IMPROVED WATERMARK DETECTION RELIABILITY USING FILTERING BEFORE CORRELATION

Geert Depovere *Ton Kalker* *Jean-Paul Linnartz*

Philips Research
Prof. Holstlaan 4
5656 AA Eindhoven
The Netherlands
{depovere,kalker,linnartz}@natlab.research.philips.com

ABSTRACT

A digital watermark is a perceptually unobtrusive signal embedded in some multimedia asset carrying additional information: e.g. copyright information of a video clip. For nearly all watermarking schemes published so far, detection is based on some form of correlation. In this paper, we show that the detection reliability can be significantly improved by applying filtering prior to correlation. This improvement is analysed using a theoretical model based on statistical communication and detection theory. Finally, the improvements predicted by the theory are verified in a number of experiments.

1. INTRODUCTION

Electronic watermarking is an enabling technology for the distribution of digital multimedia content. It aims at embedding additional data into clear content (images, audio etc) in a way that is difficult to remove [1, 2, 3, 4].

Principal applications of electronic watermarks are in copyright enforcement, automatic metering and monitoring of asset usage in multi-media applications, piracy tracing, and in providing additional information, such as image captions. In many cases watermark detection amounts to thresholding a correlation computation. 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 pattern of absent and present watermark patterns carries the embedded information.

Correlation detection is only optimal in the case that the signal can be modelled as additive white Gaussian noise. In this paper a correlation detection receiver is proposed which includes prefiltering to obtain optimum detection in the case of real images in which the power spectrum is not white. The analysis presented in this paper is based on the approach presented in the paper [5]. Some experiments are described which verify the theoretical improvement of the watermark detection reliability.

2. FORMULATION OF THE MODEL

2.1. Introduction

For the formulation of the theory, we consider two stochastical processes: $\mathcal{W}$ which generates watermarks and $\mathcal{P}$ which generates images. The watermarks and images have a size of N_1 by N_2 pixels with a total of $N = N_1 N_2$ pixels. The intensity level of the pixel with coordinates $n = (n_1, n_2)$, $(0 \leq n_1 \leq N_1 - 1, 0 \leq n_2 \leq N_2 - 1)$ is denoted as $p(n)$. The set of all pixel coordinates is denoted as A. We restrict our discussion to gray scale images in which $p(n)$ takes on real or integer values in a certain interval.

Whenever convenient we will represent $p(n)$ as a z-expression $p(z)$ defined by

$$\begin{aligned} p(z) &= \sum_{n \in A} p(n) z^{-n} \\ &= \sum_{n \in A} p(n) z_1^{-n_1} z_2^{-n_2} \end{aligned}$$

We will assume that both stochastic processes $\mathcal{W}$ and $\mathcal{P}$ are wide-sense stationary and ergodic. By wide-sense stationarity the statistical k^{th} moment

$$\mu_k[p(n)] = \mathrm{E}[\, p^k(n) \,]$$

becomes $\mu_k(p)$ and the statistical autocorrelation function

$$R_{p,p}(n, m) = \mathrm{E}[\, p(n) p(m) \,]$$

becomes $R_{p,p}(n-m)$. The correlation only depends on the difference vector $\Delta = (\Delta_1, \Delta_2) = (n - m)$. By ergodicity we are allowed to approximate the statistical k^{th} moment $\mu_k(p)$ by the spatial k^{th} moment

$$m_k(p) = \frac{1}{N} \sum_{n \in A} p^k(n)$$

0-8186-8821-1/98 $10.00

and the statistical autocorrelation function $R_{p,p}(\Delta)$ by the spatial autocorrelation function

$$\langle p,p\rangle(\Delta) = \frac{1}{N}\sum_{n\in A} p(n)p(n+\Delta),$$

where we assume $n+\Delta$ to wrap around when it formally falls outside the set A. Note that we used a different notation to distinguish between the statistical and spatial moments and correlations.

