Watermark Detection after Quantization Attacks

Joachim J. Eggers and Bernd Girod

Telecommunications Laboratory, University of Erlangen-Nuremberg
Cauerstr. 7/NT, 91058 Erlangen, Germany
{eggers,girod}@LNT.de

Abstract. The embedding of additive noise sequences is often used to hide information in digital audio, image or video documents. However, the embedded information might be impaired by involuntary or malicious "attacks." This paper shows that quantization attacks cannot be described appropriately by an additive white Gaussian noise (AWGN) channel. The robustness of additive watermarks against quantization depends strongly on the distribution of the host signal. Common compression schemes decompose a signal into sub-signals (e.g., frequency coefficients) and then adapt the quantization to the characteristics of the sub-signals. This has to be considered during watermark detection. A maximum likelihood (ML) detector that can be adapted to watermark sub-signals with different robustness is developed. The performance of this detector is investigated for the case of image watermark detection after JPEG compression.

1 Introduction

In the last decade, the digital representation of continuous signals (e.g., audio, images, video) has become very popular due to the possibility of efficient transmission and copying without quality degradation. On the other hand, unauthorized copying is also simplified. One approach to solve this problem is to mark a valuable digital document such that a copyright can be proven or the distribution path be reconstructed. The watermarking process produces a perceptually equivalent digital document rather than a bit-exact copy.

A general watermark embedding scheme can be described by

$$\underline{s}_k = \underline{x} + \underline{w}_k\,, \tag{1}$$

where $\underline{w}_k$ denotes the signal modification introduced by the watermarking process, $\underline{x}$ the original document, and $\underline{s}_k$ the published document (watermarked document). $\underline{x}$ is also called "host signal" or "private document". In many schemes $\underline{w}_k$ is explicitly given, but there are also schemes where the signal modification depends on the private document $\underline{x}$. In the remainder of this article, signals are denoted by vectors (e.g. $\underline{x}$), th nth signal sample by $x[n]$, and random variables by boldface. Here, the index k allows for the possibility of embedding different watermarks. Later, the index will be also used to denote sub-signals.

The watermark detector receives a signal

$$\underline{r}_k = \underline{s}_k + \underline{e} = \underline{x} + \underline{w}_k + \underline{e}, \tag{2}$$

where $\underline{e}$ denotes the distortion that might be introduced by the watermark channel. Here, only independent watermarks are considered. That is, $\underline{w}_i$ and $\underline{w}_j$, $j \neq i$, are independent of each other. In this case, a different watermark than the one to be detected appears as additive noise. This noise can be included in $\underline{e}$ and thus, the index k can be neglected.

A complete characterization of the watermark channel is still an open problem. In contrast to many other communications problems, the channel distortion might be introduced intentionally to remove or obscure the transmitted information (the watermark). Besides attacks that exploit possible weaknesses of protocols for watermarking schemes, desynchronization and compression attacks usually are most successful. The latter will be discussed in this article. Therefore, we assume perfect synchronization of the watermark detector. This assumption is not too restrictive, since many desynchronization attacks can be counter-attacked by improved detectors [5, 13].

In Section 2, watermark detection is discussed. We derive a decision rule that can be adapted to the different robustness of different parts of an embedded watermark. A detailed analysis of watermark detection after scalar quantization is presented in [1]. This analysis is based on the theory of dithered quantizers, as described e.g. in [3, 6, 10]. Due to space constraints, here only the main aspects of this analysis are summarized in Section 3. To demonstrate the importance of the detection problem after quantization attacks, we discuss an example image watermarking scheme. This scheme is described in Section 4, and, in Section 5, the corresponding detection results are discussed. Section 6 concludes the paper.

2 Watermark Detection

Signal detection has been intensively analyzed by communication engineers. However, the quantization attack is a very special transmission channel, thus we derive a special watermark detection scheme.

2.1 Bayes' Hypothesis Test

The watermark detection problem can be stated as a simple hypothesis test.

hypothesis H_0 : the watermark $\underline{w}$ is not present,
hypothesis H_1 : the watermark $\underline{w}$ is present.

The problem of hypothesis testing is to decide which of the hypotheses is true, when a document $\underline{r}$ ist given. Usually it is not possible to separate all watermarked and unwatermarked documents perfectly; a received signal $\underline{r}$ might be watermarked with probability $p\left(H_1|\underline{r}\right)$ or not watermarked with probability $p\left(H_0|\underline{r}\right)$. We can trade off the probability p_{FP} of accepting H_1 when H_0 is true (*false positive*) and the probability p_{FN} of accepting H_0 when it is false (*false negative*). Bayes' solution is the decision rule

$$\frac{p_{\mathbf{r}}\left(\underline{r}|H_1\right)}{p_{\mathbf{r}}\left(\underline{r}|H_0\right)} \begin{cases} > K \Rightarrow \text{accept } H_1 \\ \leq K \Rightarrow \text{accept } H_0, \end{cases} \tag{3}$$

where $K = \text{cost}_{p_{FP}} p_{H_0} / (\text{cost}_{p_{FN}} p_{H_1})$ is a constant depending on the a priori probabilities for H_1 and H_0 and the cost connected with the different decision errors [2]. For

$K = 1$, the decision rule (3) forms a **maximum-likelihood (ML) detector**. For equal a priori probabilities, the overall detection error probability is $p_e = \frac{1}{2}(p_{FP} + p_{FN})$. Receiver operating characteristic (ROC) graphs, as proposed in [8], can be computed using different thresholds K.

2.2 Correlation Detection

The watermark information is spread by an independent, mean-free, pseudo-noise sequence over the complete original signal. For an AWGN channel and $K = 1$ the hypothesis test (3) can be implemented as a correlation detector [4]:

$$H_1 : \quad \frac{\underline{r}^T \underline{w}}{M\sigma_{\mathbf{w}}^2} \; > \; \frac{1}{2}, \tag{4}$$

where $\sigma_{\mathbf{w}}^2$ denotes the watermark power and M is the signal length. The AWGN channel model implies that $\mathbf{x}$ and $\mathbf{e}$ are jointly Gaussian random processes and statistically independent from a possibly included watermark $\mathbf{w}$. In Section 3, it is shown that this assumption is not valid for a quantization attack.

