ICASSP 2001
Salt Lake City
Utah, USA
May 7–11, 2001

BLIND WATERMARKING APPLIED TO IMAGE AUTHENTICATION

Joachim J. Eggers

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

Bernd Girod

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

ABSTRACT

To prevent image manipulations and fraudulent use of modified images, the embedding of semi-fragile digital watermarks into image data has been proposed. The watermark should survive modifications introduced by random noise or compression, but should not be detectable from non-authentic regions of the image. The original image cannot be used by the watermark detector to verify the authenticity of the image. In this paper, we investigate the application of a recently developed quantization based watermarking scheme to image authentication. The watermarking technology, called *Scalar Costa Scheme* (SCS), allows reliable blind watermark detection from a small number of pixels, and thus enables the detection of local modifications to the image content.

1. INTRODUCTION

This paper focuses on image authentication with respect to the originality of image content. That is, modification of the content, like exchanging a head of a person with someone else's or erasing an image object, should be detectable. It is assumed that modification of the content requires the replacement of the pixels in an entire region of the image. The image is considered authentic if no such modifications are detectable.

One approach to tackle the authentication problem is based on watermarking. Watermark information generated with a *key* $\mathbf{k}$ is spread all over the host image. The marked image is subject to operations like D/A and A/D conversion, lossy compression or additive noise, leading to a distorted image. The watermark is designed such that it is reliably detectable as long as the distorted image has a sufficiently high quality and the correct key is known. However, if some image region is replaced by somebody not knowing the key $\mathbf{k}$, the watermark information will not be reliably detectable from the modified image region. Therefore, reliability of watermark detection can be used as a measure of authenticity. To be able to locate non-authentic image parts, the watermark should be reliably detectable from a small neighborhood of authentic pixels. Watermarking schemes that require many signal samples for reliable detection will not be appropriate.

Watermarks designed for authentication methods are called "semi-fragile" watermarks [1]. This term reflects that the watermark is robust against one group of attacks and fragile against other attacks. Here, we demand robustness against attacks facing a constraint on the introduced distortion, and fragility against local replacement of data.

We propose a two-step design procedure for semi-fragile watermarks. First, a watermark **communication** scheme is selected which offers high robustness against distortion-constrained attacks, where *mean squared error* (MSE) is used as distortion measurement throughout this paper. Using this watermarking scheme, a specific codeword $\mathbf{d}$ is embedded into the host data $\mathbf{x}$ dependent on a key $\mathbf{k}$. Without loss of generality, we choose the all-zero codeword $\mathbf{d}^0 = \mathbf{0}$.

Second, we consider **authentication** of a sub-set $\mathbf{r}_{\mathcal{R}}$ of the received, possibly attacked data $\mathbf{r}$, where $\mathcal{R}$ of size $R = |\mathcal{R}|$ is the index set of the considered elements of $\mathbf{r}$. Watermark detection from $\mathbf{r}_{\mathcal{R}}$ is formulated as an hypothesis test.

In this paper, the "Scalar Costa Scheme (SCS)" [2, 3] is applied for watermarking. This method was previously developed for efficient blind watermarking in the context of copyright protection or fingerprinting. In Sec. 2, we briefly review the design of SCS watermarking. The application of SCS watermarking for authentication purposes is discussed in Sec. 3. Experimental results for image data are presented in Sec. 4, where also comparisons with authentication using a *spread-spectrum* (SS) type watermarking scheme are shown.

