ICIP 2000
Special Session on WM
Vancouver, Canada
Sep. 10–13, 2000

ROBUSTNESS OF A BLIND IMAGE WATERMARKING SCHEME

Joachim J. Eggers, Jonathan K. Su

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

Bernd Girod

Information Systems Laboratory
Stanford University
Stanford, CA 94305-9510, USA
girod@ee.stanford.edu

ABSTRACT

Recently it was shown that, in some situations, blind watermarking can perform as well as watermarking schemes with the host signal available to the decoder. In this paper, blind watermarking of colored Gaussian host signals in the presence of filtering and additive Gaussian noise attacks is discussed. Three suboptimal but practical schemes are compared with a scheme where the host signal is available at the decoder. The performance is analyzed theoretically and experimentally for image watermarking.

1. INTRODUCTION

The goal in digital watermarking is to communicate information by embedding it in digital data, called "host data," to produce "marked data," from which the information should be retrieved. The embedded *watermark* conveys this information, which should be reliably decodable even after *attacks*, processing of the marked data that may inadvertantly or deliberately impair communication. *Robustness* refers to the ability of a watermarking scheme to maintain communication in the presence of attacks. In addition, it is often desired to retrieve the embedded information without reference to the host data; this is known as *blind watermarking*. This paper combines two recent developments, one in blind watermarking, another in the study of robustness.

2. BLIND WATERMARKING

It has been shown recently that blind watermarking can be considered communication with side information (the host signal x) at the encoder [1]. This insight leads to a new group of blind watermarking schemes; some of them are discussed here. Throughout this section we consider sending a watermark letter $d \in \mathcal{D}$, where $\mathcal{D}$ is a finite discrete alphabet, over an AWGN channel with noise $z \sim \mathcal{N}(0, \sigma_{\mathbf{z}}^2)$.

2.1. Communication with Side Information at the Encoder

For a host signal $x \sim \mathcal{N}(0, \sigma_{\mathbf{x}}^2)$, available at the encoder but not the decoder, Costa [2] has shown that the capacity is $C = 0.5 \log(1 + \sigma_{\mathbf{w}}^2/\sigma_{\mathbf{z}}^2)$, independent of $\sigma_{\mathbf{x}}^2$. However, a codebook $\mathcal{U}$ must be designed that accounts for the side information x. Costa presented a capacity-achieving scheme based on a random codebook $\mathcal{U}$, where the code sequences are $u = w + \alpha x$ with independent Gaussian signals w and x, and a scalar factor α. The codebook must be designed for a certain channel noise power $\sigma_{\mathbf{z}}^2$ and a given watermark power $\sigma_{\mathbf{w}}^2$, where capacity can be achieved for $\alpha^* = \sigma_{\mathbf{w}}^2/(\sigma_{\mathbf{w}}^2 + \sigma_{\mathbf{z}}^2)$. Note that α^* is independent of the host power $\sigma_{\mathbf{x}}^2$. The decoder quantizes the received signal $y = x+w+z$ to the closest entry in the codebook $\mathcal{U}$ and produces the index i of the quantized signal. The index i is mapped to the decoded letter $\hat{d}$, e.g., $\hat{d} = i \bmod |\mathcal{D}|$ for a regular mapping, where $|\mathcal{D}|$ is the size of the alphabet. The encoder perturbs the host x by w to form the sent signal $s = x + w$ so that, with high probability, y will fall into the correctly indexed quantization bin.

For low channel noise ($\alpha^* \approx 1$), the codebook must describe the host signal almost perfectly. In this case the codebook contains different vector quantizers for the host signal dependent on the watermark letter d to be sent. For large noise power ($\alpha^* \approx 0$) the codebook contains almost no information about the host signal x. Here, additive embedding of a pseudo-noise sequence is nearly optimal.