2.3 Generalized Watermark Detection

The description of the watermark detection problem is generalized in this subsection, such that the characteristics of the quantization attack described in Section 3 can be exploited. We do not restrict the channel distortion to be AWGN. However, we assume to know a signal decomposition

$$x[n] = \sum_{i=1}^{i_{\max}} \sum_{m=0}^{\tilde{M}_i - 1} \tilde{x}_i[m] \psi_i \left[n - m\frac{M}{\tilde{M}_i} \right] \tag{5}$$

so that all $i_{\max}$ signals $c_i[m] = \tilde{r}_i[m]\tilde{w}_i[m]$ of length $\tilde{M}_i$ are white and stationary. The function $\psi_i\left[\cdot\right]$ denotes the ith function of the set of basis functions used for the decomposition. For instance, the decomposition could be a block-wise frequency transform, where m denotes the block index and i the frequency component. $\underline{\tilde{r}}_i$ and $\underline{\tilde{w}}_i$ are the sub-signals of $\underline{r}$ and $\underline{w}$ that are defined just like $\underline{\tilde{x}}_i$ in (5). The received sub-signals $\underline{\tilde{r}}_i$ are different for both hypotheses:

$$H_1 : \; \underline{\tilde{r}}_i = \underline{\tilde{x}}_i + \underline{\tilde{w}}_i + \underline{\tilde{e}}_{1i} \tag{6}$$

$$H_0 : \; \underline{\tilde{r}}_i = \underline{\tilde{x}}_i + \underline{\tilde{e}}_{0i}. \tag{7}$$

The channel distortion depends on the considered hypotheses and can even depend on the watermark in case of hypothesis H_1. The proper choice of a signal decomposition is not discussed further in this article. *The goal of the decomposition is to separate signal components that can hide watermarks with different robustness.* For the experiments described in Section 4, the 8×8 block-DCT is used. In this case, the basis functions are two-dimensional.

With help of a signal decomposition, watermark detection can be separated into two steps:

1. Estimate the expectations $\mathrm{E}\left\{\mathbf{c_i}\right\}$ with $i = 1, \ldots, i_{\max}$ by

$$\mathrm{E}\left\{\mathbf{c_i}\right\} \approx C_i = \frac{1}{\tilde{M}_i} \sum_{m=1}^{\tilde{M}_i} c_i[m], \tag{8}$$

   and combine these values to form the vector $\underline{C} = (C_1, C_2, \ldots, C_{i_{\max}})^T$. C_i is equal to the correlation of the sub-signals $\underline{\tilde{r}}_i$ and $\underline{\tilde{w}}_i$. For sufficiently large $\tilde{M}_i$, we can assume that the C_i are normally distributed with variance [9]

$$\mathrm{Var}\left\{\mathbf{C_i}\right\} = \frac{1}{\tilde{M}_i} \mathrm{Var}\left\{\mathbf{c_i}\right\}. \tag{9}$$

   Thus, the collection of all sub-channels $\underline{C}$ can be described by a multivariate Gaussian random variable $\underline{\mathbf{C}}$ with mean vector $\underline{C}_\mu$ and covariance matrix $\underline{\underline{\Phi}}_{\mathbf{CC}}^{H_1}$ in the case of hypothesis H_1 or with mean vector[1] $\underline{0}$ and covariance matrix $\underline{\underline{\Phi}}_{\mathbf{CC}}^{H_0}$ for H_0, respectively.

2. Apply the Bayesian hypothesis test (3) with the sample vector $\underline{C}$. Using the multivariate Gaussian PDF, the decision rule is given by

$$H_1: \quad K < \frac{(2\pi)^{-\frac{i_{\max}}{2}} \left|\underline{\underline{\Phi}}_{\mathbf{CC}}^{H_1}\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\underline{C} - \underline{C}_\mu)^T \underline{\underline{\Phi}}_{\mathbf{CC}}^{H_1\,-1} (\underline{C} - \underline{C}_\mu)\right)}{(2\pi)^{-\frac{i_{\max}}{2}} \left|\underline{\underline{\Phi}}_{\mathbf{CC}}^{H_0}\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\underline{C} - \underline{0})^T \underline{\underline{\Phi}}_{\mathbf{CC}}^{H_0\,-1} (\underline{C} - \underline{0})\right)}$$

   or equivalently

$$\begin{aligned} H_1: \quad \log(K) < \frac{1}{2} &\left(\log\left(\left|\underline{\underline{\Phi}}_{\mathbf{CC}}^{H_0}\right|\right) - \log\left(\left|\underline{\underline{\Phi}}_{\mathbf{CC}}^{H_1}\right|\right) - \underline{C}_\mu^T \underline{\underline{\Phi}}_{\mathbf{CC}}^{H_1\,-1} \underline{C}_\mu\right) \\ &+ \frac{1}{2}\underline{C}^T \left(\underline{\underline{\Phi}}_{\mathbf{CC}}^{H_0\,-1} - \underline{\underline{\Phi}}_{\mathbf{CC}}^{H_1\,-1}\right) \underline{C} + \underline{C}^T \underline{\underline{\Phi}}_{\mathbf{CC}}^{H_1\,-1} \underline{C}_\mu. \end{aligned} \tag{10}$$

The decision rule (10) has quadratic terms in $\underline{C}$ and thus defines a parabolic hypersurface in the $i_{\max}$-dimensional space. The analytic computation of the decision error probability is hard due to the quadratic terms. In addition, decision boundaries obtained from (10) are not very robust to an inaccurate estimation of the channel parameters. This will be demonstrated in Section 5. A simplification of rule (10) results by assuming equal covariance matrices $\underline{\underline{\Phi}}_{\mathbf{CC}}^{H_1}$ and $\underline{\underline{\Phi}}_{\mathbf{CC}}^{H_0}$, in which case the decision hyper-surface becomes a hyper-plane.