2. SCS WATERMARKING

A general model for the communication of a message via watermarking can be described as follows: The encoder derives from the watermark message and the host data $\mathbf{x}$ an appropriate watermark sequence $\mathbf{w}$ which is added to the host data to produce the watermarked data $\mathbf{s}$. $\mathbf{w}$ must be chosen such that the distortion between $\mathbf{x}$ and $\mathbf{s}$ is negligible. Next, an attacker might modify the watermarked data $\mathbf{s}$ into data $\mathbf{r}$ to impair watermark communication. The attack is only constrained with respect to the distortion between $\mathbf{x}$ and $\mathbf{r}$. Finally, the decoder must be able to detect the watermark message from the received data $\mathbf{r}$. In *blind* watermarking schemes, the host data $\mathbf{x}$ are not available to the decoder. The codebook used by the watermark encoder and decoder is randomized dependent on a key $\mathbf{k}$ to achieve secrecy of watermark communication. In this paper, $\mathbf{x}$,$\mathbf{w}$,$\mathbf{s}$,$\mathbf{r}$ and $\mathbf{k}$ are vectors, and x_n,w_n,s_n,r_n and k_n refer to their respective nth elements.

It has been shown that blind watermarking can be considered communication with side information at the encoder [4]. Moulin and O'Sullivan [5] showed that for white Gaussian host data and MSE distortion measurement, the Gaussian test channel (GTC) is the worst possible attack in the sense that the rate of reliable communication is minimized for a constrained distortion of $\mathbf{r}$. The design of a watermark encoder and decoder in case of a GTC attack can be translated into the design for an effective additive white Gaussian noise (AWGN) attack [3]. For the latter case, Costa

[6] showed theoretically that for a Gaussian host signal of power $\sigma_{\mathbf{x}}^2$, a watermark signal of power $\sigma_{\mathbf{w}}^2$, and AWGN of power $\sigma_{\mathbf{v}}^2$ the maximum rate of reliable communication (capacity) is $C = 0.5\log(1+\sigma_{\mathbf{w}}^2/\sigma_{\mathbf{v}}^2)$, independent of $\sigma_{\mathbf{x}}^2$. The result is surprising since it shows that the host signal $\mathbf{x}$ need not be considered as interference at the decoder although the decoder does not know $\mathbf{x}$.

Costa's scheme involves a **random** codebook which must be available at the encoder and the decoder. Unfortunately, for good performance the codebook must be so large that neither storing it nor searching it is practical. Thus, we proposed replacing it by a structured codebook, in particular a product codebook of dithered uniform scalar quantizers and called this scheme *SCS* (Scalar Costa Scheme) [2]. The watermark message m is encoded into a sequence of watermark letters $\mathbf{d}$, where $d_n \in \mathcal{D} = \{0,1\}$ in case of binary SCS. Each of the watermark letters is embedded into the corresponding host elements x_n. The embedding rule for the nth element is given by

$$\begin{aligned} a_n &= \Delta\left(\frac{d_n}{2}+k_n\right)\\ s_n &= x_n+\alpha\left(\mathcal{Q}_\Delta\left\{x_n-a_n\right\}+a_n-x_n\right), \end{aligned} \tag{1}$$

where $\mathcal{Q}_\Delta\left\{\cdot\right\}$ denotes scalar uniform quantization with step size Δ. The key $\mathbf{k}$ is a pseudo-random sequence with $k_n \in (0,1]$. This embedding scheme depends on two parameters: the quantizer step size Δ and the scale factor α. Both parameters can be jointly optimized to achieve a good trade-off between embedding distortion and detection reliability for a given noise variance of an AWGN attack. Optimal values for Δ and α are given in [2]. In case of the GTC attack with a certain constraint on the attack distortion, the parameters α and Δ are obtained from those for the equivalent effective AWGN attack.

Watermark detection is based on the pre-processed received data $\mathbf{y}$. The extraction rule for the nth element is

$$y_n = \mathcal{Q}_\Delta\left\{r_n-k_n\Delta\right\}+k_n\Delta-r_n, \tag{2}$$

where $|y_n| \le \Delta/2$. y_n should be close to zero if $d_n = 0$ was sent, and close to $\pm\Delta/2$ for $d_n = 1$.

