Lessons from the BOWS Contest

Andreas Westfeld
Technische Universität Dresden
Institute for System Architecture
01062 Dresden, Germany
westfeld@inf.tu-dresden.de

ABSTRACT

In the first BOWS Contest, a watermarking system was exposed to attacks from all over the world—not to prove that the system is resistant to attacks or to claim that it is unbreakable, rather to learn more about the level of weakness and perhaps new possible attacks. The goal was to wipe the watermark while keeping the best possible quality of the images. This paper highlights some of our tests and results. We analysed the interaction between attack strength and quality measure and described two more successful approaches to remove the watermark: (1) to reduce artefacts that are introduced by the watermarking algorithm, which renders embedded information unreadable and (2) a specialised sensitivity attack.

Categories and Subject Descriptors

H.m [**Information Systems**]: Miscellaneous

General Terms

Security

Keywords

Digital Watermarking, Robustness, Attack

1. INTRODUCTION

Recently, the European Network of Excellence in Cryptology ECRYPT organised the first BOWS Contest (Break Our Watermarking System) [4]. BOWS exposed an algorithm for digital watermarking to a world-wide massive attack. According to the announcement [3], the contest was not intended to prove how well-performing the watermarking system is, but to learn more about the degree of difficulty of breaking the watermark and finding new possible attacks.

The first phase of the attack, in which no information about the watermarking system was given, started on December 15, 2005. The BOWS website [3] offered three watermarked greyscale images in raw format and 512×512 pixels

MM&Sec'06, September 26–27, 2006, Geneva, Switzerland.

Figure 1: Three watermarked 512×512 greyscale images in raw format

in size (see Figure 1). This raw format stored the unsigned pixel values in the file (8 bits per pixel) without any header, with the image scanned from left to right and from top to bottom.

To enter the Hall of Fame, participants had to make the watermark unreadable to the BOWS detector in all three images while maintaining an image quality rated by a PSNR (peak signal to noise ratio) above 30 dB. In the first six weeks, the Hall of Fame was empty, however, the maintainer of the BOWS website reported successful attacks on the strawberry image. PSNR is a simple and widely used quality metric based on the mean squared error (MSE), which is computed by averaging the squared intensity differences of the distorted image a and its reference b:

$$\mathrm{PSNR}(a,b) \quad = \quad 10 \cdot \lg \left(\frac{255^2 \cdot 512^2}{\sum_{x=1}^{512} \sum_{y=1}^{512} (a_{x,y} - b_{x,y})^2} \right) \quad (1)$$

To test whether the watermark is still readable, the attacked images were uploaded to the BOWS server through a web interface to run the detection process. The BOWS oracle replied the request by giving the result of the detection (watermarked removed/still there) and the PSNR achieved. Note that the oracle always reports the result of the watermark detection regardless of the PSNR criterion.

After the first phase ended on March 15, 2006, the BOWS organisers revealed that the watermarking algorithm proposed by Miller et al. [7] was used to embed the watermark into the three images.

The second phase of the contest (with no prize) remained open for additional three months. This paper contains attacks from both phases and is organised as follows: Section 2 presents some basic attacks, in Section 2.1, some pixel-oriented attacks are considered, against which most watermarking systems are rather robust. These attacks will visibly degrade the image before the PSNR measure is affected. In Section 2.2, we will look at basic geometric attacks that are more effective against watermarking systems. However,

because geometric attacks will change almost all pixels, the PSNR based quality measure will decrease before the image is visually degraded. Section 3 examines a different approach that identifies artefacts introduced by the marking algorithm and tries to remove the watermark by reducing distortions. Section 4 presents the sensitivity attack, which obtained the best PSNR values. The paper is summarised in Section 5.

2. BASIC ATTACKS

A BOWS participant has two goals:

1. to make the watermark unreadable to the oracle and
2. to preserve a PSNR > 30 dB.

It is easy to find attacks that achieve the first goal. However, to judge whether the first goal is attained, we have to ask the oracle. This might be a bottleneck, since the access to the oracle is limited in response time and number of requests. When BOWS started, there was a limit of 40 requests per day and client IP address. The BOWS oracle processes one request in about two seconds or more, which depends on the server load. In January, the limit was increased to 5000 requests per day (and decreased to 3000 by end of March).

