

Watermark Removal based on Non-linear Filtering

Gerrit C. Langelaar, Reginald L. Lagendijk, Jan Biemond

Faculty of Information Technology and Systems (ITS),
Information and Communication Theory Group,
Delft University of Technology,
P.O. Box 5031, 2600 GA Delft, The Netherlands
E-mail: {gerhard, lagendijk, biemond}@it.et.tudelft.nl

Keywords: Watermarking, watermark attacks, non-linear filtering

Abstract

Many watermarking methods are based on adding pseudo-random noise in the spatial domain. In general, the robustness of the watermark is determined by measuring the resistance to JPEG-compression, adding Gaussian noise and applying linear filters. Using these processing techniques the quality of the image must be affected significantly before the watermark is removed. In this paper a method is proposed to estimate a pseudo-random spread spectrum watermark only from the watermarked image. If this estimated watermark is subtracted from the watermarked image, the watermark is removed without distorting the image significantly.

1. Introduction

Many spatial spread spectrum watermarking methods are proposed in literature [1..9]. Basically, these methods add a pseudo-random pattern to an image in the spatial domain to embed a watermark. This watermark can be detected by correlating with the same pattern or by applying other statistics to the watermarked image.

For our initial experiments we use the basic spread spectrum implementation of Smith and Comiskey [1]. They use a Direct-Sequence Spread Spectrum method, which divides the image first into $n_x n_y$ blocks to store a string of watermark bits in an image. Each $n_x n_y$ block contains one of the watermark bits. A modulation function, a constant integral valued gain factor G multiplied by a pseudo-random block of bits, either +1 or -1, is added to each image block. A positive gain factor G is used to embed watermark bit "1" in an image block, otherwise a negative gain factor G is used. The watermark is recovered by demodulating with the modulation function.

Many variants of this approach exist. Only a few will be mentioned here. Bender *et al* [2], and Pitas and Kaskalis [3] describe two precursors of this watermarking method, which add one watermark bit

to an image. Langelaar *et al* [4] extend the method described in [3] to store more watermark bits in one image and to find optimal gain factors for each pseudo-random block. Hartung and Girod [5] extend method [1] for real-time watermarking of MPEG video.

In general, the strength of the method is determined by measuring the resistance to JPEG-compression [10], to adding Gaussian noise and to linear filtering. Using these processing techniques the quality of the image must be affected significantly before the watermark is removed. The question is if we can express the strength of a watermark in terms of resistance against JPEG compression or filtering. Other attacking techniques should be investigated which use information of the image contents and the watermark itself.

In this paper a method is proposed to estimate the pseudo-random spread spectrum watermark from the watermarked image only. If a nearly perfect estimation of the watermark can be found, this estimated watermark can be subtracted from the watermarked image. In this way the watermark is removed without affecting the quality of the image.

In Section 2 the properties of spread spectrum watermarks are described. In Section 3 a correlation attack on these watermarks is discussed. The new attack is introduced in Section 4. Experimental results of this attack are presented in Section 5. Finally conclusions are drawn.

2. Properties of Spread Spectrum Watermarks

If we apply the method of Smith and Comiskey to an image I , a random pattern W consisting of the constants $-c$ and $+c$ is added to obtain the watermarked image $I_W = I + W$, where c is a positive integer value. The watermark energy resides in all frequency bands. Compression and other degradations may remove signal energy from certain parts of the spectrum, but since the energy is distributed all over the spectrum, some of the

watermark remains. The random pattern W is uncorrelated with image I , but correlated with I_W :

$$\begin{aligned} \text{cov}(W, I+W) &= \text{var}(W) + \text{cov}(I, W) \approx \text{var}(W) + 0 \\ \rho(W, I+W) &= \frac{\text{cov}(W, I+W)}{\sqrt{\text{var}(W)}\sqrt{\text{var}(I+W)}} \approx \sqrt{\frac{\text{var}(W)}{\text{var}(I+W)}} \\ \rho(W, I+W) &\approx \frac{c}{\sqrt{\text{var}(I+W)}} \end{aligned} \quad (1)$$

Evaluation of Equation 1 for typical images yields that ρ ranges from 0.02 to 0.05. However, if the watermarked images are compressed using the JPEG algorithm or distorted, the approximation in Equation 1 does not hold. Indeed, the correlation coefficients decrease by a factor 2, while the variance of $(I+W)$ nearly equals the variance of the JPEG compressed version of $(I+W)$.

