to $\hat{W}$ or H_2 in Figure 1.

3.7 Phase Only Matched Filtering

A refinement of the whitened matched filter is to adaptively whiten the spectrum of the incoming video. Instead of using a fixed H_1, the detector calculates an FFT of Q, and sets amplitudes to unity at all frequencies, thereby preserving only phase information [**?**].

This idea agrees with the result in Section 3.3. The rationale is that the watermark is only a small perturbation of the luminance. We partition the image into I spectral components. Even if it has been watermarked the suspect image gives a good estimate of the spectral components σ_i of P. It turns out that it is reasonable to assume that the embedding depth is likely to be more or less proportional to σ: even if this is not the case

[2] Formally speaking, the probability density of the amplitude of the image components do not satisfy the particular relation to the quantization step size as it was derived in [**?**]

for freshly watermarked content, this is almost always true after common processing (JPEG, MPEG). Using the results of Section 3.3, the weigh factor is found as $f_i = \phi_i E_{i,w}/\sigma_{i,p}^2$, i.e. $f_i = E_{i,w}/\sigma_{i,p}$ which corresponds to phase-only filtering. We refer to [**?**] for an analysis of the performance of the phase only matched filter.

3.8 Adaptive Threshold Setting

Up to this point we have primarily described the extraction of a real-valued decision variable. Next we will discuss the extraction of a hard *watermark present - not present* decision. Detection theory suggests that a threshold can be used, and almost all practical systems use this method. A suspect signal (or an extracted set of features) is correlated with some pattern to obtain a correlation value. If this value is larger than a signal-dependent threshold then the watermark is said to be present. Otherwise the watermark is said to be absent. The setting of the threshold determines the trade-off between the false negative and the robustness to image processing. [3].

Only if non-adaptive embedding is used and the image is not modified the detector can exploit that $d_w = E_w$. In practice, the value d_w is not known exactly to the detector, because the algorithm for adaptive embedding may not be known or can not be repeated in the detector. Also, minor shifts and scaling of the image severely affects the value of d_w [**?**]. Moreover MPEG compression, or other processing may affect the high frequency components of the watermark in the image.

This makes the determination of the false negative rate problematic, unless extensive statistical assumptions are made about the processing that is likely to affect the image. An appropriate design approach is to determine a required false positive rate and to set the threshold accordingly. The problem of guaranteeing a certain false negative rate, i.e., how to make ϕ, or more precisely the correlation $C_{\hat{w},\mathrm{MPEG}(\phi w)}$ large enough, then becomes mostly an embedding issue.

Various authors observed that d_p is in good approximation a Gaussian random variable [**?**] [**?**] [**?**] [**?**]. The ratio of the threshold setting over $m_2(d_p)$ determines the false positive rate, according to

$$P_{fp} = \mathrm{erfc}\sqrt{\frac{d_{thr}^2}{m_2(d_p)}}$$

Since $m_2(d_p)$ significantly differs from image to image, it appeares useful to estimate $m_2(d_p)$ from the image. A practical solution is to use decision variables gathered during a search for shifts of the image [**?**]. For all attempts in the search that failed, $d_w \approx 0$ so $d \approx d_p$. One can set the threshold level according to

$$d_{thr}^2 = m_2(d_p)\{\mathrm{erfc}^{-1}(P_{fp})\}^2$$

where erfc^{-1} is the inverted error function. This operation is mathematically equivalent to the concept of normalized correlation [**?**].

[3] Traditionally, the false alarm probability versus missed detection probability are compared. However, the watermark embedder has full knowledge about the original image, thus is it can ensure detection. The threshold setting determines what modifications can be tolerated.