The PSNR score is said to privilege classes of attacks that do not change pixel positions. Geometric distortions, which stretch, shear, shift, or rotate the image by an unnoticeable amount [9], were at a disadvantage.

A better (higher) PSNR value does not mean that the attacked image is visually close to the marked original. For example, the woodpath in Figure 2 (left) is shifted to the right by one pixel, which is rather unobtrusive, while the clearly visible distortions in the sky (right) maintain a PSNR of more than 30 dB. (The watermark is still detected in both images.) The only "correct" method of quantifying visual image quality is through subjective evaluation [10]. Here, however, we need a simple, quick, and automated evaluation.

Figure 2: Two images with counterintuitive PSNR: 19.84 dB (left) and 30.83 dB (right)

2.1 PSNR-Friendly Attacks

2.1.1 Additive White Gaussian Noise

Noise adding is a standard evaluation procedure when developing watermarking systems. We scaled the standard deviation of 512×512 normal distributed random values and added them to the image. Table 1 shows the detection boundary determined using the oracle and repeated interval bisection [1] for three different sets of random numbers from the same generator and their respective PSNR. The present watermarking system is very resistant against added noise. The visual quality of the image is heavily degraded by the noise as we can see in the magnified detail from the strawberry image (Figure 3).

Table 1: Standard deviation σ for Gaussian noise to remove the watermark and the achieved PSNR for three repeated tests (saturated values clipped)

measurement	strawberry	woodpath	church
first	$\sigma = 26.7$	$\sigma = 37.4$	$\sigma = 30.0$
	19.7 dB	17.2	18.8 dB
second	$\sigma = 26.8$	$\sigma = 33.5$	$\sigma = 29.9$
	19.7 dB	18.1	18.8 dB
third	$\sigma = 31.7$	$\sigma = 29.3$	$\sigma = 26.5$
	18.3 dB	19.2 dB	19.8 dB

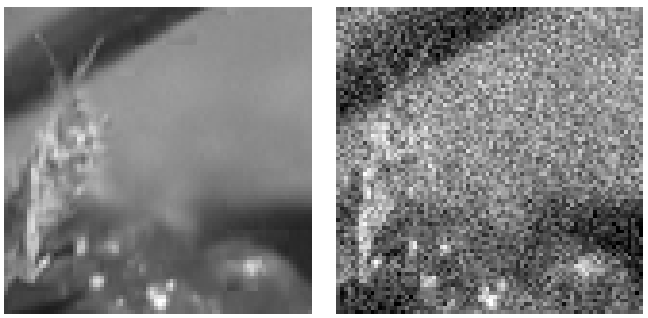

Figure 3: The bug on the strawberry before and after adding minimal noise to remove the watermark (first measurement)

2.1.2 Scaling the Brightness

The watermarking system is even more robust against brightness manipulation. We reduced the brightness of all pixels by a scale factor, rounded the scaled values, and searched for the detection boundary. Table 2 lists the scale factors for the resulting images. These images contain only four shades $\{0, 1, 2, 3\}$. However, the detection boundary is blurred by rounding effects or even deliberately randomised. It is possible to find smaller scale factors for which the oracle still detects the watermark: Figure 4 shows an example for the strawberry that has only three grey levels $\{0, 1, 2\}$ for the scale factor 0.00758. The contrast has been maximised because otherwise the image appears black.

Table 2: Brightness scaled to remove the watermark

Image	scale factor	Watermark	PSNR
strawberry	0.01035	still there	6.136 dB
	0.01034	removed	6.135 dB
woodpath	0.01102	still there	6.858 dB
	0.01101	removed	6.857 dB
church	0.01191	still there	4.745 dB
	0.01190	removed	4.745 dB

2.1.3 Cropping

Finally, we determined the detection boundary for cutting off stripes from the left, top, right, and bottom margin, re-

