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. We extend existing models for evaluating the reliability of detecting electronic watermarks in digital images. Probabilities of incorrect detections (missed detection and false alarms) are expressed in terms of the watermark-energy-to-image-variance power ratio. We present some counterintuitive results showing for instance that the reliability of detection significantly depends on spatial correlation in watermark and the luminance values of pixels in the image. Moreover we find that a random DC component in the watermark may have a signifant effect on the performance.

llncs

Abstract. We extend existing models for evaluating the reliability of detecting electronic watermarks in digital images. Probabilities of incorrect detections (missed detection and false alarms) are expressed in terms of the watermark-energy-to-image-variance power ratio. We present some counterintuitive results showing for instance that the reliability of detection significantly depends on spatial correlation in watermark and the luminance values of pixels in the image. Moreover we find that a random DC component in the watermark may have a signifant effect on the performance.

I Background

New multi-media networks and services facilitate the distribution of content, but at the same time make copying and copyright piracy simple. Here we see a clear need to embed copyright data, such as the ownership or the identity of the authorized user in an indelible way. Particularly if watermarking is part of an active copy control concept, typical requirements include:

1. Erasing or altering the watermark should be technically difficult.
2. The watermarking scheme should be 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.

3. It should be robust against typical attacks, e.g. those described in [?].
4. False alarms, i.e., positive responses for content that does not contain a watermark should not occur more often than electronic or mechanical product failures. Consumer electronics devices should not fail to work because a watermark detector was erroneously triggered.
5. The watermark should be unobstructive, and not be annoying to bona-fide users.

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

The organization of the paper is as follows. Section I provides an introduction to the problem of watermarking, its potentials and limitations. Section II introduces our model of the image and the watermark. It extends the idea proposed in [1] to regard the original content as interference during the detection of weak wanted signal (namely the watermark). However, we consider known properties on the image and addresses randomness in the watermark generation. Section III discusses the use of a correlator detector, as is now common practice in many watermark systems. The relaibility of such generic detector is derived, and three special cases are dealt with. The model is verified with experiments in Section IV. Section V concludes this paper.

The aim of this paper is to contribute to the modelling of the reliablility of watermark detectors, which involves the development of a mathematical framework and verification of critical assumptions. A few counterintuitive results are found and discussed.

II Formulation of Model

We approach the problem of watermark detection by assuming a stationary process p as a model for our set of images. We assume certain relevant spatial properties of the image to be known. Watermarking is done by modifying the image such that a detector is triggered. Given a reference copy of the watermark $w(\mathbf{n})$ the detector decides whether an image is watermarked or not by computing a decision variable y and comparing y to a threshold y_{thr}. We will derive expressions for the statistical properties of y and the reliability of detection.

II.1 Image Model

We address an image of size $N = N_1 N_2$ pixels. The intensity level 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

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

In color pictures, $p(\mathbf{n})$ is a YUV or RGB vector, but for sake of simplicity we restrict our discussion to gray scale images, 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 $\mu_k = \mathrm{A}[p^k(\mathbf{n})]$, where A is a spatial averaging operator. In particular, μ_1 represents the average value or expected "DC-component" in an image and $\mu_2 = \mathrm{A}[p^2] = \frac{1}{7}N\sum_{(n)\in A_N} p(\mathbf{n})$ represents the average power in a pixel and $E_p = N\mu_2$ is the average total energy in an image. The variance is $\sigma^2 = \mathrm{E}[p(\mathbf{n}) - \mu_1]^2 = \mu_2 - \mu_1^2$.