If an arbitrary random pattern W_x is used, the correlation coefficient will be very small:

$$\begin{aligned} \text{cov}(W_x, I+W) &= \text{cov}(W_x, W) + \text{cov}(W_x, I) \approx 0 + 0 \\ \rho &= \frac{\text{cov}(W_x, I+W)}{\sqrt{\text{var}(W_x)}\sqrt{\text{var}(I+W)}} \approx 0 \end{aligned} \quad (2)$$

This holds only if W and W_x are orthogonal and W_x is not correlated with I . Typical values for another random watermark W_x and I_W are a factor 10^2 smaller.

3. Correlation Attack

A simple attack would be to search for all possible random patterns and take the one with the highest correlation value as possible watermark pattern. This approach has several disadvantages. In the first place the search space is huge. Even if the watermark should meet the requirement that the number of $-c$'s and the number of $+c$'s are equal, more than 4×10^{306} possible patterns have to be checked for a 32×32 pixel watermark. As a first step, we carried out experiments with a genetic algorithm to search the random pattern with the highest correlation coefficient with $I_W = I+W$. In some cases the genetic algorithm found a pattern with a relative high correlation (0.3) with I_W and no correlation with W (10^{-5}). This means that the pattern is adapted to the image contents and not to the watermark.

To avoid that the genetic algorithm finds random patterns with higher correlation coefficients than the embedded watermark we must adapt our fitness function. From the properties of spread spectrum watermarks we know the following about W :

- $\rho(W, I_W) \in [0.01 \dots 0.05]$
- $\rho(W, I) \approx 0$

- W is pseudo-random and has a flat spectrum

If the image is distorted by compression, $\rho(W, I_W)$ is unknown. Too many patterns meet the requirement $\rho(W, I) \approx 0$. The additional information that W is random and has a flat spectrum is also not enough to create a suitable fitness function. If we have several different images with the same watermark on it to our disposal, there are some possibilities (e.g. collusion attacks). A fitness function for the genetic algorithm dependent on all images can be used, or if there are enough images, the average of the images can be taken as estimation of the watermark. But if different watermarks are used for each image, we have to follow another approach.

4. Estimating Watermark using Image Information

In general, a watermark can be regarded as a perceptually invisible enforced distortion in the image. In most cases, this distortion is not correlated to the image contents. If we could apply a nearly perfect image model to the watermarked image $I_W = I+W$, we could predict the image content \hat{I} and find back an estimate of the watermark $\hat{W} = I_W - \hat{I}$. Because perfect image models and perfect noise filters do not exist, \hat{I} will be different from I and \hat{W} will be different from W . Our objective is to separate $I_W = \hat{I} + \hat{W}$ in such a way that the watermark is totally removed from \hat{I} and resides completely in \hat{W} . This means that image contents may remain in the predicted watermark.

A NSHP-causal-AR-model, linear smoothing filters (3x3 and 5x5), Kuwahara filters (several sizes), non-linear region based filters and filters based on thresholding in the DCT-domain are tested to separate I_W in \hat{I} and \hat{W} . In some cases, the watermark can be retrieved from both \hat{I} and \hat{W} , while \hat{I} has still a reasonable quality and \hat{W} does not contain any image information. In other cases the watermark can only be retrieved from \hat{W} , but the quality of \hat{I} is significantly affected and the image contents, especially the edges, remain in \hat{W} .

We select some candidates from the separation operations that totally destroy the watermark in \hat{I} , $\rho(W, \hat{I}) \approx 0$. From these candidates we select the operation that has the highest correlation coefficient $\rho(\hat{W}, W)$ in a test set of 9 images. In Table 1 the correlation coefficients for several separation operations are listed. The 3x3 median filter turns out to be the best separation operation and is used for the rough estimation of $\hat{W} = I_W - \text{med}_{3 \times 3}(I_W)$. However, correlation coefficients $\rho(\hat{W}, W)$ between 0.13 and 0.22 are still too low and \hat{W} must be

refined further by using information about the watermark properties.

Table 1. Correlation coefficients $\rho(\hat{W}, W)$ using different separation operations.

Separation Operation	$\rho(\hat{W}, W)$
Misc. Noise Reduction Filters	0.08-0.12
Auto Regressive Model	0.10-0.17
Median 3x3	0.13-0.22

The estimate \hat{W} does still contain edge information. To protect the edges in I_W we limit the range of \hat{W} from $[-128..128]$ to $[-2..2]$ before we subtract \hat{W} from I_W . In Figure 1 the modulus of the Fourier Transform of the truncated \hat{W} is presented.