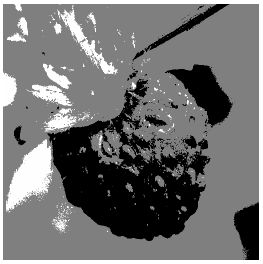

Figure 4: The watermark in the strawberry image is still detected if the brightness is reduced to 3 levels of grey

spectively. The width of the stripes depends on the image content (see Table 3). Consequently, the watermarking sys-

Table 3: Amount of pixels to cut off from the four margins in order to remove the watermark and respective PSNR

Image	left	top	right	bottom
strawberry	220	210	250	307
	16.1 dB	18.7 dB	17.0 dB	14.4 dB
woodpath	117	210	163	117
	20.0 dB	15.7 dB	14.8 dB	18.9 dB
church....	237	241	65	123
	19.3 dB	19.2 dB	23.3 dB	18.9 dB

tem is adaptive to image content. The narrowest stripe was found for the right margin of the church image (see Figure 5). However, the stripe cut from the top margin has to be almost four times wider.

Figure 5: To remove the watermark, either 241 rows of pixels from the top or 65 columns of pixels from the right have to be replaced by neutral grey in the church image

2.2 Geometric Attacks

Geometric attacks remove the watermark by introducing a visually unnoticeable quality loss. Most marking systems available on the market are confused by simple geometrical distortions [9]. In this section, we will rotate and shift the images and see whether this is a promising approach or foiled by the PSNR measure.

2.2.1 Rotation

A rotation of the image by an angle as small as $\alpha = 1^\circ$ already corrupts the watermark. We can minimise this angle and increase the PSNR using the bisection method. Table 4 shows the optimal results for left rotation. The image is rotated by a small angle around its centre and then resampled by bi-linear interpolation. The area outside the image that appears inside after rotation was supposed to be neutral grey.

Table 4: Images rotated as a whole by a minimum angle to remove the watermark

Image	α	PSNR
strawberry	0.534°	28.94 dB
woodpath	0.217°	22.9 dB
church....	0.264°	24.99 dB

2.2.2 Translation

Let $R = \{0, \ldots, 255\}^{512,512}$ be the set of raw images. Let $m \in R$ be the watermarked image. $m_{x,y}$ is the pixel in m at position (x, y), $x, y \in \{1, 2, \ldots 512\}$. Translation leads to an image s, shifted by width w:

$$s_{x,y} := m_{x+w,y}$$

Because PSNR is a based on the MSE (mean squared error), doubled magnitude in changes leads to quadrupled penalty in the quality measure. To avoid large changes, we define a limiting parameter l that keeps pixels unchanged if the absolute change is not smaller than l:

$$s_{x,y} := \begin{cases} m_{x+w,y} & \text{for } |m_{x,y} - m_{x+w,y}| < l \\ m_{x,y} & \text{else} \end{cases}$$

Table 5: Limited translation

shift width w:	1	2	3	4
strawberry:				
l for PSNR>30 dB	∞	46	34	30
l' to remove WM	—	60	42	39
PSNR/dB for l' and w	—	28.94	29.1	28.62
woodpath:				
l for PSNR>30 dB	21	21	22	22
l' to remove WM	59	33	40	37
PSNR/dB for l' and w	23.1	26.22	24.5	25.09
church:				
l for PSNR>30 dB	28	26	26	25
l' to remove WM	121	50	38	41
PSNR/dB for l' and w	24.57	25.51	27.02	26.27

2.3 Summary

Although the PSNR discriminates against geometric attacks, the effectiveness of these attacks compensates for this. The decision to score the attacks on the basis of the PSNR did not invert the strength of the attacks considered here. Geometric attacks preserve a PSNR, which is about 10 dB higher than for the "PSNR-friendly" attacks.

3. REDUCTION OF VISIBLE ARTIFACTS

Figure 6 shows blockiness artefacts from the watermarking process. These artefacts are mainly visible on the contours of solid areas in the image. It appears as if these distortions come from manipulated low-frequency coefficients of

the 8×8 pixel DCT. The goal of this section is to identify the coefficients and the amount they are modified by the watermark to revert the watermarking process and increase the image quality.