Fig. 1 shows a two-dimensional example. Here, the channel distortion is caused by the quantization of watermarks in two different DCT coefficients. More details are explained in Section 4. The measured correlations $\underline{C}$=$(C_{24}, C_{40})^T$ are plotted for both hypotheses and denoted by "$\times$" and "$+$" for H_1 and H_0, respectively. The samples leading to detection errors when using (10) are circled. The figure also depicts the parabolic and planar decision boundaries. Both rules are almost identical in the range where both hypotheses might be confused. Therefore, in practice both rules perform similarly, since they differ mainly in the regions of low probabilities.

[1] The conditional expectation $\mathrm{E}\left\{\underline{\mathbf{C}}|H_0\right\}$ is zero, since the watermark $\mathbf{w}$ is mean-free and independent from the unmarked document.

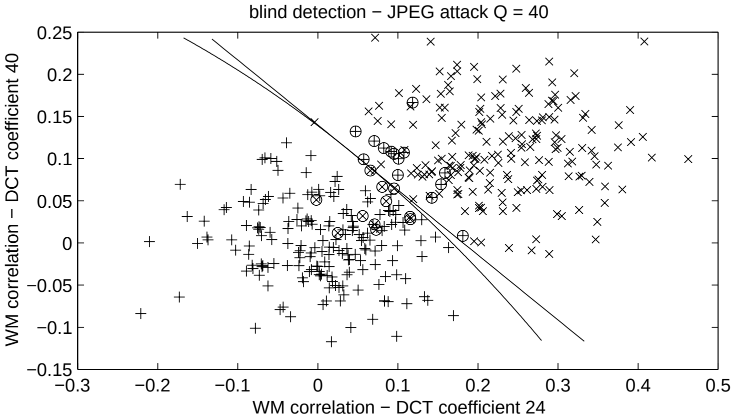

Fig. 1. Combined detection of two watermark components after a quantization attack

In Section 4 we will treat the 8×8 block-DCT as a decomposition that yields almost uncorrelated sub-channels. For uncorrelated sub-channels the covariance matrices $\underline{\underline{\Phi}}^{H_1}_{\mathbf{CC}}$ and $\underline{\underline{\Phi}}^{H_0}_{\mathbf{CC}}$ are diagonal and thus given by the conditional variances $\mathrm{Var}\left\{\mathbf{C_i}|H_1\right\}$ and $\mathrm{Var}\left\{\mathbf{C_i}|H_0\right\}$ for each channel. We assume $\sigma^2_{\mathbf{C_i}} = \mathrm{Var}\left\{\mathbf{c_i}\right\}/\tilde{M}_i = \mathrm{Var}\left\{\mathbf{c_i}|H_1\right\}/\tilde{M}_i = \mathrm{Var}\left\{\mathbf{c_i}|H_0\right\}/\tilde{M}_i$ to obtain a decision hyper-plane. Defining $\mu_{\mathbf{C_i}} = \mathrm{E}\left\{\mathbf{C_i}|H_1\right\}$ and setting $K = 1$ yields the detection rule

$$H_1:\ \sum_{i=1}^{i_{\max}} \left(\frac{C_i}{\mu_{\mathbf{C_i}}} - \frac{1}{2}\right)\frac{\mu^2_{\mathbf{C_i}}}{\sigma^2_{\mathbf{C_i}}} > 0. \tag{11}$$

The ratio $\alpha_i = \mu^2_{\mathbf{C_i}}/\sigma^2_{\mathbf{C_i}}$ can be interpreted as a weight for the correlation result computed for sub-channel i. This weight is largest for the sub-channels that provide the most robust watermark detection. The error probabilities for this detector are given by

$$p_{FP} = \frac{1}{2}\mathrm{erfc}\left(\frac{\frac{1}{2}\sum_{i=1}^{i_{\max}}\alpha_i}{\sqrt{2\sum_{i=1}^{i_{\max}}\frac{\alpha_i^2}{\mu^2_{\mathbf{C_i}}}\mathrm{Var}\left\{\mathbf{C_i}|H_0\right\}}}\right) \tag{12}$$

$$p_{FN} = \frac{1}{2}\mathrm{erfc}\left(\frac{\sum_{i=1}^{i_{\max}}\alpha_i\left(\frac{1}{\mu_{\mathbf{C_i}}}\mathrm{E}\left\{\mathbf{C_i}|H_1\right\} - \frac{1}{2}\right)}{\sqrt{2\sum_{i=1}^{i_{\max}}\frac{\alpha_i^2}{\mu^2_{\mathbf{C_i}}}\mathrm{Var}\left\{\mathbf{C_i}|H_1\right\}}}\right) \tag{13}$$

where $\mathrm{erfc}\,(x) = \frac{2}{\sqrt{\pi}}\int\limits_x^\infty \exp(-\xi^2)\ \mathrm{d}\xi$.

The detector (11) is completely determined by the α_i and $\mu_{\mathbf{C_i}}$. These values can be defined independently from $\mathrm{E}\left\{\mathbf{C_i}|H_1\right\}$, $\mathrm{Var}\left\{\mathbf{C_i}|H_1\right\}$, and $\mathrm{Var}\left\{\mathbf{C_i}|H_0\right\}$, which, of course, does not give an optimal detector. However, in this case (13) and (12) can also be used to compute the error probabilities of mismatched detection[2].

3 The Quantization Channel

Quantization of a watermarked signal can decrease the robustness of watermark detection. We investigate scalar uniform quantization following the additive embedding of a watermark. The considered scheme is depicted in Fig. 2. Although every watermark can be described by an additive signal, the special property of the investigated watermark is its independence from the host signal $\underline{x}$. In this section, we assume that the samples of the watermark and the host signal are independent identically distributed and $\varDelta$ is the quantizer step size. Therefore, no signal decomposition is necessary.