2.2. Practical Blind Watermarking

Particularly for low channel noise, the random codebook in Costa's scheme becomes large so that neither storing it nor searching it is practical. Thus, we proposed [3] replacing it by a structured codebook, e.g., a product codebook of dithered uniform scalar quantizers, and sending one watermark letter $d_n \in \mathcal{D} = \{0, 1\}$ per host sample x_n. We denote this scheme by SCS (scalar Costa scheme); the same approach was considered by Chen and Wornell [1]. Fig. 1 depict the corresponding embedding process which embeds the nth watermark letter d_n in the nth signal sample x_n. Ramkumar

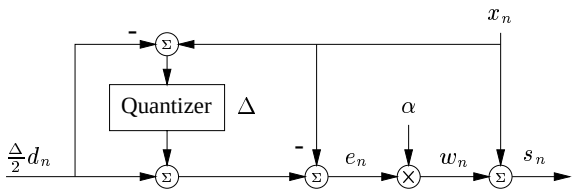

Fig. 1. Watermark embedding using Costa's scheme with a scalar component codebook (SCS). The watermark symbol $d_n \in \mathcal{D} = \{0, 1\}$ is embedded after dithered uniform scalar quantization of x_n and the embedding of the scaled quantization error αe_n as watermark w_n.

[4] describes a similar scheme based on continuous periodic functions for self noise suppression (CP-SNS), where each embedded watermark sample is thresholded to $|w_n| \leq \beta/2$. A special case of both schemes is dither modulation (DM), proposed earlier by Chen and Wornell [5]. Fig. 2 depicts the PDFs of the sent signal s, for all three methods.

For the AWGN channel, the PDFs $p_{\mathbf{y}}(y)$ and $p_{\mathbf{y}}(y|d)$ of the

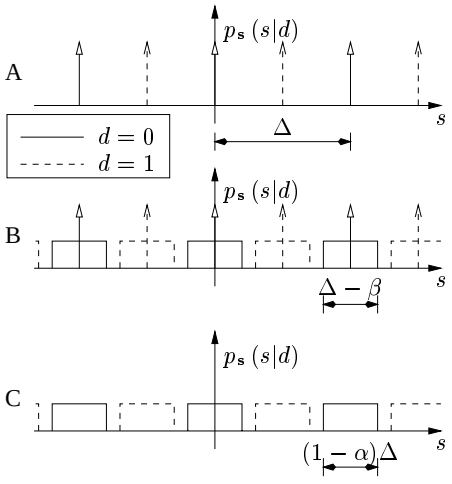

Fig. 2. PDFs of the sent value s for a given watermark symbol $d \in \mathcal{D} = \{0, 1\}$ in the case of (A) DM proposed in [5], (B) CP-SNS with thresholding proposed in [4], and (C) SCS proposed in [1, 3].

received signal y are obtained by convolving the conditional PDF $p_{\mathbf{s}}(s|d)$ of the sent signal s for a given watermark symbol d and the PDF $p_{\mathbf{z}}(z)$ of the additive channel noise: $p_{\mathbf{y}}(y|d) = p_{\mathbf{s}}(y|d) * p_{\mathbf{z}}(y)$. For small quantizer step sizes Δ, we can assume that $p_{\mathbf{s}}(s|d)$ is periodic with period Δ; then the convolution can be computed precisely by the discrete Fourier transform of the product of the the characteristic functions for $p_{\mathbf{s}}(s|d)$ and $p_{\mathbf{z}}(z)$ [3].

Knowing the PDFs of the received signal, we compute the mutual information $I(y; d)$ and the uncoded bit error rate p_e. Fig. 3 shows the mutual information obtained for all three considered methods with binary signaling. DM performs poorly for negative watermark-to-noise ratios WNR= $10 \log_{10} \sigma_{\mathbf{w}}^2/\sigma_{\mathbf{z}}^2$. SCS and CP-SNS are much more robust since α and β can be optimized to achieve better noise resistance. Optimal values for α and β are given in [3, 4]. SCS performs slightly better than CP-SNS. These results are independent of the host power $\sigma_{\mathbf{x}}^2$ for reasonably large $\sigma_{\mathbf{x}}^2$, thus, no general comparison with conventional blind schemes like spread spectrum watermarking is possible.

