A Regression-Based Restoration Technique for Automated Watermark Removal

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

ABSTRACT

The second Break our Watermarking System (BOWS-2) contest exposed a technology named "Broken Arrows" to attacks from all over the world. For a successful attack participants had to render the watermark unreadable to the online detector in three given images while preserving a minimum quality level of 20 dB PSNR. We applied a generic approach to remove the independent elements of the watermark from dependent media elements. Its core part does not need to interact with the online detector to remove the watermark. In a postprocessing step with ten detector calls the quality was increased by about 1 dB.

The first episode of the second Break Our Watermarking System (BOWS-2) contest was aiming to investigate when an image watermarking system can be broken by image processing operations. During this period, from July to October 2007, no information about the watermarking algorithm was provided and only 30 trials per day were allowed. We describe the attack that preserved the highest quality of the content.

Categories and Subject Descriptors

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

General Terms

Security

Keywords

Digital Watermarking, Robustness, Attack

1. INTRODUCTION

It is difficult to evaluate a digital watermarking system, since still many attacks are not predictable at the moment of evaluation during design. The European Network of Excellence ECRYPT has exposed a watermarking system to world-wide attacks in order to learn more about the degree of difficulty of breaking the watermark and perhaps finding new attacks. The second BOWS contest ("Break our Watermarking System") challenged a large number of attackers from all over the world, to break the watermarking system "Broken Arrows" (BA), which was specifically developed for BOWS-2 [8]. The contest is not intended to prove how well-performing BA is, but to have a better estimate how robust watermarking (in general) is in such a scenario [5].

In its first episode, running from mid July until mid October 2007, only ten out of several hundred participants (450 registered IP addresses) from all over the world maintained the required minimum quality (20 dB PSNR) in their attacked images.

The goal was to remove the invisible watermarks from three greyscale images (landscape motives with 512×512 pixels, high contrast and large share of high spatial frequencies, cf. Figure 1).

Figure 1: Images from the first episode of the contest: *fall*, *sheep*, and *souvenir*

This short paper is organised as follows: Section 2 briefly motivates the principle of the proposed attack, the structure of which is described in more detail in Sect. 3. Section 4 concludes the paper.

2. HINTS AND INSPIRATION

The marked images of the contest are given in PGM format (portable greymap). In the header of these image files, there is a comment: "Generated by libit." Watermarks are rarely added directly to the pixel values. For better robustness they are brought in a transformed space. In most cases this is the DCT or the wavelet domain [3]. After the DCT domain was already used in the first BOWS contest [4 10], wavelet was likely to be next in BOWS-2. The "libit" from the comment mentioned earlier, a library for signal processing, provides exactly one two-dimensional transformation—the wavelet transform. Of course this comment could have been a feint, however, the achieved results rather support the use of this library.

MM&Sec'08, September 22–23, 2008, Oxford, United Kingdom.