The intensity levels of pixels $\mathbf{n_i}$ and $\mathbf{n_j}$ are correlated, with

$$\mathrm{E}[p(\mathbf{n_i})p(\mathbf{n_j})] = \mathit{\Gamma}_{p,p}(\mathbf{n_i} - \mathbf{n_j}).$$

The correlation only depends on the difference vector $\boldsymbol{\Delta} = (\Delta_1, \Delta_2) = (n_{i,1} - n_{j,1}, n_{i,2} - n_{j,2})$, as we assume that the image has homogeneous statistical properties (wide-sense spatial stationarity). If the image size is large enough ($N_1 \gg 0, N_2 \gg 0$) and if the process $p(\mathbf{n})$ is assumed to be ergodic we are allowed to approximate the statistical autocorrelation $\mathit{\Gamma}_{p,p}(\Delta)$ by a spatial autocorrelation $R_{p,p}(\Delta)$

$$\mathit{\Gamma}_{p,p}(\boldsymbol{\Delta}) \approx R_{p,p}(\boldsymbol{\Delta}) = \frac{1}{N}\sum_{\mathbf{n}\in A} p(\mathbf{n})p(\mathbf{n}+\boldsymbol{\Delta}).$$

In order to make calculations for our examples tractable, we simplify the image model assuming the first-order separable autocorrelation function (acf) [**?**]

$$\mathit{\Gamma}_{p,p}(\Delta_1, \Delta_2) = \mu_1^2 + \sigma^2\alpha^{|\Delta_1|+|\Delta_2|}$$

where α can be interpreted as a measure of the correlation between adjacent pixels in the image.

Experiments, e.g. in [**?**] reveal that typically $\alpha \approx 0.9...0.99$. We denote $\tilde{p}(\mathbf{n})$ as the non-DC components of the image, that is $p(\mathbf{n}) = \mu_1 + \tilde{p}(\mathbf{n})$, so $\mathit{\Gamma}_{\tilde{p},\tilde{p}} = \sigma^2\alpha^{|\Delta_1|+|\Delta_2|}$.

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 appeared that reliability estimates based on this crude model can be reasonably accurate for the purpose of this evaluation. These assumptions, however, exclude certain images, such as binary images or computer-generated images with a limited number of colors.

II.2 Watermark Model

To detect a watermark in a suspect image, some proposed methods only use the suspect image and reference data on the watermark, while other methods also require the availability and use of the original image. We assume here that $p(\mathbf{n})$ is not available at the detector. Watermarked images have similar properties as unmarked images, except that perceptially invisible modifications have been made. The watermark is represented by $w(\mathbf{n})$ which takes on real values in all

pixels $\mathbf{n} \in A$. This watermark $w(\mathbf{n})$ is added to the original image. This results in the marked image $q(\mathbf{n}) = p(\mathbf{n}) + \gamma(\mathbf{n})w(\mathbf{n})$, where we intentionally do not specify the embedding depth $\gamma(\mathbf{n})$. In the analysis we assume $\gamma(\mathbf{n}) \equiv 1$forall $\mathbf{n} \in A$.

This model implicitly assumes that no spatial transformation of the image (resizing, cropping, rotation, etc.) is conducted. We aim at detecting whether a particular watermark is present or not, based on knowledge of $w(\mathbf{n})$. A watermark detector has to operate on the observation $q(\mathbf{n})$, while having knowledge on the a priori statistical behaviour of $p(\mathbf{n})$.

For two watermarks watermark w_1 and w_2 the (deterministic) spatial inner-product is

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

where we assume for simplicity that $n + \Delta$ wraps around when it formally falls outside of the set A. If we consider an ensemble of many watermarks generated by a particular watermark generation algorithm, the statistical correlation

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

The total energy in watermark equals $E_w = \sum_{\mathbf{n} \in A} w^2(\mathbf{n}) = N\Gamma_{w,w}(\mathbf{0})$.

II.3 DC components

The DC content of the watermark is $D_0 = \frac{1}{N} \sum_{\mathbf{n} \in A} w(\mathbf{n})$.

Some watermarks are generated by randomly generating a $+k$ or $-k$ pixel value for $w(\mathbf{n})$, independently for each pixel $\mathbf{n}$. Averaged over a large collection of such watermarks, the mean DC component $\mathrm{E}w(\mathbf{n})$ is zero, however each individual watermark not necessarily has a zero DC component. We call a *watermark generation process* "statistically DC-free" or "DC-free in the mean" iff $\mathrm{E}\left[\sum_{\mathbf{n} \in A} w(\mathbf{n}) = 0\right]$. This is a necessary, but not a sufficient condition for all individual watermarks to be DC free. An individual watermark is DC-free iff $D_0 = 0$.