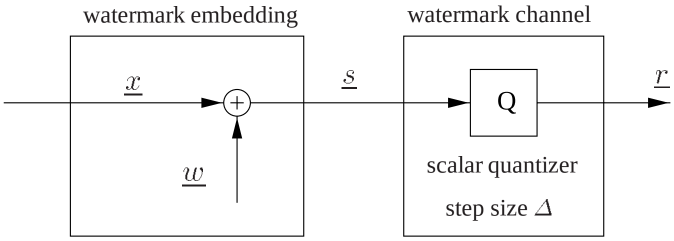

Fig. 2. Additive watermark embedding and subsequent quantization

It is known that the quantization channel can be modeled by an AWGN channel in the case of *fine quantization* [1]. However, for *coarse quantization*, as common for compression, better models are necessary. Here, only the most important results are summarized. A detailed description is given in [1].

The channel distortion $\underline{e}$ equals the quantization error $\underline{e} = \underline{r} - \underline{s}$. The watermark is detected via correlation according to (4). We introduce a slight extension to improve the detection with help of the original signal $\underline{x}$. This is expressed by the subtraction of $\underline{x}$, weighted by a factor γ_x. Thus we define

$$c[n] = (r[n] - \gamma_x[n]x[n])w[n]. \tag{14}$$

The normalized conditional expectation $\mathrm{E}\left\{\mathbf{c}|H_1\right\}$ is given by

$$\frac{\mathrm{E}\left\{\mathbf{c}|H_1\right\}}{\sigma^2_{\mathbf{w}}} = \frac{\mathrm{E}\left\{(\mathbf{r} - \gamma_x \mathbf{x})\mathbf{w}\right\}}{\sigma^2_{\mathbf{w}}} = \frac{\mathrm{E}\left\{\mathbf{e}\mathbf{w}\right\}}{\sigma^2_{\mathbf{w}}} + 1. \tag{15}$$

We would like to obtain the value 1, meaning the correlation $\mathrm{E}\left\{\mathbf{e}\mathbf{w}\right\}$ between the quantization error $\mathbf{e}$ and the watermark $\mathbf{w}$ should be zero. This is assumed when the quantization attack is modeled by an AWGN channel.

[2] Here, mismatched detection means to use a detector that was designed for the case of a different attack.

The value of $\mathrm{E}\{\mathbf{ew}\}/\sigma_{\mathbf{w}}^2$ can be determined for a given quantizer if the PDFs of the original signal $\mathbf{x}$ and the watermark $\mathbf{w}$ are known. Here, we investigate a Gaussian and Laplacian host signal $\mathbf{x}$ with zero mean and unit variance. We consider watermarks with Gaussian, uniform or bipolar ($w[n] = \pm\sigma_{\mathbf{w}}$) distributions. All three signal characteristics are frequently used in watermarking schemes. For convenience, the standard deviations of the input signal and the watermark are normalized by the quantizer step size $\varDelta$. This defines the normalized parameters $\chi = \sigma_{\mathbf{x}}/\varDelta$ and $\zeta = \sigma_{\mathbf{w}}/\varDelta$.

We are mainly interested in the robustness of an additive watermark against quantization of different coarseness. Fig. 3 shows the cross-correlation $\mathrm{E}\{\mathbf{ew}\}$ for a constant watermark-to-host-signal ratio ζ/χ and increasing quantizer step size $\varDelta$. In this case ζ/χ represents the embedding strength and an increasing step size $\varDelta$ equals a decreasing value of χ.

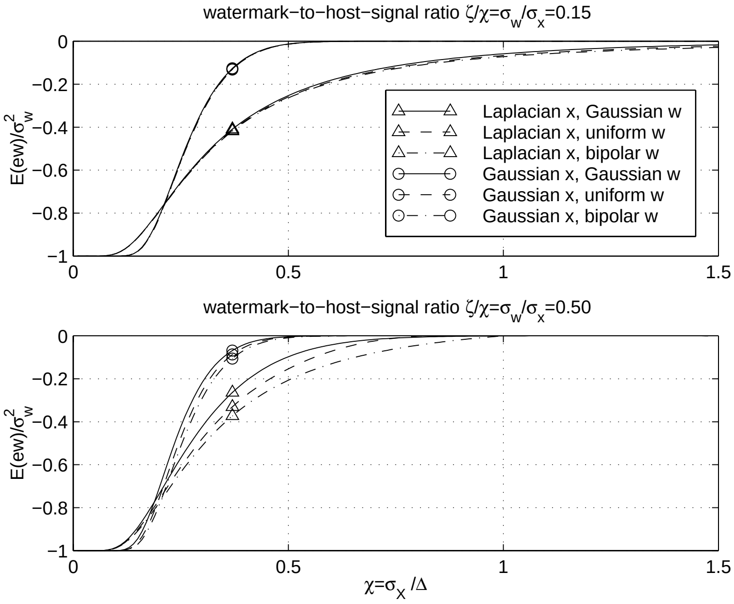

Fig. 3. Predicted cross-correlation $\mathrm{E}\{\mathbf{ew}\}/\sigma_{\mathbf{w}}^2$ for $\zeta/\chi = 0.15$ and $\zeta/\chi = 0.5$.

For a fixed ratio ζ/χ and varying χ, the correlation $\mathrm{E}\{\mathbf{ew}\}$ becomes zero for sufficiently large χ (fine quantization) and converges towards -1 for the limit $\chi \to 0$. The behavior of $\mathrm{E}\{\mathbf{ew}\}$ for large χ is intuitively clear, since in this case the host signal has an approximately constant PDF over the range of a step size $\varDelta$. At the limit $\chi \to 0$, the quantizer step size $\varDelta$ becomes arbitrarily large, which leads to a zero quantizer output, assuming zero is a reconstruction value, and thus to the quantization error $e[n] = -x[n] - w[n]$. As a result, the normalized expectation $\mathrm{E}\{\mathbf{ew}\}/\sigma_{\mathbf{w}}^2$ converges to -1 and the conditional expectation $\mathrm{E}\{\mathbf{c}|H_1\}$ becomes zero.