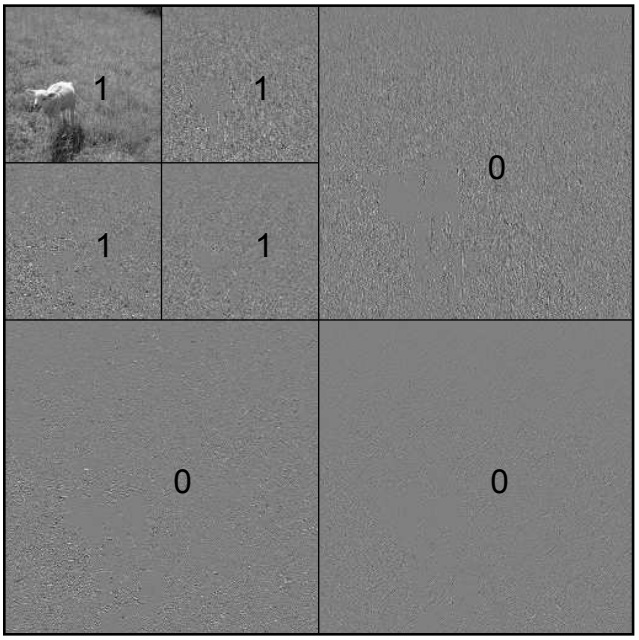

Figure 2: Wavelet coefficients for *sheep*

The wavelet transform halves the image resolution and stores the rest of information in one subband with horizontal (LH), vertical (HL), and diagonal wavelet coefficients (HH). This transformation can be recursively applied to the reduced image (cf. Figure 2). The watermark is not detected anymore if we set the coefficients in the first and second level of decomposition to zero (nullify LH0, HL0, HH0, LH1, HL1, HH1)[1]. The image appears smoothed then (cf. Figure 3). The required minimum quality of 20 dB is maintained for

Figure 3: Detail of *sheep* before (left) and after erasing two levels of decomposition (right), 21.41 dB

fall (20.11 dB) and *sheep* (21.41 dB), but not for *souvenir* (18.61 dB). These qualities indicate that *souvenir* has more spectral energy in the nullified subbands than *fall* and *fall* more than *sheep*. When we look at the results by Engel, we can find that he obviously applied the same attack on 9th of August 2007 with 20.11 dB, 21.41 dB, and (enhanced to be just successful) 20.01 dB (cf. Figure 4) [7]. Sometimes it is possible to recognise the performed attack by its achieved PSNR.

[1]This does not imply that all other coefficients have no impact on the watermark.

			average	fall	sheep	souvenir
08/09/2007	Dominik Engel	University of Salzburg	20.46 dB	20.11 dB	21.41 dB	20.01 dB

Figure 4: Entry in the Hall of Fame by Dominik Engel, the PSNR of which indicates nullifying two levels of decomposition

Although all three images have been watermarked with the same PSNR (about 42.5 dB), the participants agree on that *souvenir* was much harder to attack successfully.

On 30th August 2007 Bennour and Dugelay submitted a successful attack to *souvenir* with impressive 21.12 dB. This was the first rank for more than a month. Bennour et al. [1] used the self similarities attack [9] during the first BOWS contest (cf. Figure 5). Certainly, this attack is dependent on an image database that randomly limits the performance. Effectively, the image database is modelling dependencies of pixels in an image. In the hope of finding a well-performing replacement, we developed a mathematical model for dependencies of wavelet coefficients.

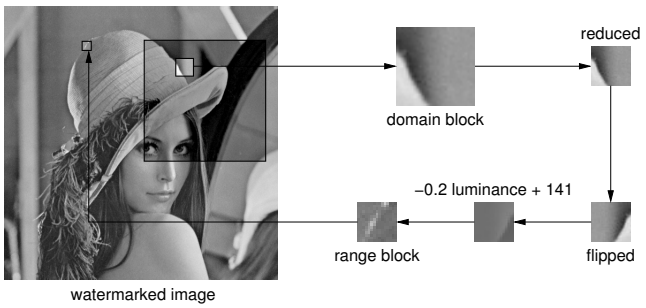

Figure 5: Self similarities attack after [9]

3. MODEL-BASED ATTACK TO WATERMARKS IN THE WAVELET DOMAIN

3.1 The Model

As guessed from the comment in the PGM files, we use the wavelet transform provided by libit. The source code comes with an example that applies the 2D wavelet transform (Daubechies 9/7) to images. Our images are 512×512 matrices X. Let W be the wavelet transform of X. We tried 2, 3, and 4 levels of decomposition. The attack achieves the best PSNR for 3 levels.