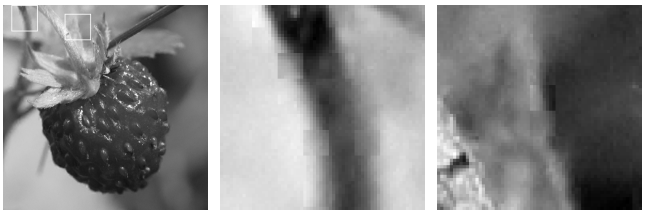

Figure 6: Clearly visible blockiness artefacts

3.1 Detection of DCT Modes with Influence

First we identify the DCT modes (all coefficients of one frequency are called mode here) that are used by the watermark. So we have to transform the three images to the DCT domain and replace all coefficients of one particular AC mode by 0 (the DC mode is replaced by 1024), the values that we obtain when transforming a neutral grey image. If the watermark is removed by this action, we know that a part of the watermark is embedded in this mode, assumed that the DCT domain is where the watermark is embedded (and this was confirmed after the watermarking algorithm was revealed by the BOWS organisers). Replacing single modes was not successful, so we continued by replacing all pairs, triples (successful pairs excluded), and quadruples of modes and found that in all three images only the modes depicted in Figure 7 (left) host the watermark.

3.2 Simple Blockiness Reduction

To reduce the blockiness, we can define a simple blockiness measure based on the difference of pixels at block boundaries. In Figure 7 (right) these pixels are marked grey. Each block has a north, west, south, and east boundary. The simple blockiness b_1 is defined by the following sum of squared differences:

$$b_1 = \sum_{i=1}^{8} (n_{1i} - n_{2i})^2 + (w_{1i} - w_{2i})^2 + (s_{1i} - s_{2i})^2 + (e_{1i} - e_{2i})^2$$

Note that the corner pixels belong to several boundaries at the same time, e. g., $n_{28} = e_{21}$. To change the blockiness, we modify the 12 DCT coefficients shown in Figure 7 (left) and find the minimum of b_1 using the conjugate gradients

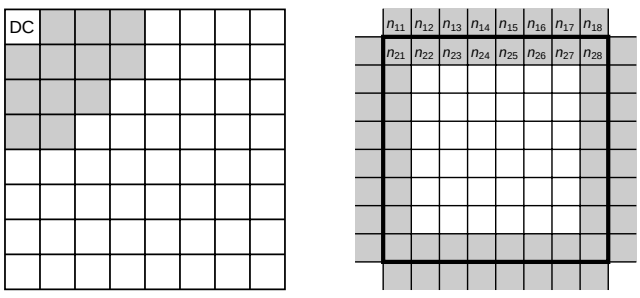

Figure 7: DCT modes used for the watermark (l.) and pixels considered for blockiness reduction (r.)

method by Fletcher and Reeves [5]. There are 64×64 blocks per image, for each block we get a modification vector with 12 elements. Pixels at the image boundary are assumed to have difference 0 to the outside. The surrounding of each block is taken from the original marked image. All blocks are optimised independently. We overcompensate a bit, when we apply the changes in full extent, since the gap is always seen from both blocks at the boundary. Therefore, our approach is a kind of generation model: Apply the changes with factor 0.5 and optimise the resulting image again. The best PSNR is achieved after 2 or 3 generations and this leads to successful attacks for all three images.

3.3 Blockiness Reduction with Gradient Correction

The blockiness measure b_1 is optimal if there is no gradient at the boundary. Suppose an image without watermark but with gradients. The optimisation of a gradient would yield a flattening at the boundary, because gradients cause larger b_1. Instead, the difference at the boundary should be derived from the surrounding. Figure 8 shows additional

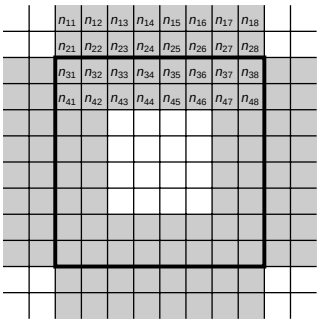

Figure 8: Pixels considered for blockiness reduction with gradient correction

pixels considered to estimate the boundary gradient. The mean of the differences of the surrounding should be equal to the difference at the boundary (2). This leads to a new, gradient corrected version of the blockiness measure b_2 (5).