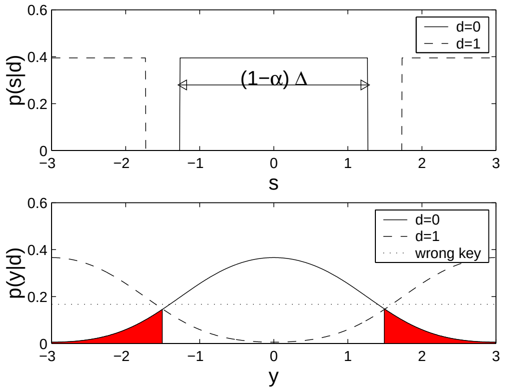

Fig. 1. One period of the PDFs of the sent and the received signal for binary SCS ($\sigma_{\mathbf{w}}^2$=1, WNR = 3dB, $\Delta = 6$, $\alpha = 0.58$). The filled areas represent the probability of detection errors assuming $d = 0$ was sent. The dotted line in the lower plot depicts the PDF when detecting with a wrong key $\mathbf{k}$.

The upper plot of Fig. 1 depicts one period of the PDF of the sent elements s conditioned on the sent watermark letter and $k_n = 0$. The lower plot shows the PDF of the pre-processed received elements y after AWGN attack conditioned on the sent watermark letter. The derivation of $p_{\mathbf{y}}\left(y_n|d_n\right)$ is given in [2]. In case of using an incorrect key $\mathbf{k}$ at the receiver, the distribution of $p_{\mathbf{y}}\left(y_n|d_n\right)$ will be uniform for any possible $\mathbf{r}$. This is indicated by the dotted line in the lower plot of Fig. 1.

3. SCS WATERMARKING FOR AUTHENTICATION

SCS watermarking is applied to the authentication problem. First, detection from one data element is considered, which is then generalized to more robust detection from a group of data elements.

3.1. Detection from One Data Element

A detector is designed that decides for r_n between the correctness of the

- test hypothesis H_0: the watermark letter $d_n = 0$ was embedded with key k_n – meaning the data has not been changed severely – and the
- alternative hypothesis H_1: the watermark letter $d_n = 0$ was **not** embedded with key k_n – meaning non-authentic data has been detected.

In this context, the *false positive* probability p_{FP} denotes the probability that r_n is considered non-authentic, although it is authentic. Conversely, the *false negative* probability p_{FN} denotes the probability of deciding for H_0 although H_1 would be correct. It is possible to trade off both error probabilities. Bayes' solution is the decision rule

$$\frac{p_{\mathbf{r}}\left(r_n|H_1\right)}{p_{\mathbf{r}}\left(r_n|H_0\right)} \begin{cases} > T & \Rightarrow \quad \text{accept } H_1\\ \le T & \Rightarrow \quad \text{accept } H_0, \end{cases} \tag{3}$$

where the decision threshold T is a constant depending on the a priori probabilities for H_1 and H_0 and the cost connected with the different decision errors [7]. For $T = 1$, the decision rule (3) forms a **maximum-likelihood (ML) detector**. For equal a priori probabilities, the overall detection error probability is $p_e = \frac{1}{2}(p_{\mathrm{FP}}+p_{\mathrm{FN}})$. We apply ML detection which can be formulated also as

$$L = \frac{p_{\mathbf{r}}\left(r_n|H_1\right)}{p_{\mathbf{r}}\left(r_n|H_0\right)+p_{\mathbf{r}}\left(r_n|H_1\right)} \begin{cases} > 0.5 & \Rightarrow \quad \text{accept } H_1\\ \le 0.5 & \Rightarrow \quad \text{accept } H_0, \end{cases} \tag{4}$$

which is convenient since $0 \le L \le 1$.