The characteristic of the watermark PDF does not have a significant influence for small ratios ζ/χ. More important is the influence of the host signal's distribution, especially, since this cannot be modified in watermarking schemes. The plots in Fig. 3 show that the watermark embedded in a signal with a Gaussian distribution resists quantization better than an equivalent watermark embedded in a signal with Laplacian distribution. In general we observe that with more peaky host signal PDFs – everything else being equal – the watermark is somewhat less robust against quantization attacks.

The expressions for the variances Var $\{\mathbf{c}|H_1\}$ and Var $\{\mathbf{c}|H_0\}$ are slightly more complicated and thus not derived here. The formulas given in [1] reveal that Var $\{\mathbf{c}|H_1\}$ and Var $\{\mathbf{c}|H_0\}$ are indeed different. However, they are approximately equal for common signal settings. The variances depend on the interference from the original document and, therefore, on the choice of γ_x. The weight $\gamma_x = 0$ has to be used when no knowledge about the original is available at the watermark decoder ("blind" detection). In applications, where full knowledge about the original can be exploited, the weight can be determined by the correlation of the received signal $\underline{r}$ with the original signal $\underline{x}$, yielding

$$\gamma_x = \frac{\mathrm{E}\{\mathbf{rx}\}}{\sigma_{\mathbf{x}}^2} = \frac{\mathrm{E}\{\mathbf{ex}\}}{\sigma_{\mathbf{x}}^2} + \frac{\mathrm{E}\{\mathbf{x}\}^2}{\sigma_{\mathbf{x}}^2} + 1. \tag{16}$$

Fig. 4 depicts the resulting detection error probalities after quantization of different strength. Here, the quality of the received signal $\underline{r}$, is measured by the host-signal-to-noise ratio after quantization. Again, we observe that the Laplacian host signal provides less robust watermark detection.

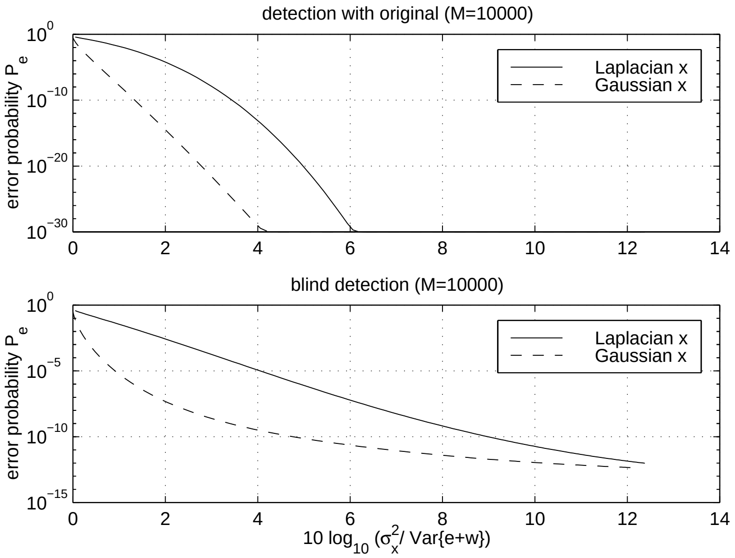

Fig. 4. Predicted error probabilities for $\zeta/\chi = 0.15$ and Gaussian watermark $\mathbf{w}$.

4 An Example Image Watermarking Scheme

The investigation of image watermark detection after quantization reveals the importance of the problems discussed in Section 2 and Section 3. The presented scheme is not fully optimized, but sufficiently good to give realistic results.

4.1 Host Data

The theoretical analysis of watermark detection after quantization has been made without specifying the data to be watermarked. Therefore, the results can be applied easily to many different signals. The following examples are for natural images. The watermark is embedded into the coefficients of an 8×8 block-DCT of the luminance component. Many different domains for the watermark embedding process have been proposed in recent publications, where, besides the DCT domain, wavelet domains are very popular [7, 8, 15]. We choose the block-wise DCT since this is also used by JPEG compression. We do not claim that the block-wise DCT is the optimal image decomposition for watermarking purposes. Two advantages of the proposed watermarking domain are:

- During JPEG compression, the coefficients of the 8×8 block-DCT are quantized with a uniform scalar quantizer, where the step size Δ_i can be different for each of the 64 frequencies. Therefore, defining the watermark in the DCT domain simplifies the optimization of detection after the compression attack.
- Quantizer step sizes for JPEG-baseline compression are optimized for subjectively quality and can be parameterized via a quality factor QF ($QF = 100$: highest quality with step size $\Delta = \Delta_{\min}$ for all coefficients; $QF = 1$: lowest quality with step size $\Delta = 256\Delta_{\min}$). Therefore, an invisible watermark can be achieved by adapting its strength to the quantization noise produced via JPEG compression with a sufficiently high quality factor.

4.2 Embedding Scheme

Fig. 5 depicts the scheme for the signal dependent additive watermark embedding. The signal decomposition is performed as in JPEG compression [14]. Image samples are denoted by $I(u, v, m)$, where (u, v) are the row and column indices of the m-th block (where the blocks are numbered in row-scan). All blocks are DCT transformed and the coefficients for the same frequency from all blocks are grouped into a sample sequence – a *sub-signal*. This sub-signal can be described relatively accurately by a white, stationary random variable $\tilde{\mathbf{x}}_i$. Since each sub-signal can be quantized differently, each sub-signal has its own *sub-channel*. Due to the 8×8 blocks, this scheme gives 64 vectors $\underline{\tilde{x}}_i$, where the index i denotes the sub-channel number. The sub-channels are numbered according to the common zigzag-scan of the DCT coefficients. The length of the vectors $\underline{\tilde{x}}_i$ equals the number of 8×8 blocks in the given image.