Our proposed attack follows the self similarities attack [9] in the figurative sense. Parts from the image are restored from the surrounding. Because locally close values in images strongly depend on each other, but the elements of the watermark do not, the image can be preserved by estimation from the surrounding while the watermark is completely removed (cf. Figure 6). Let n be the number of wavelet coefficients in a particular subband. We use simple linear regression to predict n wavelet coefficients w_s (where s is a "serial number" of the coefficient in its subband, $1 \leq s \leq n$) from k "neighbouring" coefficients $w_{s,1}, \ldots, w_{s,k}$:

$$w_s = \beta_1 w_{s,1} + \beta_2 w_{s,2} + \cdots + \beta_k w_{s,k} + \epsilon_s. \quad (1)$$

The number of terms k depends on the decomposition level

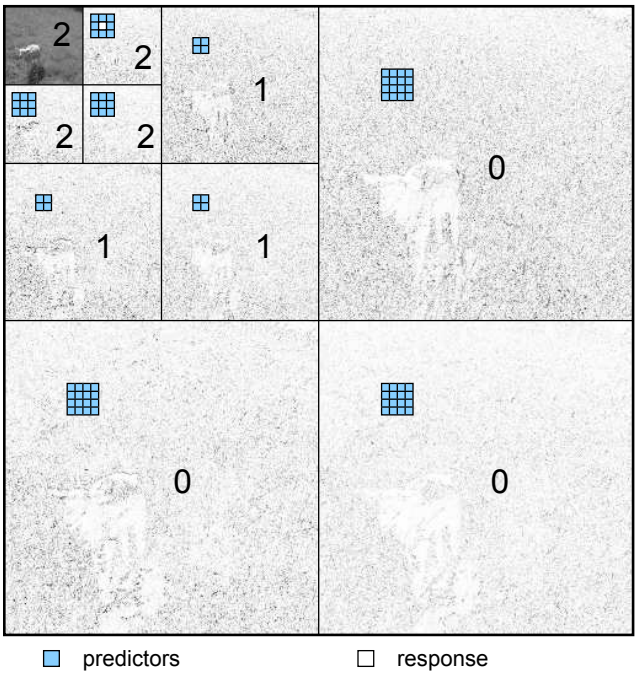

Figure 6: Model for estimating the absolute value of wavelet coefficients in LH2 from the surrounding

Table 1: Number of terms k for prediction

response in	number of terms from level 0	level 1	level 2	**total terms**
level 0	$3 \times 9 - 1$	0	0	**26**
level 1	3×4	$3 \times 9 - 1$	0	**38**
level 2	3×16	3×4	$3 \times 9 - 1$	**86**

that the coefficient belongs to (cf. Table 1). The regression model collects the local dependencies between the wavelet coefficients. We determine the predictor parameters $\hat{\beta}_1, \ldots, \hat{\beta}_n$ for which we find the minimum sum of squared residuals $\sum_{s=1}^{n} \epsilon_s^2$ (ordinary least squares). This condition is equivalent to the maximum PSNR, which is a logarithmic measure based on the mean squared error (MSE). In a first experiment, we predicted the signed values of the coefficients directly:

$$\hat{w}_s = \hat{\beta}_1 w_{s,1} + \hat{\beta}_2 w_{s,2} + \cdots + \hat{\beta}_k w_{s,k}. \quad (2)$$

We can restrict the model to estimate only the absolute value of the coefficients, while taking the signs from the marked original:

$$|w_s| = \beta'_1 |w_{s,1}| + \beta'_2 |w_{s,2}| + \cdots + \beta'_k |w_{s,k}| + \epsilon'_s, \quad (3)$$

$$\hat{w}_s = \mathrm{sign}(w_s) \cdot (\hat{\beta}'_1 |w_{s,1}| + \cdots + \hat{\beta}'_k |w_{s,k}|). \quad (4)$$

This leads to increased precision in the estimation and preserves a higher PSNR. Figure 6 marks a predicted coefficient in LH2 and the corresponding terms used for prediction. Every coefficient in LH2 is estimated from

- its direct neighbours in LH2,
- its counterpart in the subbands HL2 and HH2 together with their direct neighbours,
- its superior counterparts from the first and second level of decomposition (4 and 16 per subband, respectively).

One of the key properties of this attack is its non-interactivity. The attacked images are produced without using the detector. All computations can be done locally on the attacker's side.