For an arbitrary value of D_0,

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

II.4 Watermark Spectrum

This has consequences for DC-free and spectrally white watermark, which have an correlation function that is a δ-function. For a watermark with D_0, one can consider $\Gamma_{w,w}(\boldsymbol{\Delta_1}) = \Gamma_{w,w}(\boldsymbol{\Delta_2}) = \eta$ for $\boldsymbol{\Delta_1}, \boldsymbol{\Delta_2} \neq 0$, where η is some constant ($|\eta| << E_w/N$). It follows that $\eta = (N^2 D_0^2 - E_w)/(N(N-1))$. In particular, we see that for a DC-free watermark ($D_0 = 0$), the values of $w(\mathbf{n_i})$ and

$w(\mathbf{n_j}), \mathbf{n_i} \neq \mathbf{n_j}$ cannot be statistically uncorrelated ($\eta < 0$). * This is also seen using a statistical argument regarding the observation that some pixel $\mathbf{n_0}$ has some non-zero value $w(\mathbf{n_0}) = k_0$ requires that the $N-1$ other pixels in the image must compensate for this through $\sum_{n_i \in A \setminus \mathbf{n_0}} w(\mathbf{n_i}) = D_0 - k_0$. Using the spatial randomness that $\Gamma_{w,w}(\boldsymbol{\Delta_1}) = \Gamma_{w,w}(\boldsymbol{\Delta_2})$ for $\boldsymbol{\Delta_1}, \boldsymbol{\Delta_2} \neq \mathbf{0}$, we find $\mathrm{E}[w(\mathbf{n_i})|w(\mathbf{n_0}) = k_0] = (D_0 - k_0)/(N-1)$, and using $\Gamma_{w,w}(\boldsymbol{\Delta}) = \mathrm{E}[\mathrm{E}[w(\mathbf{n_i})w(\mathbf{n_i} + \boldsymbol{\Delta})|[w(\mathbf{n_i})]]$, we get

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

We will call a watermark generation process "white and DC-free" if its autocorrelation function is as described above. Its spatial spectrum components are flat (except at DC).

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

¿¿¿ D_0

We will call a watermark "white and Dc-free" if $\Gamma_{w,w} = NE_w$ and

A "purely white" watermark requires that the correlation equals exactly zero outside $\boldsymbol{\Delta} = \mathbf{0}$. We have seen that purely white marks cannot be absolutely DC free, but $D_0 =$

¿¿¿ NB white a statistical property of the generation process, not of the watermark

As an other example, we will treat the case that the watermark has a low-pass spatial spectrum. This method has been advocated by for instance by Cox et al. [?]. In such situation, a potential attacker can not easily remove the watermark by low-pass filtering. Moreover, JPEG compression typically removes or distorts high-frequency components. A low-pass a watermark can be generated by spatially filtering a spatially white watermark. Perceptually this appears as a smoothing. A first-order two dimensional IIR spatial smoothing filter computes

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

It can be shown that in case of a statistically DC-free watermark $\mathbf{w_1}$, a first-order filter generates a new watermark $\mathbf{w_2}$ with correlation function

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

Another method of generating a spatially shaped watermark is to use a random generator which gives a correlated output for neighboring pixels.

* 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 fundamental to the requirement of the DC value.

figure=../waterm1.epsi,height=3cm

Fig. 1. Watermark Embedder and Correlation Detector

III Correlator detector

Correlator detectors are interesting to study, for several reasons. They are a mathematical generalization of the simple scheme in which watermarks with $w \in \{-1, 0, +1\}$. 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 values in which the watermark is negative, i.e., $s_- = \sum_{\mathbf{n}\in A_-} q(\mathbf{n})$ and the sum of all pixel values in which the watermark is positive, i.e., $s_+ = \sum_{\mathbf{n}\in A_+} q(\mathbf{n})$. Then, $y = s_+ - s_-$ is used as a decision variable, e.g. [**?**][**?**]. From our more general results to follow it can be concluded that