Note that the described hypothesis test is different from the decision between the watermark letter "0" or "1" as in the case of watermark communication. In the latter case, the embedder can modify the host data such that the value y_n is concentrated around 0 or $\pm\Delta/2$. In authentication applications, the received signal in case of H_1 can have any structure and cannot be influenced by the authentication mechanism. The hypothesis H_1 is equivalent to detection with a wrong key $\mathbf{k}$. Thus, the hypothesis test is based on the PDFs

$$\begin{aligned} p_{\mathbf{y}}\left(y_n|H_0\right) &= p_{\mathbf{y}}\left(y_n|d_n=0\right) \qquad (5)\\ p_{\mathbf{y}}\left(y_n|H_1\right) &= \frac{1}{\Delta}. \qquad (6) \end{aligned}$$

The decision between H_0 and H_1 cannot be as reliable as the decision between $d_n = 0$ and $d_n = 1$. Fig. 2 shows the minimal error probabilities p_e for both detection cases at different watermark-to-noise power ratios (WNR=10 $\log_{10} \frac{\sigma_{\mathbf{w}}^2}{\sigma_{\mathbf{v}}^2}$), assuming an attack with AWGN $\mathbf{v}$. Particular at high WNRs, the flat PDF $p_{\mathbf{y}}(y_n|H_1)$ intersects more strongly with $p_{\mathbf{y}}(y_n|H_0)$, which leads to the much higher error rates in case of authentication. Due to the different detection scenario, the SCS parameter Δ and α have been optimized with respect to the minimum p_e for each WNR. However, the resulting values of Δ and α are similar to those given in [2], and thus we can use the latter one.

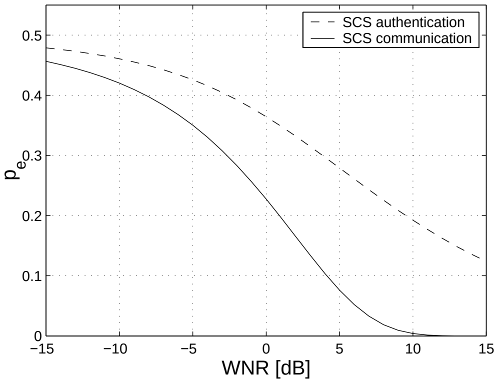

Fig. 2. The minimum detection error probability for binary SCS in case of communication and authentication for different WNRs after AWGN attack. Significantly higher error probabilities occur in the authentication case since the non-authentic data can have any structure.

3.2. Detection from a Group of Data Elements

The error probabilities shown in Fig. 2 are definitely too high for practical applications. Thus, the detection values from several data elements have to be combined. Consider detection from the preprocessed received elements $\mathbf{y}_{\mathcal{R}}$. Assuming that the host data and the distortion constrained attack is independent identically distributed (I.I.D.), detection from $\mathbf{y}_{\mathcal{R}}$ is based on

$$p_{\mathbf{y}}(y_{\mathcal{R}}|H_0) = \prod_{n \in \mathcal{R}} p_{\mathbf{y}}(y_n|H_0) \quad (7)$$

$$p_{\mathbf{y}}(y_{\mathcal{R}}|H_1) = \prod_{n \in \mathcal{R}} p_{\mathbf{y}}(y_n|H_1). \quad (8)$$

The L-value of ML detection can be effectively computed by

$$L = \frac{1}{1 + \exp\{\sum_{n \in \mathcal{R}} \log p_{\mathbf{y}}(y_n|H_0) - \log p_{\mathbf{y}}(y_n|H_1)\}}. \quad (9)$$

Unfortunately, it is possible that some of the elements selected by $\mathcal{R}$ belong to H_0 and some to H_1. The outcome of L depends on the dominating hypothesis. For ordered data $\mathbf{x}$, the elements from a local neighborhood or region should be put into $\mathcal{R}$. A sliding window can ensure that $\mathcal{R}$ matches to a local manipulation of the data.