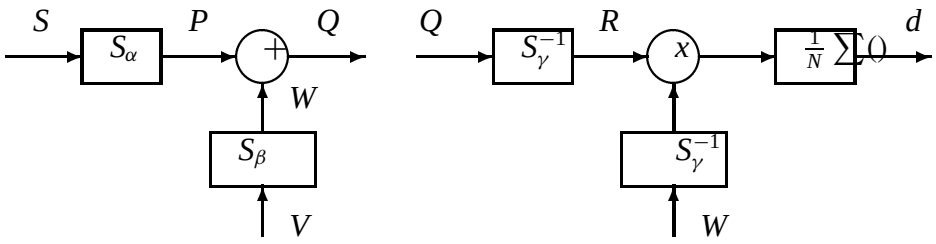

Fig. 5. Watermark Embedding and Correlation Detection.

4 Error Performance for Whitening

This section presents a theoretical study of the whitened matched filter. We consider a low-pass watermark W, which is generated by spatially filtering a white watermark V with a first-order two dimensional spatial smoothing IIR filter $S_\beta(\vec{z})$, We further consider a low-pass image P, which is generated by spatially filtering a white image S with a first-order two dimensional spatial smoothing IIR filter $S_\alpha(\vec{z})$. After watermark embedding, the image is denoted $Q = P + W$. We apply a first order *whitening filter* at the input of the correlation receiver. This filter aims at transforming the non-white input signal of the receiver Q to a signal with a constant power spectrum. As shown in Section III, also the watermark has to be filtered in the same way. In order to keep the model as general as possible, we use the filter $G_\alpha(\vec{z}) = S_\gamma^{-1}(\vec{z})$ as whitening filter.

In the correlator detector, a decision variable d is extracted by correlating the filtered received image Q with a filtered locally stored copy of the watermark W, i.e. $d = d_p + d_w$, where

$$d_p = \frac{1}{N}[S_\alpha(\vec{z}^{-1})S_\gamma^{-1}(\vec{z}^{-1})S(\vec{z}^{-1})S_\beta(\vec{z})S_\gamma^{-1}(\vec{z})V(\vec{z})]_0,$$

and

$$d_w = \frac{1}{N}[S_\beta(\vec{z}^{-1})S_\gamma^{-1}(\vec{z}^{-1})V(\vec{z}^{-1})S_\beta(\vec{z})S_\gamma^{-1}(\vec{z})V(\vec{z})]_0.$$

We introduce:

$$T_\alpha(\vec{z}) = S_\alpha(\vec{z})S_\alpha(\vec{z}^{-1}).$$

Using results from the previous sections, the error probability is given by

$$P = \frac{1}{2}\,\mathrm{erfc}\left(\frac{\mathrm{E}[d_w]}{2\sqrt{2}\sigma_{d_p}}\right),$$

with

$$\begin{aligned}\mathrm{E}[d_w] &= \frac{1}{N}[\frac{T_\beta(\vec{z})}{T_\gamma(\vec{z})}]_0 E_w \\ &= \frac{E_w}{N}[\frac{1+\gamma^2-2\alpha\beta}{1-\gamma^2}]^2.\end{aligned}$$

For the variance we find

$$\begin{aligned}\sigma_{d_p}^2 &= \frac{1}{N^2}\mathrm{E}[[S_\alpha(\vec{z}^{-1})S_\beta(\vec{z})T_\gamma^{-1}(\vec{z})S(\vec{z}^{-}1)V(\vec{z})]_0]^2,\\ &= \frac{\sigma_p^2 E_w}{N^2}[\frac{T_\alpha(\vec{z})T_\beta(\vec{z})}{T_\gamma(\vec{z})^2}]_0\\ &= \frac{\sigma_p^2 E_w}{N^2}[\frac{(1+\alpha\beta)(\gamma^4+2\gamma^2(2-\alpha\beta)+1)-2(\alpha+\beta)(2\gamma^3-(\alpha+\beta)\gamma^2+2\gamma)}{(1-\gamma^2)^2(1-\alpha\beta)}]^2\end{aligned}$$