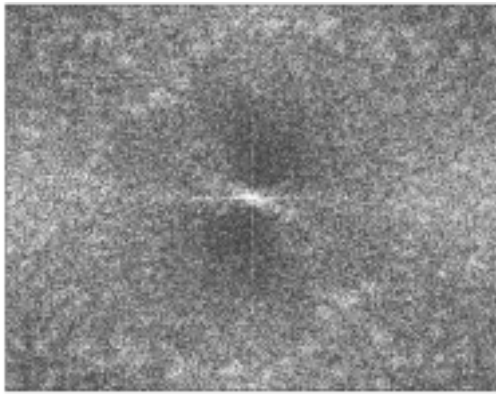


Figure 1. Power density spectrum of $\hat{W}_{[-2..2]}$.

The horizontal, vertical and diagonal patterns in Figure 1 clearly indicate that some dominating low frequency components are present in the spectrum. Since a spread spectrum watermark should not contain such dominating components, these come certainly from the image contents. To remove these components a 3x3 linear high pass filter is applied to the non-truncated \hat{W} . After the filtered \hat{W} is truncated to the range $[-2,2]$ the Fourier spectrum as presented in Figure 2 is obtained.

The correlation coefficients between the high pass filtered \hat{W} and W , $\rho(\hat{W}, W)$, increase to values around 0.4.

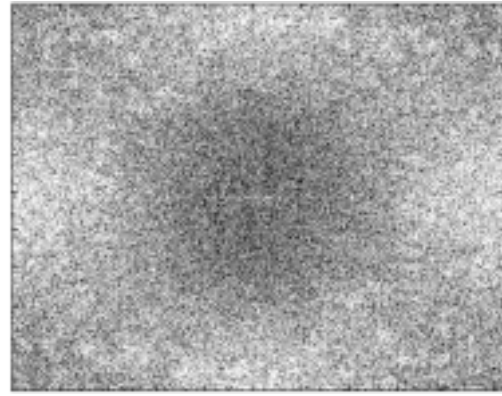


Figure 2. Spectrum of high-pass(\hat{W}) $_{[-2..2]}$.

If the so found watermark \hat{W} is subtracted from the watermarked image I_W the watermark is not completely removed. This is not surprising, since we are not able to predict the low frequency components of the watermark. These components can be left in \hat{I} by the median filter or are discarded during the high pass filtering stage of \hat{W} . To compensate for the fact that we have only found a part of the watermark, \hat{W} must be amplified with a certain gain factor A . The complete scheme for removing a watermark is represented in Figure 3.

The value of A is dependent on the image content and the amount of energy in the embedded watermark. If A is chosen too high, the watermark inverts and can still be retrieved from \hat{I} by inverting the image before retrieving the watermark.

The value A is experimentally determined. A watermark is added to an image using the method of Smith and Comiskey [1], 32x32 pixels are used to store one bit of watermark information and the watermark carrier consists of the integers $\{-2,2\}$. The watermark removing scheme is applied to the watermarked image with several values for A . The percentage watermark bit errors is plotted as function of A in Figure 4. If 50% bit errors are made, the watermark is removed, if 100% bit errors are made, the watermark is totally inverted.

According to Figure 4 the gain factor A should have a value between 2 and 3 to remove the watermark from this image.

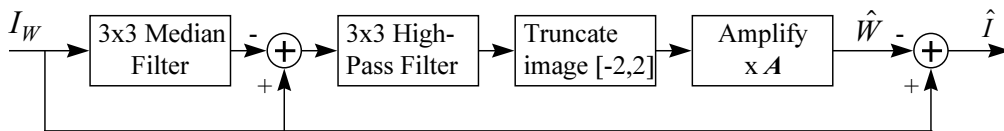


Figure 3. Complete watermark removing scheme.

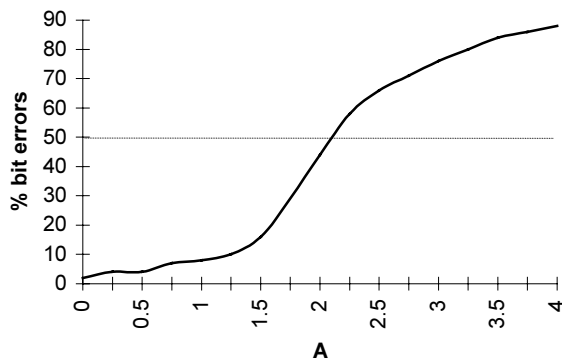


Figure 4. Bit errors as function of gain factor A .

5. Experimental Results

We tested the watermark removing scheme (WRS) represented in Figure 3 on a set of 9 true color images. Informal subjective tests were performed to determine the quality of the images. Some images hardly contain any textured areas and sharp edges, some contain many sharp edges and much detail, others contain both smooth and textured areas.

First, the WRS ($A=2.5$) is applied to the methods of Bender *et al* [2] and Pitas and Kaskalis [3]. The watermarks in the 9 test images are all removed without reducing the quality of the images significantly.