$$n_{2i} - n_{3i} \approx \frac{(n_{1i} - n_{2i}) + (n_{3i} - n_{4i})}{2} \quad (2)$$

$$0 \approx n_{1i} - 3n_{2i} + 3n_{3i} - n_{4i} \quad (3)$$

$$b_{2,n} = \sum_{i=1}^{8} (n_{1i} - 3n_{2i} + 3n_{3i} - n_{4i})^2 \quad (4)$$

$$b_2 = b_{2,n} + b_{2,w} + b_{2,s} + b_{2,e} \quad (5)$$

The gradient aware reduction of blockiness results in 35.89 dB, 31.6 dB, and 32.2 dB PSNR for the three images.

4. SENSITIVITY ATTACK

The most promising attack against watermarking schemes with correlation based detector is the sensitivity attack introduced by Cox and Linnartz [2]. We applied the attack as follows:

1. We set up a marked image near the detection boundary. We replaced some of the 12 DCT modes by 0 until the watermark was removed. Then we restored one of these modes, and gradually replaced individual

coefficients of this mode by 0 until the watermark was removed. Finally we restored the last coefficient that we replaced, so that the watermark is still there.

2. To find individual coefficients that remove the watermark, we walked through the 4096 coefficients in one of the DCT modes that was not modified in the first step. Each single coefficient was incremented and decremented by 20 to trigger the oracle's decision with the resulting image.

3. Combining the knowledge on how sensitive the detector is to a modification of each coefficient, we estimated the combination of successful coefficients with the largest (expected) influence on the oracle decision. We used the original marked image and added the combination of modifications scaled by a minimal factor that just removes the watermark.

4.1 Misleading Results

We did not reach this third step in our first try. We found contradictory sensitivity for many coefficients. Out of 4096 coefficients that we tested, 801 removed the watermark when increased by 20, 801 removed the watermark when decreased by 20, and 2802 did not make the watermark unreadable to the oracle, neither for positive nor for negative modification. 308 coefficients removed the watermark for both, positive *and* negative modification.

Probably, this is an effect of the Viterbi decoder that randomises the detection boundary. Randomisation is a countermeasure proposed against the sensitivity attack [6]. Before we knew the underlying watermarking system [7], we suspected that dither modulation or another quantisation based data hiding method was used. This seemed very likely because the watermarking community in the ECRYPT network proposed such systems recently (e. g. RDM, [8]).

4.2 Minimalist Evidence

The required PSNR > 30 dB is maintained, when we change all pixels of the image by 8 (30.07 dB). How many pixels can be changed by the possible maximum (255) for the same PSNR? Since (1) is based on the squared error, we derive this number n of pixels from the equation $512 \cdot 512 \cdot 8^2 = n \cdot 255^2$, yielding $n \approx 258$. This means that we could replace a black 16×16 region of an image by white colour and still have the required quality. So we can keep a high PSNR when we change only few coefficients by a large amount and leave most of the image untouched. The coefficients that we can change by the largest amount must have a large absolute value. When we determined the maximum and the minimum coefficient for each of the 12 DCT modes and replaced them by the opposite saturated value (± 707 if the coefficient is in the DC row or column in the matrix shown in Figure 7 [left], otherwise ± 500), the watermark disappeared. This gave evidence that the scheme *is* correlation based.

We removed some of the 24 coefficients, which did not contribute but only decreased the PSNR. Then we tried to find better replacements for coefficients with small contribution We sorted the 12×4096 DCT coefficients by decreasing absolute value to find the most relevant coefficients first. In the following coefficients are referred to by their rank numbers $1 \ldots 49152$. After some iterations, we found out that only three coefficients must be changed in the strawberry image to remove the watermark. The best triple that we found is (1006, 2363, 4595) in rank numbers. Changing these coefficients by 608.5 (subtract this value if the coefficient is positive, add it otherwise) yielded a PSNR of 42.555 dB, which is better than the PSNR of the strawberry submitted by the winning team. In the woodpath change (17, 29, 998) by 820.4, and in the church (42, 133, 218, 415, 3395) by 516.2. The changes are clearly visible, however, honoured by 39.597 dB and 41.247 dB PSNR.