- its performance highly depends on whether the probability that $\mathbf{n_i} \in A_-$ statistically depends on whether $\mathbf{n_j} \in A_-$ for some pair of differing pixel locations $\mathbf{n_i} \neq \mathbf{n_j}$. High correlation between pixels in A_- (and those in A_+) substantially reduces reliability.
- 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, this is significantly worse than when the number of elements is always precisely the same. This is in contrast to our intuition that if pixels are put in A_- and A_+ with probability $1/2$, the statistical effect of a differing number of elements in each class becomes negligibly small for increasing image sizes. Our results will show that this is not the case.

Another reason to address correlators is that are these are known to be the optimum detector for particular situations often encountered in radio communication, namely the Linear Time-Invariant (LTI), frequency non-dispersive, Additive Gaussian Noise (AWGN) channel, when the receiver has full knowledge about the alphabet of waveforms used to transmit a message.

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 $\hat{w}(\mathbf{n}) = w(\mathbf{n})$, so $y = R_{w,q}(\mathbf{0})$, with

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

Figure 2 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 watermark data is added to DCT coefficients. For our analysis, we

separate y into a deterministic contribution y_w from the watermark,

$$y_w = \frac{1}{N}\sum_{\mathbf{n}\in A} \hat{w}(\mathbf{n})w(\mathbf{n}) = R_{w,w}(\mathbf{0}) = \frac{E_w}{N}$$

plus filtered noise from the image y_p

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

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

$$\mathrm{E}y_p = \frac{1}{N}\mathrm{E}\sum_{\mathbf{n}\in A} \hat{w}(\mathbf{n})p(\mathbf{n}) = \frac{\mathrm{E}p(\mathbf{n})}{N}\sum_{\mathbf{n}\in A} \hat{w}(\mathbf{n}) = \mu_1\hat{D}_0$$

This result appears to be irrespective of the correlation in pixels. To find the second moment, we compute

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

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

Here, E_p denotes an expectation over all images. 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 in the following sections we will show that for correlated pixels ($\alpha > 0$) and spectrally non-white watermarks, non-zero cross terms substantially affect the results, even if $D_0 = 0$.

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 in section V. 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. Since y_w equals E_w/N,

$$P_{md} = \frac{1}{2}\mathrm{erfc}\frac{E_w + \mu_1 N\hat{D}_0 - N_1N_2y_{thr}}{\sqrt{2E_w}\sigma}$$

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

On the other hand, given that no watermark is embedded, a *false alarm* 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: White and DC-free watermark

The 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

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

For a 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 the ignorance of boundary effects. To be more precise, we consider $\Gamma_{w,w}(\boldsymbol{\Delta})R_{\tilde{p},\tilde{p}}(\boldsymbol{\Delta})$ to vanish rapidly enough with increasing $\boldsymbol{\Delta}$ to allow the following approximation: we consider the summings over $\boldsymbol{\Delta}$ to cover the entire plane R^2 even though the size of the image is finite and $\mathbf{n}+\boldsymbol{\Delta}$. This allows us to write

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

We assume $\Gamma_{\tilde{p},\tilde{p}} \approx R_{\tilde{p},\tilde{p}}$.

For a binary watermark with embedding depth k, thus with $w(\mathbf{n}) \in \{-k, +k\}$, this gives

$$\mathrm{E}y_p^2 = \frac{k^2\sigma^2}{N} - \sum_{\boldsymbol{\Delta}\neq\mathbf{0}}\frac{k^2}{N(N-1)}[\sigma^2\alpha^{|\Delta_1|+|\Delta_2|}]$$

We get

$$Ey_p^2 = \frac{k^2\sigma^2}{N-1} - \frac{k^2\sigma^2}{N(N-1)}\left[\frac{1+\alpha}{1-\alpha}\right]^2$$