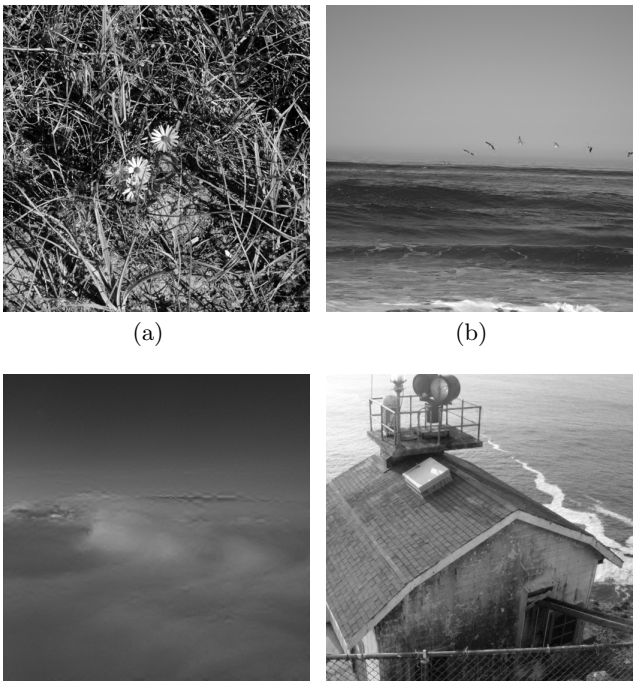

Figure 7: The regression-based attack is likely to succeed if the resulting PSNR is between the minimum 19.7 dB (a) and about 35 dB (b), and likely to fail above these 35 dB to maximum PSNR 45.0 dB (c). The median PSNR is 30.4 dB (d)

3.2 Postprocessing

If we reduce the changes made by the attack towards the detection boundary, we can improve the quality by about 1 dB PSNR.

weakened image = p·attacked image+$(1-p)$·marked original

By interval bisection the optimum parameter $p \in [0, 1]$ is found in about 10 detector uses. The resulting quality for this weakened attack is listed in Table 2. Figure 8 shows a detail of the attacked *sheep* resulting from regression-based attack with postprocessing.

Table 2: Quality (PSNR) of attacks before and after weakening

	after estimation	after weakening
average	**22.87 dB**	**24.30 dB**
fall	22.82 dB	24.90 dB
sheep	24.60 dB	25.45 dB
souvenir	21.67 dB	22.97 dB

Figure 8: Detail of *sheep* before (left) and after regression-based attack with postprocessing (right), achieved quality: 25.45 dB PSNR

We restricted the attack to two levels of wavelet decomposition as well as extended it to four levels. After weakening, the attack produced the highest PSNR with three levels, which are obviously used for the watermark.

3.3 Evaluation of the Attack

We applied the attack to 10,000 marked images downloaded from [6]. The resulting PSNR ranged from 19.7 to 45.0 dB (cf. Figures 7 and 9).

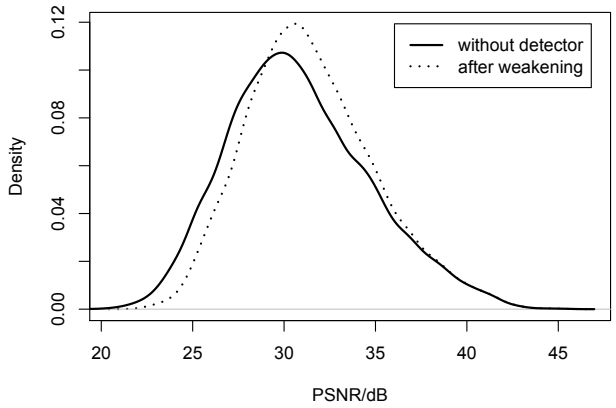

Figure 9: Estimated density curve of PSNR for attacked images from the BOWS-2 database (solid curve) and after postprocessing (dotted curve)

The attack is succesful for 85.4 % of the BOWS-2 database. Figure 10 shows the dependency of the PSNR and the detector value, which is compared to the detector threshold. The randomness of this threshold has almost no influence to the results presented here. We used its minimum for best reproducibility. If the quality of the resulting image of the regression-based restoration is about 35 dB and above, the attack is likely to fail. Figure 11 shows the decreasing percentage of successfully broken images for increasing PSNR (solid curve). Above 30 dB PSNR we can still break more than 40 % of the images. After postprocessing, the quality is slightly increased and we break almost 49 % of the images above 30 dB PSNR.