Fig. 4 depicts the probability of bit errors for uncoded transmission. An additive bipolar random watermark sequence is used for the scheme with host signal at the decoder. SCS and CP-SNS perform comparably, and are significantly better than DM.

In practice, SCS will be used in combination with error correction codes. To verify our results on the achievable rate, binary SCS was combined with turbo codes [6], which allow near-capacity performance. The results for code rates R=1/2, R=1/3, and R=1/5 are shown in Fig. 5. We observe that the typical error floor of the turbo codes is achieved within about 0.7 dB of the achievable rate. These results agree closely with the performance of turbo codes in the case of 2PSK transmission over an AWGN channel. Note that such good performance can be obtained only for large blocks of host signal samples.

The mutual-information results and detection-error rates reveal that a significant performance gap remains between the suboptimal scalar codebook and schemes with the host signal at the decoder. Chou et al. [7] have shown that duality between blind watermarking and distributed source coding exists, which can be exploited to design better codebooks.

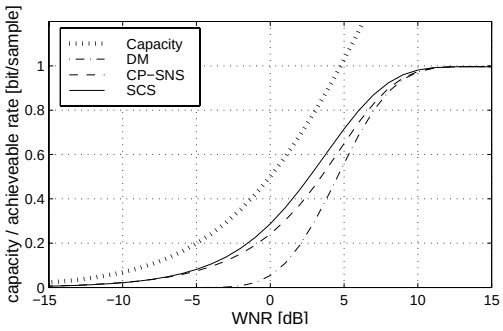

Fig. 3. Comparison of mutual information for different blind watermarking schemes.

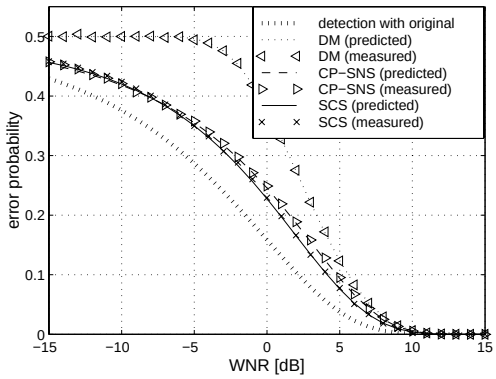

Fig. 4. Measured and predicted bit error rate for uncoded binary transmission.

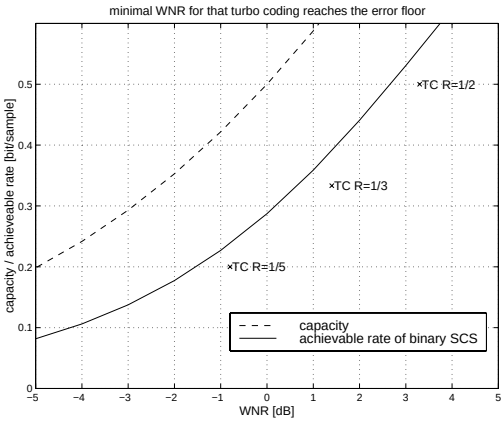

Fig. 5. SCS with turbo coding. Standard turbo codes with interleaver length 10,000 are combined with binary SCS. The minimum WNR necessary to reach the error floor of the turbo codes is depicted for different rates. In all cases, the error floor is at $p_e \leq 10^{-5}$.

3. THEORETICAL ROBUSTNESS ANALYSIS

The robustness of digital watermarks against linear filtering and additive noise attacks was studied theoretically in [8]. All signals are treated as (colored) Gaussian random processes. With

the power spectra of the host signal and watermark given, the attack finds the best combination of LSI filtering and additive colored Gaussian noise to minimize the channel capacity for a desired attacked-signal distortion. It was shown in [9, 8] that LSI filtering and noise yields a more effective attack than additive noise alone. Equations for the filter transfer function and noise power spectrum appear in [8].