We see that the effect of pixel correlations is significant only if α is very close to unity (little luminance changes) in a small-size image. If the image is large enough, that is, if $N >> \left[\frac{1+\alpha}{1-\alpha}\right]^2$, we may approximate

$$Ey_p^2 \approx \frac{k^2\sigma^2}{N}$$

In practical situations, this appears a reasonable approximation. Inserting value of the standard deviation σ_{y_p}, the error probability becomes

$$P_{fa} = \frac{1}{2}\mathrm{erfc}\frac{Ny_{thr} - \mu_1 N\hat{D}_0}{\sqrt{2E_w}\sigma}$$

In Figure 3, 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 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 to image noise ratio E_w/σ^2, expressed in dB ($10\log_{10}\left(E_w/\sigma^2\right)$). We have chosen this definition as it best matches common practice in statistical communication to use E_b/N_0 where E_b is the average energy per bit, and N_0 is the spectral power density of the noise. Note that E_w/σ^2 typically is much larger that unity for reliable detection, but this does not imply that the watermark $w(\mathbf{n_i})$ in a particular pixel exceeds the luminance variations of the image. 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, it is 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, examples 3 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 "waveform" used. This is in contrast to spread spectrum radio over AWGN channels.

III.2 Example 2: Non-DC free watermarks

A popular method to generate a watermark is to independently randomly choose pixel values $w(\mathbf{n}) \in \{-k, +k\}$, which then may or may not be filtered by a two-dimensional first-order filter. That is, $\mathrm{E}[D_0] = 0$, but individual realizations of the watermark may not be DC-free. Intuitively one may believe that if N is large enough, setting a fixed detection threshold accounting for $\mathrm{E}[D_0] = 0$ but without specifically compensating for D_0 will not affect the performance significantly. Here we prove different.

We address the ensemble-mean behavior, that is we average over all possible watermarks generated this way. We assume that image and watermark are independent, so

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

If we ignore boundary effects, we get

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

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

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

If we ignore boundary effects, we get

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

and, as we saw befor $\mathrm{E}_p[y_p] = \mu_1 D_0$, so

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

This result is surprising. Inituitively one may expect that for $\beta \to 0$, this would reduce to

$$\mathrm{E}[y_p^2] = \frac{E_w\sigma^2}{N^2}$$

but it tends to $E_w\mu_2/N^2$ where $\mu_2 = \sigma^2 + \mu_1^2$. This difference is due to the fact that the watermark has a random DC-component. Note that for randomly chosen watermark pixel values, the running DC components conducts a random walk. In fact, the term containing μ_1 accounts for the fluctuations in the DC component $\sum w(\mathbf{n})$.

For a fixed threshold $y_{thr} = E_w/(2N)$, the error rate goes into

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

whereas for the threshold setting $y_{thr} = \mu_1 D_0 + E_w/(2N)$, we find

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

In practice we found that typically μ_2 is about four times larger than σ^2. Performance is different by about 5 to 10 dB. 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.

III.3 Example 3: Low-pass and DC-free watermark

Similar to the above analysis, one can show that for a DC-free low-pass watermark,

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

The error rate goes into

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

figure=../unfiltered.epsi,height=10cm

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": "×": absolutely DC-free "+": random watermark, independent pixels, DC-free i.m. Solid lines: corresponding theoretical curves.

In other words, the variance of the interference increases with increasing β. The stronger the watermark is limited to low-pass spatial components, the more difficult or unreliable the detection becomes.

This result clearly shows that if the watermark is confined to low-pass components of the image, this significantly affects the reliability of detection. 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).

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. In our experiments, we approximated white and absolutely DC-free watermarks through 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 [**?**] [**?**] of length $2^{14} - 1 = 127 * 129$. Such sequences have a negligibly small DC component $\sum_{\mathbf{n}} w(\mathbf{n}) = -1$ and a correlation function that 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.