4.3 Towards 60 dB

We restarted the sensitivity attack from the minimalist tuples found by the method described in Section 4.2. Unlike the description of step 2 in Section 4 we did not just add or subtract a fixed amount to the current coefficient. Because oracle calls are a limited resource, we made sure that the PSNR will increase on the local system. Suppose, we have the three aforementioned coefficients for the strawberry image c_{1006}, c_{2363}, and c_{4595}, together with the current coefficient under sensitivity test c_1. We replace the four coefficients $\vec{c} = (c_{1006}, c_{2363}, c_{4595}, c_1)$ in our (DCT transformed) image by $\vec{c} - p \cdot \mathrm{sign}(\vec{c})$ and search for the minimal parameter p that yields a PSNR of at least 42.555 dB. For $(c_{1006}, c_{2363}, c_{4595}, c_1)$, $(c_{1006}, c_{2363}, c_{4595}, c_2)$, ..., the minimal parameter p is 501.5, 501.6, 501.5, 502.3, ..., i. e., p is not a constant value. We prepared the images with the locally determined values of p before we sent them to the oracle. If the oracle answered that the watermark has been removed, the currently tested coefficient contributes to a better quality and we remember it as a sensitive one. Note that we only checked one direction of sensitivity to this point. The other direction was not accessible yet, because p is still too large (and the change would get lost due to saturation). After about 2500 oracle calls we found 30 sensitive coefficients. We combined these 30 coefficients in $\vec{c'}$ and replaced them by $\vec{c'} - p \cdot \mathrm{sign}(\vec{c'})$, with the parameter p only about 80 and 49.028 dB PSNR.

Table 6: Results for sensitivity attack

	strawberry	woodpath	church
number of coefficients	424	279	185
absolute change	5.09	10.21	13.02
postprocessing limit .	1.83	2.08	3.19
PSNR/dB	60.74	57.05	57.29

Table 7: Histogram of the absolute value of differences between attacked and marked image

Image	0	1	2	3	4
strawberry	248052	14016	76	—	—
woodpath	248668	6872	6570	30	4
church	252656	4082	4223	1173	10

Tables 6 and 7 list the results after about 100000 oracle calls per image. We identified about 300 coefficients per image, the absolute change of the coefficients is about 10, and the PSNR is only 2 dB away from the 60 dB boundary. In the spatial domain this means that no grey shade is changed by more than 4. Over 95 % of the pixels remain unchanged. The changes are unobtrusive, if noticeable at all. Figure 9 shows the attacked strawberry image and an amplified difference to the marked original. Interestingly, the changes are not concentrated on specific areas. Although the watermarking system uses informed coding and informed embedding, the attack does not modify the images adaptively.

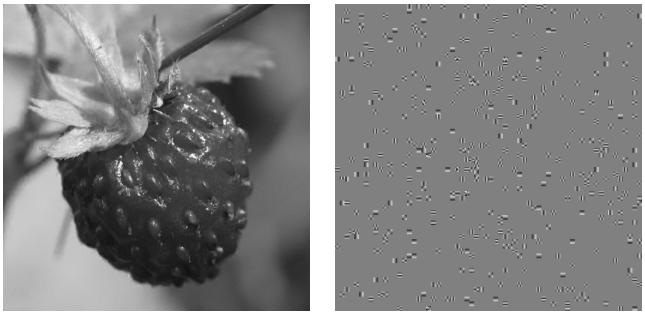

Figure 9: Attacked strawberry image with 60.74 dB PSNR and (amplified) difference to marked image

5. CONCLUSION

From all attacks that we tried in the contest, the sensitivity attack is the most general and the setup was the same for all three images. The pixel oriented and geometric attacks only worked image dependently. Figure 10 shows the learning