The investigation produced a rule of thumb for resisting the attack when using mean-squared error (MSE) distortion. Namely, "white watermarks have near-optimal robustness at low distortions, while *power-spectrum condition-compliant* (PSC-compliant) watermarks have near-optimal robustness at high distortions." A PSC-compliant watermark has a power spectrum that is directly proportional to that of the host.

The power spectrum of a signal can often be approximated as a collection of parallel, independent, memoryless Gaussian channels. This approximation is useful for practical application of the analysis in [8]. For example, a discrete-time/space signal can be transformed into a frequency representation, where each discrete frequency represents one channel. Frequencies could also be grouped to form a channel. Within this framework it is also possible to consider different signal statistics in different image objects.

Within each channel, the watermarking and attack model has the form shown in Fig. 6. Here, $d_{n,k}$ is the component of the message d in the kth channel at the nth use of the channel; $d_{n,k}$ is embedded into the host-signal component $x_{n,k}$ to produce $s_{n,k}$, the kth watermarked-signal component at channel usage n. The embedding distortion between s and x is measured over all channels and channel usages.

Next, for each channel, the attack behaves like the "Gaussian test channel" [9]. The attack multiplies $s_{n,k}$ by $g_{n,k}$ and adds noise $v_{n,k}$ to yield $r_{n,k}$, the kth component of the attacked-signal during the nth use of the channel. Because the channel is memoryless, $g_{n,k} = g_k$, $\forall n$. The distortion of the attacked signal is measured between r and the host signal x over all channels and channel usages.

Finally, the receiver is given r. If $g_n \neq 0$, the receiver compensates for filtering by dividing $r_{n,k}$ by g_k to produce $y_{n,k} = s_{n,k} + v_{n,k}/g_k$. For decoding, we can define an *effective AWGN channel* with noise $z_{n,k} = v_{n,k}/g_k$.

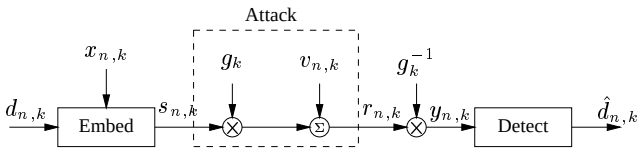

Fig. 6. Transmission over a Gaussian test channel.

4. EXPERIMENTS

In Sec. 2 and Sec. 3, theoretical results were discussed and verified for synthetic data. Here, watermarking of gray-scale images is considered. We used an 8×8 block DCT to decompose the image into 64 sub-channels; each frequency is considered a sub-channel. In practice the high-frequency coefficients may not be used for secure watermarking, thus, only the first 21 coefficients in zigzag scan were used for watermarking. The schemes described in Sec. 2 were implemented for each sub-channel, where the parameters are chosen dependent on the watermark power and effective noise power per sub-channel. The attack described in [8] was applied. Following the arguments given in Sec. 3, white watermarks and PSC watermarks were used.

First of all, we discuss experiments without coding, except for repetition. A repetition code over all 21 sub-channels was used to embed one bit per block. For attacks introducing high distortions, many errors occur, since embedding of one bit per block means operating above capacity. Nevertheless, the relative performance of the different schemes can be observed.

For the scheme with the host signal available at the decoder (WH), the combined detection from all sub-channels described in [10] was used. We also implemented a conventional scheme without host interference suppression (BL). Here, the same detection principles as for detecting with host signal were used except that the host signal was not subtracted. Note that the DCT coefficients are non-Gaussian and, thus, the interfering noise is also non-Gaussian in this case. For SCS, a maximum-likelihood detector based on the numerically computed PDFs $p_{\mathbf{y_k}}(y_k|d_k)$ was used to detect one bit jointly from all sub-channels.