Figure 3 compares the above theoretical results with measurements of the "Lenna" image. In the figure, we combined the results from one image and many watermarks to get statistical results. We computed the components of the decision variable and estimated which signal-to-noise ratio would be needed to achieve reliable detection. Any particular image with a particular watermark gives a step-wise transition at $y_w = 2y_p$. Combination of this step for many (10^4 or more) 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, we must choose $k \geq 2y_p$, so $E_w = k^2 N_1 N_2 \geq 4y_p^2 N_1 N_2$. To express this in terms of the signal-to-noise ratio $\gamma = E_w/\sigma^2$, we used sample moments of the "Lenna" image to estimate σ. Other images gave essentially the same curves. The shape of the curve confirm the approximately Gaussian behaviour of y_p.

To get consistent results, we had to generate a large set of watermarks fully independently. If one simply shifts a single 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 leads to a significant deviation from the theoretical curve and to a more stepwise (non-Gaussian) decrease of errors rates with increasing SNR. Images with higher correlation (such as "Teeny 1") are more sensitive to correlations among different watermark during the experimental evaluation.

V Conclusions

In this paper, we proposed a mathematical framework to model electronic watermarks embedded in digital images. The model regards the process of embedding and transferring a watermark to be similar to that of communication channel. It treats the original contents (the image itself) as interference or noise.

We observe that many detectors proposed for watermarks are of the correlator type, though often with minor modifications. Several essential differences appear with the case of 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, absolutely DC-free 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 content of the watermark appears another critical parameter. If the watermark is non-white, the spectral properties of the images are also of significant influence.

References

1. B.M. Macq and J.J. Quisquater, "Cryptology for digital tv broadcasting" Proc. of the IEEE, Vol. 83 No. 6, 1993, pp. 944-957
2. R.G. van Schyndel, A.Z. Tirkel, C.F. Osborne: "A Digital Watermark", Int. IEEE Conf on Image Processing, Vol.2., 13-16 Nov. 1994, IEEE Comput. Soc. Press, Los Alamitos, CA, USA, pp. 86-90
3. W. Bender, D. Gruhl, N. Morimoto, "Techniques for Data Hiding", Proceedings of the SPIE, 2420:40, San Jose CA, USA, February 1995
4. I. Pitas, T. Kaskalis : "Signature Casting on Digital Images", Proceedings IEEE Workshop on Nonlinear Signal and Image Processing, Neos Marmaras, June 1995

5. E. Koch, J. Zhao : "Towards Robust and Hidden Image Copyright Labeling". Proceedings IEEE Workshop on Nonlinear Signal and Image Processing, Neos Marmaras, June, 1995
6. Caronni G.: "Assuring Ownership Rights for Digital Images", Proceedings of Reliable IT Systems, VIS '95, Vieweg Publishing Company, Germany, 1995
7. F.M. Boland, J.J.K. O Ruanaidh, C. Dautzenberg, "Watermarking Digital images for Copyright Protection", Proceedings of the 5th IEE International Conference on Image Processing and its Applications, no. 410, Edinburgh, July, 1995, pp. 326-330.
8. J. Zhao, E. Koch : "Embedding Robust Labels into Images for Copyright Protection", Proceedings of the International Congress on Intellectual Property Rights for Specialized Information, Knowledge and New Technologies, Vienna, Austria, August 1995
9. I.J. Cox, J. Kilian, T. Leighton, T. Shamoon "Secure Spread Spectrum Watermarking for Multimedia", NEC Research Institute, Technical Report 95 - 10
10. W. Bender, D. Gruhl, N. Morimoto and A. Lu, "Techniques for data hiding", IBM Systems Journal, Vol. 35. No. 3/4 1996
11. 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
12. 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
13. I.J. Cox and J.P.M.G. Linnartz, "Public watermarks and resistance to tampering", accepted for presentation at Int. Conf. on Image Processing (ICIP) 1997.
14. N.S. Jayant and P. Noll., "Digital Coding of waveforms" Prentice Hall, 1984.
15. Ch. W. Therrien, "Discrete Random Signals and Statistical Signal Processing" Prentice Hall, 1992.
16. F.J. McWilliams and N.J.A. Sloane, "Pseude-Random Sequences and arrays", Proc . of IEEE, Vol. 64, No.12, Dec. 1976, pp. 1715-1729
17. 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

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