2.2. Image Model

In this paper images are modelled by assuming a first-order separable autocorrelation function [6]

$$\langle p,p\rangle(\Delta) = m_1^2(p) + s_p^2\alpha^{|\Delta|},$$

where $|\Delta| = |\Delta_1| + |\Delta_2|$. The standard deviation s_p is defined as $s_p^2 = m_2(p) - m_1^2(p)$. This assumption seems a crude approximation of the typical properties of images. However, from experiments such as those to be reported in Section 5, it appears that error rates based on this crude model can be reasonably accurate for the purpose of this evaluation. This assumption excludes certain images, such as binary images or computer-generated images with a limited number of colors.

The quantities $m_1(p)$ and $m_2(p)$ are referred to as the *DC-component* and the *energy* of the image p, respectively. The value α can be interpreted as a measure of the correlation between adjacent pixels in the image. Experiments reveal that typically $\alpha \approx 0.8 \ldots 0.99$.

It should be noted that for $\alpha \to 0$ the autocorrelation approaches a Dirac distribution:

$$\langle p,p\rangle(\Delta) = m_1^2(p) + s_p^2\delta(\Delta)$$

and the pixels become totally uncorrelated. We denote $\tilde{p}(n)$ as the non-DC component of the image, that is $\tilde{p}(n) = p(n) - m_1(p)$, so

$$\langle \tilde{p},\tilde{p}\rangle(\Delta) = s_{\tilde{p}}^2\alpha^{|\Delta|}. \qquad (1)$$

In the following, we will always assume that in the watermark detector all signals have been processed by subtracting the DC-component such that e.g.

$$\tilde{p} = p.$$

2.3. Watermark Model

A watermark $w(n)$ is modelled as a sample image drawn from the stochastical process $\mathcal{W}$ which is often implemented as a cryptographic process.

The energy in a watermark w equals $\langle w,w\rangle(0) = m_2(w)$ and is denoted as $E(w)$. In the following, we will always assume that the watermark w is DC-free, i.e. $m_1(w) = 0$. Similarly as in the case of images we can write that:

$$\tilde{w} = w.$$

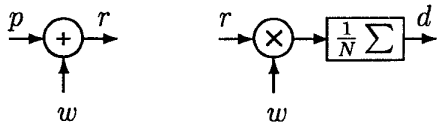

Figure 1: Watermark Embedding and Correlation Detection.

White watermarks reasonably model most of the early proposals for increasing and decreasing the pixel luminance according to a pseudo random process. The spatial autocorrelation function of a white watermark approaches the Dirac distribution when the image size is large enough:

$$\langle w,w\rangle(\Delta) = E(w)\delta(\Delta)$$

Low-pass watermarks are generated by spatially filtering a white watermark source $\mathcal{W}$ with a first-order two dimensional spatial smoothing IIR filter $S_\beta(z)$,

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

Such watermarks have been proposed amongst others in [4] because they are more robust against filtering and compression [7]. In this case the autocorrelation becomes:

$$\langle w,w\rangle(\Delta) = E(w)\beta^{|\Delta|} \qquad (2)$$

3. CORRELATION DETECTION

In Figure 1 the watermark embedding is depicted as

$$r(n) = p(n) + w(n)$$

Correlator detectors are a mathematical generalization of the simple device in which watermarks with $w(n) \in \{-1,0,+1\}$ are detected by computing the normalized sum of all pixel values in which the watermark is negative, i.e., $s_- = \frac{1}{N}\sum_{n:w(n)=-1} r(n)$ and the normalized sum of all pixel values in which the watermark is positive, i.e., $s_+ = \frac{1}{N}\sum_{n:w(n)=1} r(n)$. Then, $d = s_+ - s_-$ is used as a decision variable, e.g. [1].

In the correlator detector, also illustrated in Figure 1, a decision variable d is extracted from the received image $r(n)$ by correlating with a locally stored copy of the watermark $w(n)$, i.e.

$$\begin{aligned} d &= d_r(w) = \langle r,w\rangle(0) \\ &= d_p(w) + d_w(w). \end{aligned}$$