The main idea for the adaptation of the watermark strength is that the embedding should introduce roughly the same distortion as JPEG compression with a certain quality factor QF_e. Therefore, uniform scalar quantization with step size Δ_i (which is used

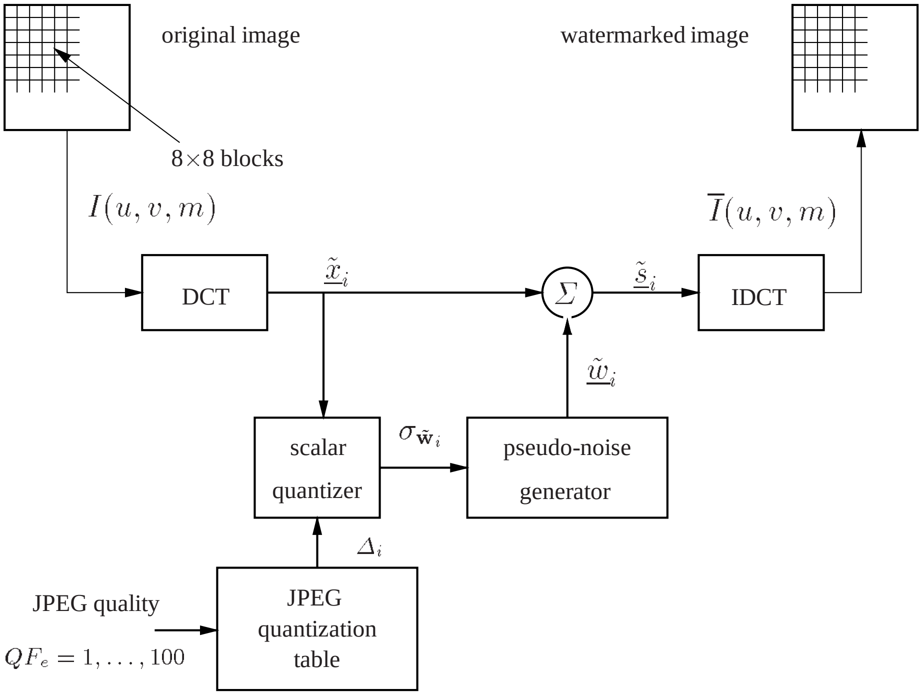

Fig. 5. JPEG-adapted additive watermarking

in JPEG compression with a certain quality QF_e) is applied to the elements of the vector $\underline{\tilde{x}}_i$. Now, the watermark variance $\sigma^2_{\mathbf{w_i}}$ for every sub-channel is chosen equal to the variance of the corresponding quantization errors. A Gaussian pseudo-noise vector $\underline{\tilde{w}}_i$ with the correspondent standard deviation is computed for each sub-channel and added to $\underline{\tilde{x}}_i$. Finally, the elements of the resulting 64 watermarked vectors $\underline{\tilde{s}}_i$ are inverse DCT transformed.

In [11, 12], it is suggested that the watermark frequency spectrum should be directly proportional to the host signal's. However, that work uses mean square error, and for subjective quality we should not satisfy this condition exactly.

4.3 Simulation Settings

In order to reduce the number of free parameters, we will discuss only the results for an embedding quality of $QF_e = 70$, which gives a watermarked image with sufficiently high quality. As a test image, we use the 256×256 gray-scale “Lenna” picture. The given image size leads to 1024 8×8 blocks, and thus to 1024 samples for each sub-channel $\underline{\tilde{x}}_i$.

200 differently watermarked images $\overline{I}(u,v,m)$ were produced, using the scheme depicted in Fig. 5, where the watermarks were obtained by different seeds for

the pseudo-random number generator. Note that, in contrast to an AWGN-channel, the quantization channel can only be investigated by varying input sequences. The watermarked images were JPEG compressed and decompressed, each with 20 different quality factors which are equally increased from $QF_a = 5$ to $QF_a = 100$.

For watermark detection, the attacked public document is transformed again by the 8×8 block-DCT. Then the signals $\underline{\tilde{r}}_i$ for the different sub-channels i are correlated with the corresponding watermarks $\underline{\tilde{w}}_i$. For a fair test, the detection process is carried out for both hypotheses H_1 and H_0, i.e., for documents that are or are not watermarked by $\underline{\tilde{w}}_i$. For simplicity we chose as reference an un-watermarked image, which was compressed in the same way the watermarked image was compressed.

5 Experimental Results

5.1 Detection Optimized for the Applied Attack

The detection rule (11) derived in Section 2 is determined by the weights $\alpha_i = \mu^2_{\mathbf{C_i}} / \sigma^2_{\mathbf{C_i}}$ and the expected correlations $\mu_{\mathbf{C_i}}$. These parameters are computed for each sub-channel and each of the 20 JPEG attacks, so that the detector defined by (11) can be optimized for a given attack. We found that the values derived experimentally by 200 different watermark realizations match to the values computed via the theoretic model presented in [1]. Therefore, it is possible to substitute the simulations by theoretic modeling.

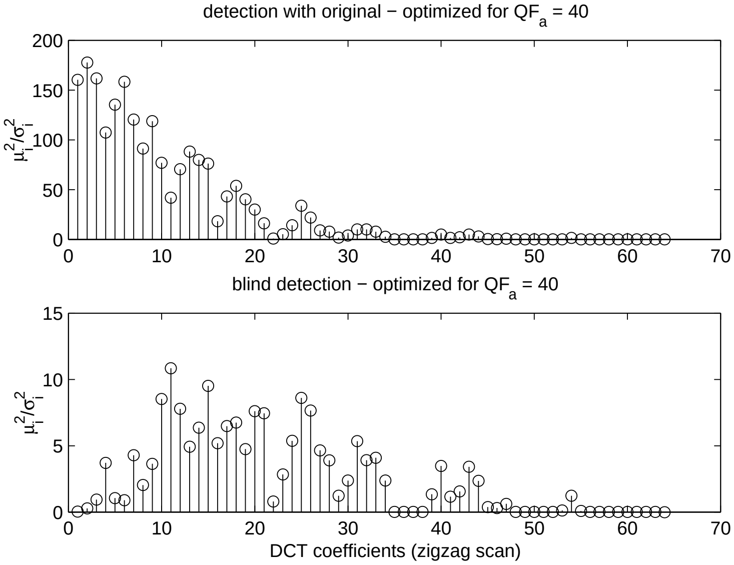

Fig. 6. Detection weights $\alpha_i = \mu^2_{\mathrm{C_i}} / \sigma^2_{\mathrm{C_i}}$ for all 64 DCT coefficients.