3.3. Authentication of Non-White Host Data

So far, white host data statistics were assumed. In practice, most data will be colored. Due to space constraints, we cannot discuss this case here. A more detailed discussion and further references for watermark communication in case of colored host data can be found in [3, 8]. There, the data is decomposed into *sub-channels* which contain approximately white data. For each sub-channel watermarking schemes like SCS and attacks like the GTC can be applied. The optimal allocation of the watermark power and the attack power can be found numerically. This approach can be used also for authentication. We only have to account for different PDFs $p_{\mathbf{y}}(y_n|H_0)$ when detecting from differently reliable sub-channels.

4. EXPERIMENTS

4.1. Sliding Window Detection for Synthetic Data

We consider a data vector $\mathbf{x}$ of length 2000 to illustrate detection from a group of data elements using a sliding window. The SCS watermark is embedded with power $\sigma_{\mathbf{w}}^2$ followed by an AWGN attack with noise power $\sigma_{\mathbf{v}}^2$. Next, two data blocks were replaced by random data. A sliding window of length $R = 140$ is moved over the ordered data elements. For each window position, the set $\mathcal{R}$ is redefined and the corresponding L-value is computed. Example results are shown in Fig. 3. Some detection errors occur when the window only partly overlaps with the non-authentic regions. Nevertheless, the non-authentic regions could be located quite accurately. In this example, only one false positive error occurred when the window covered completely authentic data.

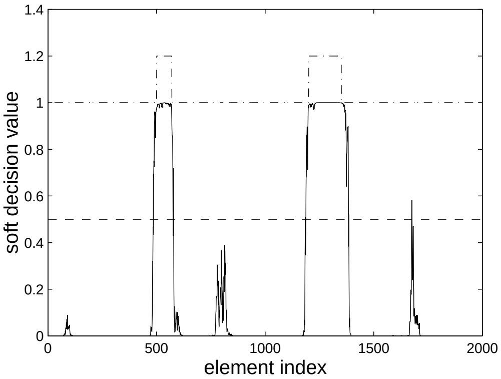

Fig. 3. Sliding window detection after AWGN attack with WNR=-3dB. Two blocks of length 70 and 150 were replaced. The solid line depicts the L-value at the position of the window center. The dashed line is the threshold for hard decision; $L > 0.5$ indicates non-authentic data. The dash-dotted line depicts the actual replacement pattern.

4.2. Detection Error Rates for Image Data

Authentication of a gray-scale image is investigated. An 8×8 block DCT was used to decompose the image into 64 sub-channels; each frequency is considered a sub-channel. Only the 2nd to 21th coefficients in zig-zag scan were selected for watermark embedding according to the results in [3, 8]. The DC-coefficient is not watermarked to avoid block artifacts due to the structure of the decomposition. Besides SCS authentication, a reference scheme based on spread-spectrum (SS) watermarking as in [9] was implemented.

The simulations were conducted for the test image "girl" of size 480×736. The watermarked image had a PSNR (peak signal-to-noise ratio) of about 40.5 dB. The watermarked image was JPEG compressed with quality factors from 10 to 100. Next, checkered arranged blocks of size 32×32 of the watermarked and already distorted image were replaced by the corresponding blocks of the unwatermarked host image. This ensures realistic statistics of the non-watermarked image blocks. The detection region $\mathcal{R}$ was also a block of size 32×32 matched to the positions of the checker board fields. The experiment was performed for 1000 different pseudo-random keys to obtain statistically meaningful results. Note that the watermark detectors were not tuned to the specific JPEG attack, e.g. the quantization step size for different DCT coefficients. Nevertheless, $p_e < 10^{-3}$ could be achieved for SCS authentication and an attacked image PSNR > 36dB as shown in Fig. 4. The performance of SS authentication is limited to $p_e \approx 0.05$ due to large host signal interference.