This model covers all detectors in which the decision variable is a linear combination of pixel luminance values in the image. The value $d_w(w)$, which is known to the detector is equal to the energy $E(w)$ of the watermark w. It should be noted that the expected value of $d_p(w)$ equals 0. Because of the Central Limit Theorem, $d_p(w)$ has a Gaussian distribution if N is sufficiently large and if the contributions in the sums are sufficiently independent. If we apply a threshold

$$d_{\text{thr}} = \frac{d_w(w)}{2} = \frac{E(w)}{2}$$

the probability of a *false negative* P_- (the watermark is present, but the detector decides it is not) equals the probability of a *false positive* P_+ (the watermark is not present, but the detector decides it is)

$$\begin{aligned} P &= P_+ = P_- \\ &= \frac{1}{2}\operatorname{erfc}\left(\frac{E(w)}{2\sqrt{2}\sigma_{d_p(w)}}\right). \end{aligned} \tag{3}$$

The standard deviation of $d_p(w)$ is computed as:

$$\begin{aligned} \sigma^2_{d_p(w)} &= \mathrm{E}[\langle p, w\rangle(0)^2] \\ &= \frac{1}{N^2}\mathrm{E}[\sum_{n,m\in A} w(n)p(n)w(m)p(m)] \\ &= \frac{1}{N}\sum_{\Delta\in A} R_{w,w}(\Delta)R_{p,p}(\Delta). \end{aligned}$$

Using the image model of Equation 1 and a low-pass watermark given by Equation 2 we can write:

$$\begin{aligned} \sigma^2_{d_{p(w)}} &= \frac{E(w)\sigma_p^2}{N}\sum_{\delta\in A}(\alpha\beta)^{|\delta|} \\ &= \frac{E(w)\sigma_p^2}{N}\left[\frac{1+\alpha\beta}{1-\alpha\beta}\right]^2. \end{aligned}$$

The error rate goes into

$$P = \frac{1}{2}\operatorname{erfc}\left(\sqrt{\frac{E(w)N}{8\sigma_p^2}\frac{1-\alpha\beta}{1+\alpha\beta}}\right), \tag{4}$$

which in the case of a white watermark $\beta \to 0$ or in the case of an image with uncorrelated pixels $\alpha \to 0$ becomes

$$P = \frac{1}{2}\operatorname{erfc}\left(\sqrt{\frac{E(w)N}{8\sigma_p^2}}\right). \tag{5}$$

We see that correlation of the pixels (the case that $\alpha \neq 0$) leads to a loss of reliability if the watermark has a low-pass spectrum. The use of a low-pass watermark also leads to a loss of reliability (the case that $\beta \neq 0$) if the pixels are correlated. However, such a low-pass watermark is less vulnerable to erasure by image processing [7], such as low-pass filtering (smoothing).

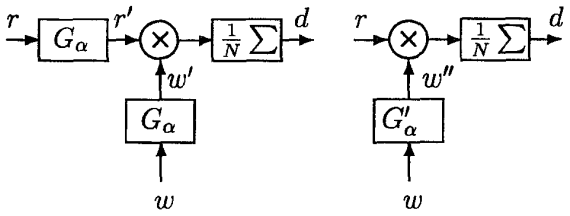

Figure 2: Correlation Detector comprising two whitening filters (left) and one filter (right).

4. WHITENING

From detection theory it follows that correlation detectors are optimum in the case of a Linear Time-Invariant (LTI), frequency non-dispersive, Additive White Gaussian Noise (AWGN) channel. However, this is not the case for real images where the pixels are correlated and the autocorrelation can be modelled by Equation 1. However, further applying standard detection theory, it is still possible to achieve optimum detection, in the case of non-white Gaussian noise, by applying a so-called *whitening filter* at the input of the correlation receiver. This filter transforms the non-white input signal of the receiver to a signal with a constant power spectrum. Of course, also the watermark signal has to be filtered in the same way before correlation, as indicated in Figure 2. If we apply the inverse first-order two dimensional spatial smoothing IIR filter $G_\alpha(z) = S_\alpha^{-1}(z)$,

$$G_\alpha(z) = \frac{1}{1-\alpha^2}(1-\alpha z_1^{-1})(1-\alpha z_2^{-1}),$$

We find that $p'(z) = G_\alpha(z)p(z)$ and $w'(z) = G_\alpha(z)w(z)$ such that

$$R_{p',p'}(\Delta) = \sigma^2_{p'}\delta(\Delta).$$

Similar as in Chapter 3 we can write:

$$\begin{aligned} d &= d_{r'}(w') = \langle r', w'\rangle(0) \\ &= \frac{1}{N}\sum_{n\in A}\sum_{\Delta\in A} g_\alpha(n-\Delta)r(\Delta)\sum_{\gamma\in A} g_\alpha(n-\gamma)w(\gamma) \\ &= \sum_{\Delta,\gamma\in A} r(\Delta)w(\gamma)g'_\alpha(\Delta-\gamma) \\ &= \frac{1}{N}\sum_{\Delta\in A} r(\Delta)w''(\Delta) \\ &= d_p(w'') + d_w(w''). \end{aligned}$$

Where we denote $g'_\alpha(\Delta) = \langle g_\alpha, g_\alpha\rangle(\Delta)$ and

$$w''(n) = \sum_{\Delta\in A} w(\Delta)g'_\alpha(n-\Delta).$$

This means that both filters $G_\alpha(z)$ in Figure 2 can be replaced by one filter $G'_\alpha(z)$. For the deterministic quantity $d_w(w'')$ it follows that:

$$d_w(w'') = \sum_{\Delta \in A} g'_\alpha(\Delta)\langle w, w\rangle(\Delta).$$

Similarly, for the stochastic quantity $d_p(w'')$ it follows that:

$$d_p(w'') = \sum_{\Delta \in A} g'_\alpha(\Delta)\langle w, p\rangle(\Delta).$$

The expected value $\mathrm{E}[\,d_p(w'')\,] = 0$ because we assume that the watermark source is DC-free. The standard deviation of $d_p(w'')$ is computed as:

$$\begin{aligned}\sigma^2_{d_p(w'')} &= \mathrm{E}[(\frac{1}{N}\sum_{\Delta \in A} g'_\alpha(\Delta)\sum_{n \in A} w(n)p(n-\Delta))^2] \\ &= \sum_{n \in A} R_{w,w}(n)\sum_{\Delta \in A} R_{p,p}(n-\Delta)\langle g'_\alpha, g'_\alpha\rangle(\Delta).\end{aligned}$$

Using the image model from Equation 1 and a low-pass watermark as presented in Equation 2 this becomes:

$$\begin{aligned}\sigma^2_{d_p(w'')} &= \frac{E(w)\sigma_p^2}{N}\left(\sum_{n,\Delta \in A} \beta^{|n|}\alpha^{|n-\Delta|}\langle g'_\alpha, g'_\alpha\rangle(\Delta)\right)^2 \\ &= \frac{E(w)\sigma_p^2}{N}\left(\frac{1+\alpha^2-2\alpha\beta}{1-\alpha^2}\right)^2\end{aligned}$$

and we find that:

$$d_w(w'') = E(w)\left(\frac{1+\alpha^2-2\alpha\beta}{1-\alpha^2}\right)^2.$$

We can again apply the Central Limit Theorem and by setting a threshold d_{thr} given by

$$d_{\mathrm{thr}} = \frac{d_w(w'')}{2}.$$

The error probability becomes:

$$\begin{aligned}P &= \frac{1}{2}\,\mathrm{erfc}\left(\frac{d_{w''}(w)}{2\sqrt{2}\sigma_{d_{p''}(w)}}\right) \qquad (6)\\ &= \frac{1}{2}\,\mathrm{erfc}\left(\sqrt{\frac{E(w)N}{8\sigma_p^2}\frac{1+\alpha^2-2\alpha\beta}{1-\alpha^2}}\right).\end{aligned}$$

In the case of $\alpha = 0$ this reduces to Equation 5. In the case of a white watermark ($\beta = 0$) the error rate becomes

$$P = \frac{1}{2}\,\mathrm{erfc}\left(\sqrt{\frac{E(w)N}{8\sigma_p^2}\frac{1+\alpha^2}{1-\alpha^2}}\right). \qquad (7)$$

If we compare Equations 6 and 4 we see that due to the whitening process, the watermark energy has been amplified by an improvement η given by

$$\eta = \left(\frac{1+\alpha^2-2\alpha\beta}{1-\alpha^2}\frac{1+\alpha\beta}{1-\alpha\beta}\right)^2.$$

For a white watermark this simplifies to:

$$\eta = \left(\frac{1+\alpha^2}{1-\alpha^2}\right)^2.$$

5. COMPUTATIONAL AND EXPERIMENTAL RESULTS

In our experiments, we approximated white watermarks through pseudo-random sequences. Non-white watermarks were generated using the smoothing filter of Section 2.3. We used the central 256 × 256 part of the image "Lena" as depicted in Figure 4 to perform the experiments. For this image we found experimentally that $\alpha = 0.9$.