Channel weights. Fig. 6 depicts the weights determined after JPEG compression with $QF_a = 40$. The robustness of watermark detection is very different for the sub-channels (the DCT coefficients). This is due to the different quantization of each sub-channel, the adapted watermark strength and the differently distributed host signals. Further, it can be observed that the weights differ for detection with original and blind detection. The small weights for the low frequency coefficients in the case of blind detection are reasonable, since here the interference from the original is large. Enlarging the watermark power for these coefficients can increase the detection weights, but also degrades the subjective quality of the watermarked image. Detection with original is not affected by interference from the original document. Therefore, the low frequency coefficients provide the most robust detection after JPEG compression.

The results are similar for compression with other quality factors. However, especially in the case of blind detection the higher frequency coefficients get larger weights for high quality compression.

Detection error probabilities. With help of the measured or computed values α_i and $\mu_{\mathbf{C}_\mathrm{i}}$, the detection error probabilities can be investigated. When the strength of the conducted JPEG compression is known, the watermark detection can be optimized by plugging the appropriate weights into (11). The error probabilities for blind detection are shown in Fig. 7 and Fig. 8. Plots with linear and logarithmic axes are provided. The values found by actually counting detection errors are only depicted in the plots with linear axis due to the relatively small number of simulations. The results in Fig. 7 are achieved by considering only the watermark components embedded in the 12-th and 13-th DCT coefficient.

The plots show the results derived by estimating the necessary means and variances from the experimentally found correlation values. In addition these values are computed by modeling the host data by a Gaussian or generalized Gaussian random variable. The plots reveal that the Gaussian model – in contrast to the generalized Gaussian model – does not agree with the experimental results. This emphasizes the importance of the host signal's PDF. A Gaussian host signal would provide much better robustness, but the actual image data is *not* Gaussian distributed. In the upper plots of Fig. 7 and Fig. 8, the circles indicate the error rates found by actually counting the detection errors of the proposed detector. When using only a subset of all sub-channels, more errors occur and the error probabilities predicted by (12) and (13) can be verified with less simulations.

The plots also depict the error probabilities that can be expected for a detector that is designed after modeling the quantization attack by an AWGN channel. The AWGN model does not consider that coarse quantization removes parts of the embedded watermark, so the normalized expected correlation is always 1. The presented results demonstrate that this assumption leads to severe degradations of the detection performance. This is especially true when all sub-channels are considered since in this case un-robust sub-channels are weighted too strongly. Here, it would be better to detect the watermark only in a small subset of all sub-channels.

With help of the proposed detector – optimized for the given attack – even blind detection of a watermark embedded with quality $QF_e = 70$ is still possible after com-

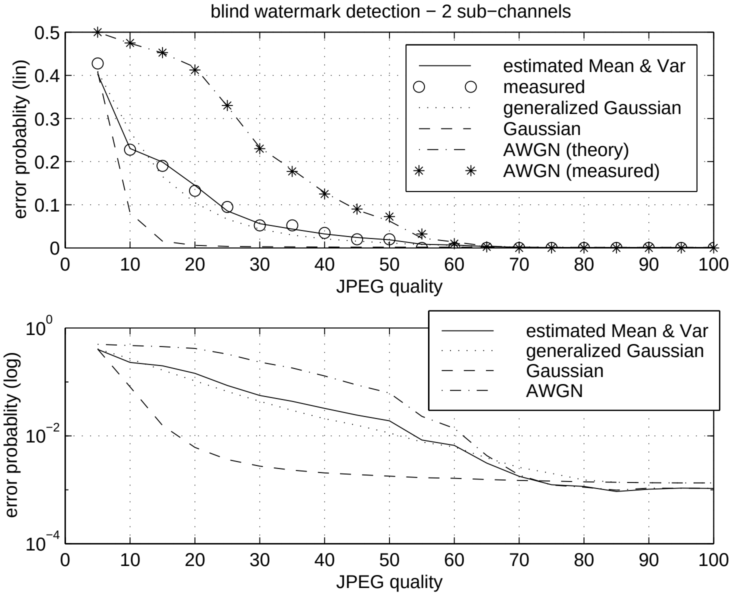

Fig. 7. Error probabilities for blind watermark detection after JPEG compression considering sub-channel 12 and 13. The detector is always optimized for the special attack.

pression attacks down to qualities about $QF_a = 20$. Naturally, detection with original is much more robust, so we do not present these results here.

5.2 Detection Optimized for a Worst Case Attack

The detection error probabilities presented in the previous subsection are very promising. However, full knowledge about the compression attack might not always be available. If this is the case, we have to find a detector that works for a large set of possible attacks. The error probabilities using a detector optimized for an attack with $QF_a = 40$ are shown in Fig. 9. These plots allow also the comparison of the performance of the parabolic and planar detection boundaries as discussed in Section 2.3. Due to the optimization of the detector for attacks with $QF_a = 40$, the parabolic detector fails completely for weak attacks ($QF_a \geq 80$). The planar detector is much more robust. We cannot expect to get error probabilities as low as with a fully optimized detector. However, the error rate is still lower than 10^{-10} for all attacks with $QF_a \geq 40$. Again, the AWGN model is in general not appropriate. It fits only in the case of very weak attacks.