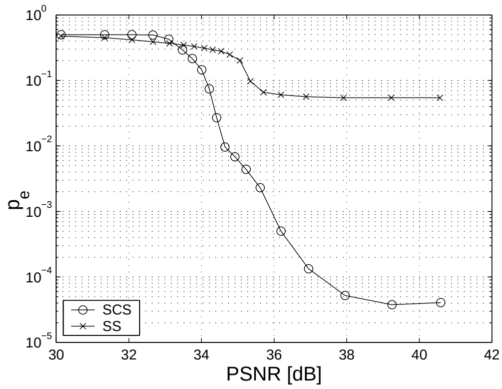

Fig. 4. Detection error rates for SCS and SS image authentication after JPEG compression and replacement of image blocks.

4.3. Example Image Manipulation

The test image "girl" was watermarked as in the previous subsection. Then, the watermarked image has been manipulated in four different regions and JPEG compressed with quality 70, which is slightly stronger than the default quality factor. The watermarked and the manipulated image are shown in the upper row of Fig. 5. Next, sliding window detection with a window of size 32×32 has been applied. Note that the window shift is restricted due to the 8×8 block DCT used for watermark embedding. The detection results are shown in the lower row of Fig. 5 with and without the manipulated image. Dark spots indicate non-authentic data. All manipulations have been detected, even the small extension of the band in the hair. One region was falsely classified non-authentic. Here, the host image has been almost white in contrast to the light gray in other background regions. Thus, JPEG compression quantized all DCT coefficients (except for the DC coefficient) to zero, which led to the detection error. This shows a fundamental problem of authentication based on semi-fragile watermarks since robust watermarking of flat image regions is almost impossible.

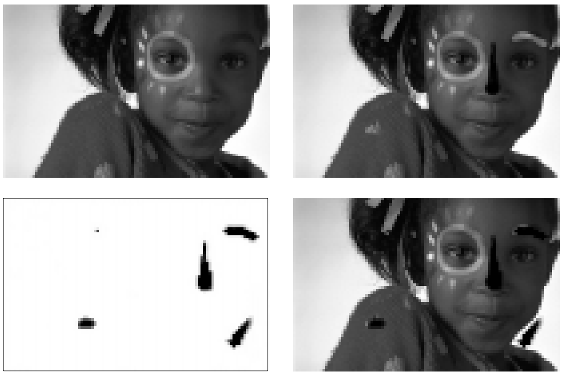

Fig. 5. Upper left: watermarked image; upper right: manipulated and JPEG compressed image; lower left: detected non-authentic regions; lower right: detected non-authentic regions on top of the manipulated image.

5. CONCLUSIONS

Authentication of image data using semi-fragile watermarks was investigated. In this application, the original image cannot be used by the watermark detector. Scalar Costa Scheme (SCS) watermarking was used since robust detection from small image regions can be achieved. In contrast, spread-spectrum watermarking is not appropriate due to large host signal interference. We demonstrated that authentication of compressed data with SCS watermarks has low error probabilities. However, it was also shown that SCS authentication cannot be as reliable as SCS communication since non-authentic data can have any structure. Further, robust watermarking of flat image regions is almost impossible, thus, leading to false detection for such regions.

6. REFERENCES

[1] R. B. Wolfgang, C. I. Podilchuck, and E. J. Delp, "Perceptual Watermarks for Digital Images and Video," *Proceedings of the IEEE, Special Issue on Identification and Protection of Multimedia Information*, vol. 87, no. 7, pp. 1079–1107, July 1999.

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

[3] J. J. Eggers, J. K. Su, and B. Girod, "Robustness of a Blind Image Watermarking Scheme," in *Proceedings of IEEE International Conference on Image Processing (ICIP 2000)*, Vancouver, Canada, September 2000.

[4] 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.

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

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

[7] I. A. Glover and P. M. Grant, *Digital Communications*, Prentice Hall, London, New York, Toronto, Sydney, 1998.

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

[9] J. J. Eggers and B. Girod, "Watermark Detection after Quantization Attacks," in *Proceedings Third International Workshop on Information Hiding*, Andreas Pfitzmann, Ed., Dresden, Germany, September/October 1999, vol. 1768, pp. 172–186, Lecture Notes in Computer Science, Springer.