Subsequently the WRS is applied to the watermarking method of Smith and Comiskey[1]. The watermarks are added using $n \times n$ pixels per bit and a gain factor of G , where $G=1$ or 2. If higher gain factors G are used the watermark becomes visible. For the values $n=8,16,32,64$ and $G=1,2$ the watermarks can be removed without affecting the quality significantly. An example is given in Figures 5,6 and 7. Figure 5 represents a watermarked image $G=2$, $n=32$. To remove the watermark completely (about 44% bit errors) using the JPEG compression algorithm, we have to use a quality factor $Q=10$. The result of this compression operation is presented in Figure 6. If we apply the WRS to Figure 5, the watermark is completely removed ($>50\%$ bit errors) and we obtain the image which is shown in Figure 7. This image hardly distorted. If the blocksize n is increased further to 128 or 256, the watermark is fully removed in smooth images, but only partially in textured images.

Finally, the WRS is applied to the method of Langelaar *et al* [4]. This method determines the gain factor G for each watermark bit automatically. Therefore only the blocksize n can be changed. All watermarks added with this method can be removed for $n=8,16,32$. For $n=64,128, \dots$ the watermarks are only partially removed.

Some methods (e.g. [9]) first subtract the original image from the watermarked image and apply the watermark retrieval operation on this difference image. However, the WRS also removes the watermarks in this case.



Figure 5. Watermarked Image.



Figure 6. Watermark removed by JPEG Compr.

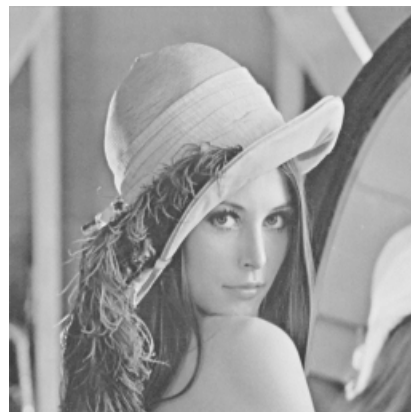


Figure 7. Watermark removed by new WRS.

Other methods using a similar approach as [1] are not tested, but we expect that their watermarks will be affected in the same way as [1], since they use the same basic principle.

6. Conclusions

In this paper we showed that determining the robustness of a watermark by measuring the resistance to JPEG-compression, adding Gaussian noise and applying linear filters is not sufficient. A simple method was proposed to remove a watermark without affecting the image quality significantly. The existence of this removal technique has certainly influence on the number of watermark bits which can be stored per image in a robust way.

References

- [1] J.R. Smith, B.O. Comiskey, "Modulation and Information Hiding in Images", Preproceedings of Information Hiding, an Isaac Newton Institute Workshop, University of Cambridge, UK, May 1996
- [2] W. Bender, D. Gruhl, N. Morimoto : "Techniques for Data Hiding", Proceedings of the SPIE, 2420:40, San Jose CA, USA, February 1995
- [3] I. Pitas, T. Kaskalis : "Signature Casting on Digital Images", Proceedings IEEE Workshop on Nonlinear Signal and Image Processing, Neos Marmaras, Greece, June 1995
- [4] G.C. Langelaar, J.C.A. van der Lubbe, R.L. Lagendijk, "Robust Labeling Methods for Copy Protection of Images", Proceedings of Storage and Retrieval for Image and Video Databases V, San Jose (CA), USA, February 1997
- [5] F. Hartung and B. Girod, "Watermarking of MPEG-2 Encoded Video Without Decoding and Re-encoding", Proceedings Multimedia Computing and Networking (MMCN 97), San Jose, CA, USA, February 1997
- [6] M. Kutter, F. Jordan, F. Bossen, "Digital Signature of Color Images using Amplitude Modulation", Proceedings of Storage and Retrieval for Image and Video Databases V, San Jose (CA), USA, February 1997
- [7] R.G. van Schyndel, A.Z. Tirkel, C.F. Osborne : "A Digital Watermark", Proceedings of the IEEE International Conference on Image Processing, volume 2, pages 86-90, Austin, Texas, USA, November 1994
- [8] G. Voyatzis and I. Pitas, "Applications of Toral Automorphisms in Image Watermarking", Proceedings ICIP-96, IEEE International Conference on Image Processing, Volume II pp 237-240, Lausanne, Switzerland, 16-19 September 1996
- [9] R.B. Wolfgang and E. J. Delp, "A Watermark for Digital Images," Proceedings of the IEEE International Conference on Image Processing, Volume III pp. 219-222, Lausanne, Switzerland, September 16-19, 1996
- [10] W.B. Pennebaker, J.L. Mitchell : "The JPEG Still Image Data Compression Standard", Van Nostrand Reinhold, New York, 1993