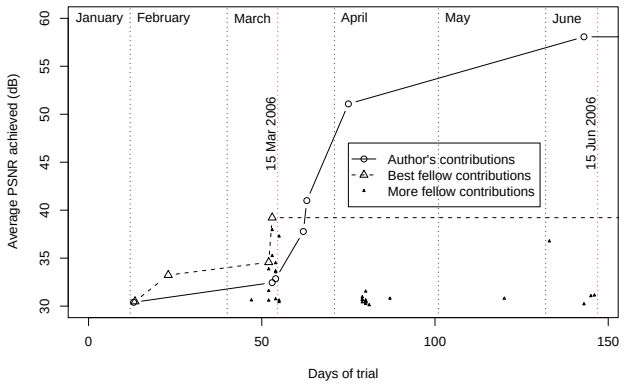

Figure 10: Learning curve in terms of quality over the two phases of the contest

curve in the contest. Our first successful attack in February was a geometric attack enhanced by a lowpass filter. This filter passes only modifications in the DCT modes shown in Figure 7 (left), which increases the PSNR by some dB. In mid-March, just before the end of phase 1, we submitted the blockiness equalised images according to Section 3.3. We started the sensitivity attack on March 24.

Knowledge of the algorithm helps to reduce the number of dead end approaches and could accelerate the selection of the right measure. In the end, however, the achieved quality would have been possible without knowing the paper of Miller et al. [7].

Although the quality defined by the PSNR measure does not always match human perception, it seems to be acceptable, since strong attacks still achieved the best PSNR. PSNR is a quality measure that is easy to implement and quickly evaluated on the local system, just suitable for optimal (iterative) preparation of the particular images that are sent to the oracle. Albeit the watermark strength is adaptive due to the informed embedding, we can see that the changes in Figure 9 are not: Sensitive coefficients can be found in areas with high and flat contrast. The subjectively perceived quality could be improved (at the expense of PSNR measured quality) by introduction of adaptive weights for the change of sensitive coefficients.

Many trials are not mentioned in this paper. Once in a while some hints of promising signs were observed, although they were not caused by the watermark and did not lead to the increase in quality that we hoped for.

Acknowledgements

The author congratulates the Team Craver for their achievement of engineering and winning the prize. The author thanks Rainer Böhme for his helpful comments on this paper as well as the protagonists who organised this motivating contest and therefore were only silent participants. Travel to the ACM Multimedia and Security Workshop was supported in part by the European Commission through the IST Programme under contract IST-2002-507932 ECRYPT.

REFERENCES

[1] Richard P. Brent. *Algorithms for Minimization without Derivatives*. Prentice Hall, Englewood Cliffs, NJ, 1973.

[2] Ingemar J. Cox and Jean-Paul M. G. Linnartz. Public watermarks and resistance to tampering. In *IEEE International Conference on Image Processing ICIP'97*, volume 3, pages 3–6, Santa Barbara, California, USA, January 1997.

[3] ECRYPT. BOWS, Break our watermarking system, 2006. Online available at `http://lci.det.unifi.it/BOWS`.

[4] ECRYPT. Network of excellence in cryptology, 2006. Online available at `http://www.ecrypt.eu.org`.

[5] R. Fletcher and C. M. Reeves. Function minimization by conjugate gradients. *Computer Journal*, 7:149–154, 1964.

[6] Jean-Paul M. G. Linnartz and Marten van Dijk. Analysis of the sensitivity attack against electronic watermarks in images. In David Aucsmith, editor, *Information Hiding (2nd International Workshop)*, LNCS 1525, pages 258–272, Berlin Heidelberg, 1998. Springer-Verlag.

[7] Matt L. Miller, Gwenaël J. Doërr, and Ingemar Cox. Applying informed coding and embedding to design a robust high-capacity watermark. *IEEE Trans. on Image Processing*, 13:792–807, 2004.

[8] Fernando Pérez-González, Carlos Mosquera, Marcos Alvarez, and Reginald Lagendijk. High-rate quantization data hiding robust to arbitrary linear filtering attacks. In Edward J. Delp and Ping W. Wong, editors, *Security, Steganography and Watermarking of Multimedia Contents VIII (Proc. of SPIE)*, San Jose, CA, January 2006.

[9] Fabien A. P. Petitcolas, Ross J. Anderson, and Markus G. Kuhn. Attacks on copyright marking systems. In David Aucsmith, editor, *Information Hiding (2nd International Workshop)*, LNCS 1525, pages 219–239, Berlin Heidelberg, 1998. Springer-Verlag.

[10] Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. on Image Processing*, 13:600–612, 2004.