Inserting these results in the formula for the error rate gives

$$P = \frac{1}{2}\,\mathrm{erfc}\left(\sqrt{\frac{E_w}{8\sigma_p^2}}\frac{(1+\gamma^2-2\beta\gamma)^2(1-\alpha\beta)}{(1+\alpha\beta)(\gamma^4+2\gamma^2(2-\alpha\beta)+1)-2(\alpha+\beta)(2\gamma^2-(\alpha+\beta)\gamma+2)}\right),$$

It should be noted that in the case the whitening filter exactly matches the the received image, i.e. $\alpha = \gamma$ the above expression reduces to

$$P = \frac{1}{2}\,\mathrm{erfc}\left(\sqrt{\frac{E_w}{8\sigma_p^2}}\frac{(1+\alpha^2-2\alpha\beta)}{1-\alpha^2}\right).$$

For a white image $\alpha = 0$ we find again the result

$$P = \frac{1}{2}\,\mathrm{erfc}\sqrt{\frac{E_w}{8\sigma_p^2}}.$$

In the case of a white watermark, i.e.$\beta = 0$, the expression for the error rate becomes

$$P = \frac{1}{2}\,\mathrm{erfc}\left(\sqrt{\frac{E_w}{8\sigma_p^2}}\frac{(1+\gamma^2)^2}{(\gamma^4+4\gamma^2+1-2\alpha\gamma(2\gamma^2-\alpha\gamma+2)}\right).$$

This reduces further to

$$P = \frac{1}{2}\,\mathrm{erfc}\left(\sqrt{\frac{E_w}{8\sigma_p^2}}\frac{1+\alpha^2}{1-\alpha^2}\right).$$

in the case of perfect whitening: $\alpha = \gamma$.

5 Conclusions

We have reviewed various recently proposed watermark detection methods from a detection theory point of view. It appeared that many improvements which have been found from experiments can be explained by extending methods and theories known from communications. These considerations have been a reference for our design of two watermarking systems, which have been demonstrated and evaluated in practical applications [**?**].

Earlier publications confirmed that the use of crude statistical models for images can be useful to create some basic understanding of typical detectors. Here we summarized such results and compiled a consistent intuition for the effect of various sophistications of watermark detectors.

As the models used here have a relatively large scope, we not only justified detector refinements proposed in previous papers, but also found limitations and possibilities for further improvements.

New mathematical results of the effect of imperfect whitening showed that the accuracy with which prefiltering is performed only has a minor effect on the reliability of the detector.

We must leave several aspects for further investigation, such as a comparison of the performance of fixed prefilters (or fixed whiteners) with phase-only matched filtering.

References

1. T. Kalker, G. Depovere, J. Haitsma, M.J. Maes, “A video watermarking system for broadcast monitoring”, Proceedings of SPIE, Security and Watermarking of Multimedia Content, Volume 3657, pp. 103–112, 1999.
2. J.P.M.G. Linnartz, “The ticket concept for copy control based on embedded signalling”, ESORICS ’98, 5th. European Symposium on research in Computer Security, Louvain-La-Neuve, September 1998, Lecture Notes in Computer Science, 1485, Springer, pp. 257-274.
3. J. Bloom, I.J. Cox, A.A.C. Kalker, J.P.M.G. Linnartz, Math Miller and B. Traw, “Copy Protection for DVD”, IEEE Proceedings, Special issue on Information and Protection of Multimedia Information, July 1999, Vol. 87, No. 7, pp. 1267-1266.
4. I. J. Cox, M. L. Miller, “A review of watermarking and the importance of perceptual modeling”, Proc. of Electronic Imaging 97, Feb. 1997.
5. I.J. Cox and J.P.M.G. Linnartz, “Some general methods for tampering with watermarks”, IEEE Journ. of Sel. Areas in Comm., Vol. 16. No. 4, May 1998, pp. 587-593.
6. I.J. Cox, M.L. Miller, and A.L. McKellips. “Watermarking as communications with side information”, IEEE Proc., Vol. 87, No. 7, pp. 1127-1141.
7. J.P.M.G. Linnartz, A.C.C. Kalker, and G.F. Depovere, “Modelling the false-alarm and missed detection rate for electronic watermarks”. Workshop on Information Hiding, Portland, OR, 15-17 April, 1998. Springer Lecture Notes on Computer Science, No. 1525, pp. 258-272, pp. 329-343.
8. 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”, Benelux Symposium on Communication Theory, Enschede, October 1997, pp. 202-209.
9. N.S. Jayant and P. Noll., “Digital Coding of waveforms”, Prentice Hall, 1984.
10. I. Pitas, T. Kaskalis, “Signature Casting on Digital Images”, Proceedings IEEE Workshop on Nonlinear Signal and Image Processing, Neos Marmaras, June 1995.
11. W. Bender, D. Gruhl, N. Morimoto and A. Lu, “Techniques for data hiding”, IBM Systems Journal, Vol. 35. No. 3/4 1996.
12. W. Bender, D. Gruhl, N. Morimoto, “Techniques for Data Hiding”, Proceedings of the SPIE, 2420:40, San Jose CA, USA, February 1995.
13. 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
14. J. Wozencraft and I. Jacobs, “Principles of Communication Engineering”, Wiley, 1965.