Due to the pixeloriented quality measure (PSNR), geometric attacks are at a disadvantage. However, the images

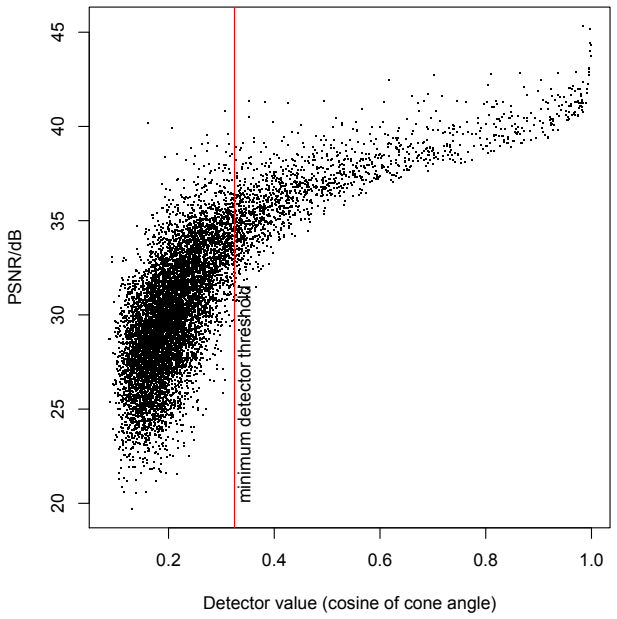

Figure 10: Dependency between resulting quality of the regression-based attack and detector response

where the regression-based attack fails, e.g., like the one in Figure 7(c), simply do not provide enough image content to extract a suitable model from. Such images are more successfully attacked by geometric attacks, because the penalty of the PSNR measure is only small for smooth areas. So we can hope for a synergy between the two types of attack under the PSNR measure: A geometric attack should be used for images that the proposed attack cannot handle. If the regression-based attack is applied to images with almost no content, the watermark is not overridden but becomes part of the model, due to overfitting. The cleanest way to solve this problem would be to take the response vector of the model from the unmarked original that is not available to the attacker in the considered security model.

We tried a very simple geometric attack that shifts the image to the right by three pixels. Figure 11 presents the poor performance of this attack (dashed curve). However, when combined with the proposed attack we can see a synergy of both (dotdash curve). For the decisive images in Episode 1 of the BOWS-2 challenge (cf. Figure 1) the proposed attack worked immediately.

4. CONCLUSION

The proposed attack removes the watermark by estimating the image from its locally dependent parts while the elements of the watermark are removed because they are independent of each other. It was ranked first by PSNR at the end of Episode 1 of the contest. The proposed attack (with postprocessing) was repeated on the very first day of Episode 3. It achieved on average 27.16 dB PSNR with only 10 detector uses in total for all three images together.

The attack can be prevented, if the watermark shows similar dependencies as the image itself. In the case of BA, one element of the watermark has to be added not only to one coefficient, but at a smaller scale also to all dependent coeffi-