Figure 3 compares the theoretical results for the error rate with measurements in the absence of a whitening filter with a white watermark ('x') and in the case of a whitening filter ($\alpha = 0.9$) and a white watermark ('o'), a low-pass watermark with $\beta = 0.3$ ('*') and a a low-pass watermark with $\beta = 0.6$ ('+'). We plotted the error rates as defined in Section 3 and Section 4 versus the Signal to Noise Ratio (SNR) defined as

$$SNR = 10\log_{10}\left(\frac{E(w)N}{\sigma_p^2}\right).$$

The figure shows a good agreement between theory and experiments.

6. CONCLUSIONS

In this paper, we showed that the detection reliability can be significantly improved by applying filtering prior to correlation. These improvements were analysed using a theoretical model based on statistical communication and detection theory. This model treats the original content (the image itself) as interference noise in a communication channel. Finally, the improvements predicted by the theory were verified in a number of experiments.

7. REFERENCES

[1] W. Bender, D. Gruhl, and N. Morimoto, "Techniques for data hiding," in *Proceedings of the SPIE*, (San Jose CA, USA), pp. 2420–2440, February 1995.

[2] I. Pitas and T. Kaskalis, "Signature casting on digital images," in *Proceedings IEEE Workshop on Nonlinear Signal and Image Processing*, (Neos Marmaras), June 1995.

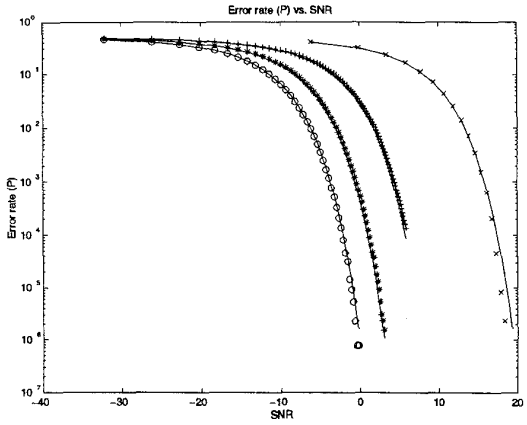

Figure 3: Theoretical results (solid lines) and experiments for the error rate without whitening filter with a white watermark ('x') and in the case of a whitening filter with $\alpha = 0.9$ for a white watermark ('o'), for a low-pass watermark with $\beta = 0.3$ ('*') and for a low-pass watermark with $\beta = 0.6$ ('+').

Figure 4: Part of the 'Lena' image used in the experiments.

[3] E. Koch and J. Zhao, "Towards robust and hidden image copyright labeling," in *Proceedings IEEE Workshop on Nonlinear Signal and Image Processing*, (Neos Marmaras), June 1995.

[4] I. Cox, J. Kilian, T. Leighton, and T. Shamoon, "A secure, robust watermark for multimedia," in *Proceedings of the Workshop on Information Hiding*, (Univ. of Cambridge, U.K.), pp. 175–190, May 1996.

[5] J. Linnartz, T. Kalker, G. Depovere, and R. Beuker, "A reliability model for the detection of electronic watermarks in digital images," in *Proceedings IEEE Fifth Symposium on Communications and Vehicular Technology*, pp. 202 – 209, October 1997.

[6] N. Jayant and P. Noll, *Digital Coding of waveforms*. Prentice Hall, 1984.

[7] I. Cox and J. Linnartz, "Public watermarks and resistance to tampering," in *Proceedings of the IEEE Int. Conf. on Image Processing (ICIP)*, October 1997. Paper appears only in CD version of proceedings.