15. M. L. Miller and J. A. Bloom, “Computing the probability of false watermark detection”, in Proc. Workshop on Information Hiding 99, Dresden.
16. 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.
17. M. Maes, “Twin peaks: the histogram attack on fized depth watermarks”, Workshop on Information Hiding, Portland, OR, 15-17 April, 1998. Springer Lecture Notes on Computer Science, No. 1525, pp. 290-305.
18. “Wireless Communication, The Interactive MultiMedia CD ROM”, Baltzer Science Publishers, Amsterdam, 3rd Edition, 1999, http://www.baltzer.nl/wirelesscd.
19. L.M. Marvel, C.G. Boncelet, C.T. Retter, “Reliable blind information hiding for images”, Workshop on Information Hiding, Portland, OR, 15-17 April, 1998. Springer Lecture Notes on Computer Science, No. 1525, pp. 48-61.
20. J.R. Hernandez, F. Perez-Gonzalez, J.M. Rodriguez and G. Nieto, “Performance analysius of a 2D Multipulse Amplitude Modulation Scheme for data hiding and watermarking of still images”, IEEE JSAC, Vol. 16, No. 4, May 1998, pp. 510-524.
21. G.C. Langelaar, J.C.A. van der Lubbe, J. Biemond,“Copy protection for multimedia data based on labeling techniques”, 17th Symposium on Information Theory in the Benelux, Enschede, The Netherlands, May 1996.
22. G.F.G. Depovere, A.C.C Kalker, and J.P.M.G. Linnartz, “Improved watermark detection reliability using filtering before correlation”, Int. Conf. on Image Processing, ICIP, October 1998, Chicago IL.
23. G. Langelaar, R. Lagendijk and J. Biemond, “Removing Spread Spectrum Watermarks”, Proceedings of Eusipco-98, Volume IV, pp. 2281–2284, Rhodes, 1998.
24. S.P. Lipshitz, R.A. Wannamaker and J. Vanderkooy, “Quantization and Dither: A theoretical survey”, J. Audio Eng. Soc., Vol. 40, No. 5, May 1992, pp. 355-375.
25. T. Kalker and A.J.E.M Janssen, “Analysis of SPOMF Detection”, Accepted at ICIP-99, Kobe, Japan, 1999.
26. J.P.M.G. Linnartz, A.A.C. Kalker, J. Haitsma, “Detecting electronic watermarks in digital video”, Paper 3010, Invited paper for special session at ICCASP-99, Phoenix, AR, March 1999.
27. M. Wu, M. Miller, J. Bloom, I. Cox, “Watermark detection and normalized correlation”, presented at ICCASP 99, Phoenix, AR, March 1999, also at DSP Conference in Florance.