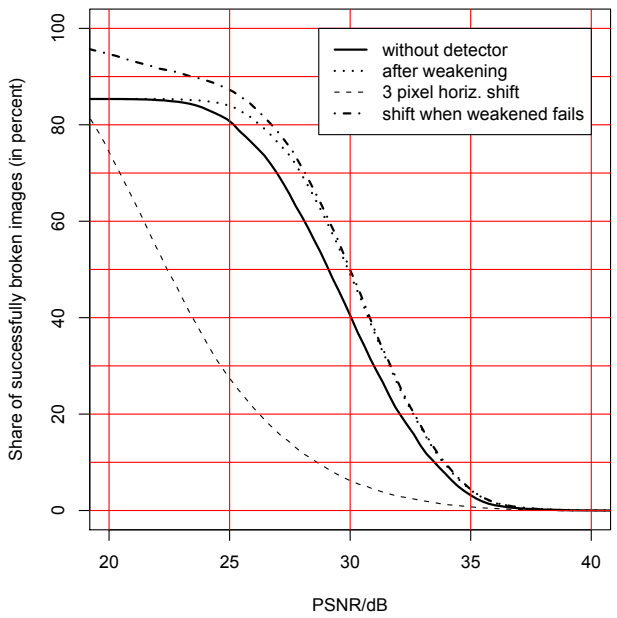

Figure 11: Operating curve of the estimated images from the BOWS-2 database

cients from the surrounding. The attack is similarly difficult when applied directly to the pixels. Like a single wavelet coefficient has impact on several pixels, so one single element of the watermark would be distributed on several pixels. An estimation from neighbouring pixels will not succeed because the neighbourhood proportionately contains the element of the watermark under removal. As a necessary prerequisite we have to guess the feature space of the watermark. The proposed attack will not work with any other space, where the elements of the watermark are not separable. After the first BOWS contest, Craver et al. have already proposed not to embed the watermark in a stereotypical feature space (like DCT or wavelet coefficients) [2]. Without knowledge of the key it should be impossible to separate the elements of the watermark.

ACKNOWLEDGEMENT

The author thanks Patrick Bas and Teddy Furon, who organised this motivating contest in a very entertaining way.

REFERENCES

[1] Jihane Bennour, Jean-Luc Dugelay, and Frederico Matta. Watermarking attack (BOWS contest). In Edward J. Delp III and Ping Wah Wong, editors, *Security, Steganography and Watermarking of Multimedia Contents IX (Proc. of SPIE)*, pages 18–1–18–6, San Jose, CA, January 2007.

[2] Scott Craver, Idris Atakli, and Jun Yu. How we broke the BOWS watermark. In Edward J. Delp III and Ping Wah Wong, editors, *Security, Steganography and Watermarking of Multimedia Contents IX (Proc. of SPIE)*, pages 1C–1–1C–8, San Jose, CA, January 2007.

[3] Scott Craver and Jun Yu. Reverse-engineering a detector with false alarms. In Edward J. Delp III and Ping Wah Wong, editors, *Security, Steganography and Watermarking of Multimedia Contents IX (Proc. of SPIE)*, pages 0C–1–0C–10, San Jose, CA, January 2007.

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

[5] ECRYPT. BOWS-2, Break our watermarking system, 2nd edition, 2007. Online available at `http://bows2.gipsa-lab.inpg.fr`.

[6] ECRYPT. BOWS-2 database of 10,000 watermarked images, 2008. Online available at `http://bows2.gipsa-lab.inpg.fr/BOWS2ImageDataBase.tgz`.

[7] ECRYPT. BOWS-2, results of the episodes, 2008. Online available at `http://bows2.gipsa-lab.inpg.fr/index.php?mode=VIEW&tmpl=resPrevEp`.

[8] Teddy Furon and Patrick Bas. Broken arrows (draft), 2008. Online available at `http://bows2.gipsa-lab.inpg.fr/BrokenArrows.pdf`.

[9] Christian Rey, Gwenaël Doërr, Gabriella Csurka, and Jean-Luc Dugelay. Toward generic image dewatermarking? In *IEEE International Conference on Image Processing ICIP 2002*, volume 2, pages 633–636, New York, NY, USA, September 2002.

[10] Andreas Westfeld. Lessons from the BOWS contest. In *Proc. of ACM Multimedia and Security Workshop 2006, MM&Sec06, Geneva, Switzerland*, pages 208–213, New York, September, 26–27 2006. ACM Press.