Fig. 7 depicts the probability of error curves achieved for the 256×256 Lenna image. The experiments were conducted for 20 different realizations of the attack. Thus, having 1024 bits per image, the lowest measurable bit error probability is 4.9×10^{-5}. The probabilities of error for the conventional blind scheme were not predicted accurately due to the non-Gaussian noise. For very strong attacks, the PSC watermark gives fewer detection errors, while the white watermark performs better for moderate or weak attacks. It is obvious that the SCS watermarking does not reach the performance of detection with host. However, in the case of weak attacks, a significant improvement over conventional blind watermarking scheme can be obtained. For strong attacks, SCS hardly performs better than the conventional approach since α becomes very small.

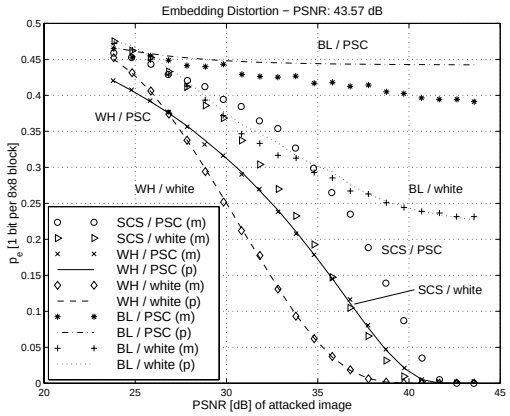

Fig. 7. Uncoded bit error rates. Measured (m) and predicted (p) results are depicted.

In practice, low rate error correction codes should be used to achieve low error probabilities even in the case of strong attacks. The choice of a specific error correction code, or the combination of several codes, is dependent on many parameters, e.g., the block size, the allowable complexity, and the statistics of the considered attack. A detailed discussion and optimization of these parameters is beyond the scope of this paper. However, we consider two simple techniques achieving good performance for stronger attacks

and discuss experimental results obtained for SCS watermarks embedded into the 256×256 Lenna image.

First of all, we use the spread transform (ST) technique, proposed in [11]. One watermark symbol is embedded into the projection of four DCT values of a certain frequency onto a random sequence. This technique gives a WNR-gain of about 6 dB for each specific frequency. Due to the joint detection from 21 differently robust sub-channels, the overall gain is in general not that high, but still several dB. Using ST of length 4, 256 bits were embedded into the test image.

Second, in addition to the ST of length 4 and the repetition code over 21 sub-channels, we used a hard-decision BCH code with 255 coded bits per 99 information bits. For this setting, 99 bits can be embedded into the given test image.

Fig. 8 depicts the measured error probabilities for the optimized filter and additive noise attack described in Sec. 3, where the results are averaged over 200 different realizations of the attack. Thus, error probabilites down to 5×10^{-5} can be measured. The embedding quality was chosen as in the previous uncoded experiments, thus, the watermarked image has a PSNR (peak signal-to-noise-ratio) of about 44 dB. For the experiment with BCH codes, no errors were found when the attacked image has a PSNR $>$ 34 dB. Thus, the 99 embedded watermark bits are robust against a loss of quality of about 10 dB.

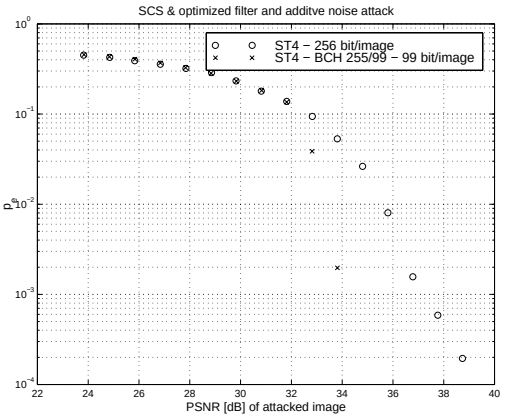

Fig. 8. Error probability for SCS watermarks embedded into the test image "Lenna". The filtering and additive noise attack is optimized for each target quality (PSNR) of the attacked image.