6 Conclusion

Quantization attacks against embedded watermarks have different severity, depending on the PDF of the host signal. The AWGN model is not appropriate for quantization

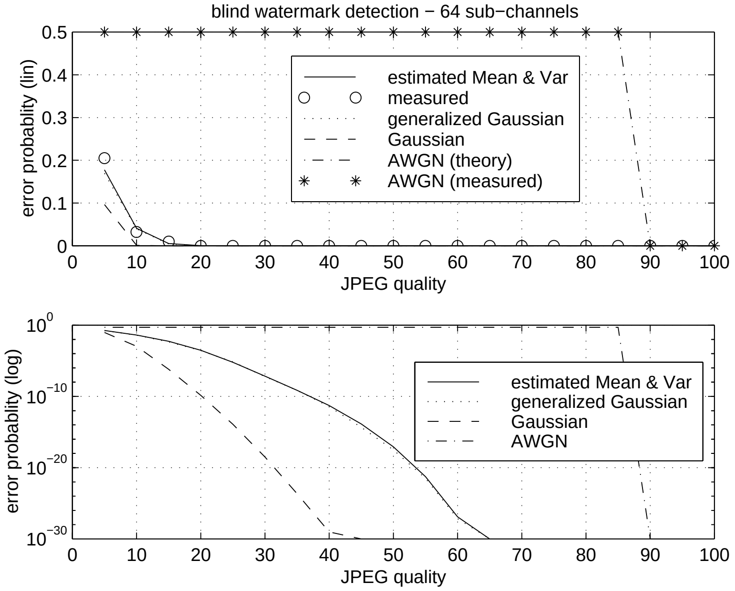

Fig. 8. Error probabilities for blind watermark detection after JPEG compression considering all sub-channels. The detector is always optimized for the special attack.

channels. This has to be considered when the watermark detector is optimized to resist certain compression attacks. An example image watermarking scheme is presented, and the proposed detection principle experimentally verified. With this scheme, it is possible to detect watermarks without using the original document with low error probabilities even after JPEG-compression with $QF_a = 20$. In practical schemes it is sufficient to use a detector that is optimized for a worst case attack, e.g., a strong attack that still provides an attacked document with good subjective quality.

References

1. J. J. Eggers and B. Girod. Quantization Effects on Digital Watermarks. *Signal Processing*, 1999. Submitted.
2. I. A. Glover and P. M. Grant. *Digital Communications*. Prentice Hall, London, New York, Toronto, Sydney, 1998.
3. R. M. Gray and T. G. Stockham. Dithered quantizers. *IEEE Transactions on Information Theory*, 39(3):805–812, May 1993.
4. J. C. Hancock and P. A. Wintz. *Signal Detection Theory*. McGraw-Hill, Inc, New York, St. Louis, San Francisco, Toronto, London, Sydney, 1966.
5. F. Hartung, J. K. Su, and B. Girod. Spread spectrum watermarking: Malicious attacks and counter-attacks. In *Proceedings of SPIE Vol. 3657: Security and Watermarking of Multimedia Contents*, San Jose, Ca, USA, January 1999.
6. N. S. Jayant and P. Noll. *Digital Coding of Waveforms*. Prentice Hall, 1984.

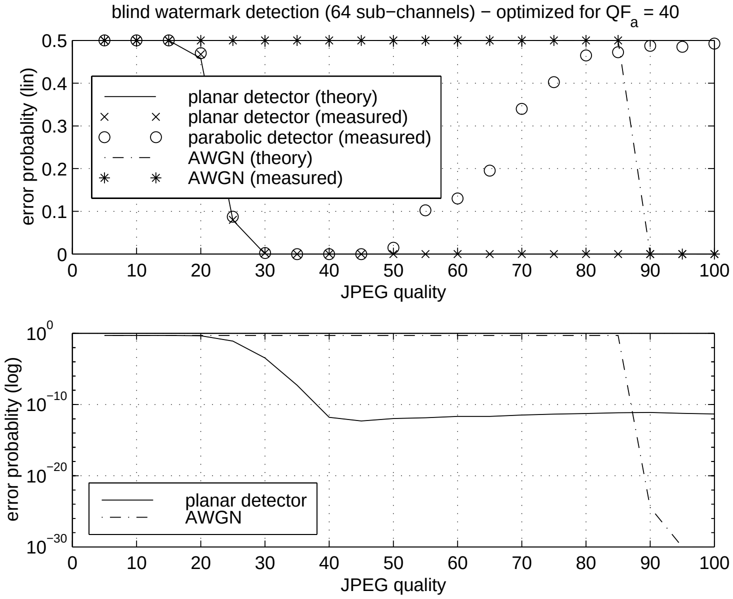

Fig. 9. Error probability of blind watermark detection after JPEG compression considering all sub-channels. The detector is optimized for an attack with $QF_a = 40$.

7. D. Kundur and D. Hatzinakos. Digital watermarking using multiresolution wavelet decomposition. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing 1998 (ICASSP 98), Seattle, WA, USA*, volume 5, pages 2969–2972, May 1998.
8. M. Kutter and F. A. P. Petitcolas. A fair benchmark for image watermarking systems. In *Proceedings of SPIE Vol. 3657: Security and Watermarking of Multimedia Contents*, San Jose, Ca, USA, January 1999.
9. A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, New York, 3rd edition, 1991.
10. L. Schuchman. Dither signals and their effect on quantization noise. *IEEE Transaction on Communication Technology (COM)*, 12:162–165, December 1964.
11. J. K. Su and B. Girod. On the imperceptibiliy and robustness of digital fingerprints. In *Proceedings IEEE International Conference on Multimedia Computing and Systems (ICMCS 99)*, Florence, Italy, June 1999.
12. J. K. Su and B. Girod. Power-spectrum condition for L_2-efficient watermarking. In *Proceedings of the IEEE International Conference on Image Processing 1999 (ICIP 99)*, Kobe, Japan, October 1999. Accepted.
13. J. K. Su, F. Hartung, and B. Girod. A channel model for a watermark attack. In *Proceedings of SPIE Vol. 3657: Security and Watermarking of Multimedia Contents*, San Jose, Ca, USA, January 1999.
14. G. K. Wallace. The JPEG still picture compression standard. *Communications of the ACM*, 34(4):31–44, April 1991.
15. L. Xie and G. R. Arce. A Blind Wavelet Based Digital Signature for Image Authentication. In *Proceedings European Signal Processing Conference (EUSIPCO 98)*, Greece, September 1998.