All experimental results were obtained for the attack discussed in Sec. 3. One topic for future research is the investigation of other types of attacks, e.g., quantization and nonlinear filtering. First experiments with JPEG compression, which is mainly a type of quantization attack, indicate that PSNRs of the compressed image down to 32 dB can be accepted for a rate of 99 watermark bits per image. However, the quantizer step sizes used in JPEG compression are optimized to achieve good compression for a given image quality. For a detailed analysis of robustness against quantization attacks, the quantizer step sizes used for different DCT coefficients have to be optimized to impair the embedded watermark as much as possible.

5. CONCLUSIONS

We have tested three blind watermarking schemes proposed in [5], [4], and [1, 3]. The latter [1, 3] was motivated by [2] and can perform significantly better than the first mentioned scheme, and slightly better than the second one. In image watermarking experiments, we have also verified the rule of thumb for watermark robustness in Sec. 3: white watermarks perform better when the attacked-signal distortion is low, PSC-compliant ones perform better when this distortion is high. The experiments also confirmed that, depending on the attacked-signal distortion, the Costa-motivated blind watermarking scheme can perform significantly better than conventional blind schemes. Future work will focus on developing more powerful blind schemes and the investigation of nonlinear attacks.

6. ACKNOWLEDGEMENT

We thank Marco Breiling for providing his software for turbo coding and his assistance.

7. REFERENCES

[1] B. Chen and G. Wornell, "Preprocessed and postprocessed quantization index modulation methods for digital watermarking," in *Proc. of SPIE Vol. 3971: Security and Watermarking of Multimedia Contents II*, San Jose, Ca, USA, January 2000, pp. 48–59.

[2] M. H. M. Costa, "Writing on Dirty Paper," *IEEE Transactions on Information Theory*, vol. 29, no. 3, pp. 439–441, May 1983.

[3] J. J. Eggers, J. K. Su, and B. Girod, "A blind watermarking scheme based on structured codebooks," in *Secure Images and Image Authentication, IEE Colloquium*, London, UK, April 2000, pp. 4/1–4/6.

[4] M. Ramkumar, *Data Hiding in Multimedia: Theory and Applications*, Ph.D. thesis, New Jersey Institute of Technology, Kearny, NJ, USA, November 1999.

[5] B. Chen and G. W. Wornell, "Digital watermarking and information embedding using dither modulation," in *Proc. of IEEE Workshop on Multimedia Signal Processing (MMSP-98)*, Redondo Beach, CA, USA, Dec. 1998, pp. 273–278.

[6] C. Berrou and A. Glavieux, "Near Optimum Error Correcting Coding and Decoding," *IEEE Transactions on Communications*, vol. 44, no. 10, pp. 284–287, October 1996.

[7] J. Chou, S. Pradhan, L. El Ghaoui, and Kannan Ramchandran, "A robust optimization solution to the data hiding problem using distributed source coding principles," in *Proc. of SPIE Vol. 3974: Image and Video Communications and Processing 2000*, San Jose, Ca, USA, January 2000.

[8] J. K. Su, J. J. Eggers, and Bernd Girod, "Analysis of digital watermarks subjected to optimum linear filtering and additive noise," Submitted to Signal Processing, Special Issue on Information-Theoretic Issues in Digital Watermarking., Apr. 2000.

[9] P. Moulin and J. A. O'Sullivan, "Information-Theoretic Analysis of Information Hiding," Preprint, September 1999.

[10] J. J. Eggers and B. Girod, "Quantization watermarking," in *Proc. of SPIE Vol. 3971: Security and Watermarking of Multimedia Contents II*, San Jose, Ca, USA, January 2000.

[11] B. Chen and G. W. Wornell, "Achievable performance of digital watermarking systems," in *Proceedings of the IEEE Intl. Conference on Multimedia Computing and Systems*, Florence, Italy, June 1999, vol. 1, pp. 13–18.