Hindawi Publishing Corporation
EURASIP Journal on Information Security
Volume 2007, Article ID 25308, 8 pages
doi:10.1155/2007/25308

Research Article

Breaking the BOWS Watermarking System: Key Guessing and Sensitivity Attacks

Pedro Comesaña and Fernando Pérez-González

Signal Theory and Communications Department, University of Vigo, 36310 Vigo, Spain

Correspondence should be addressed to Pedro Comesaña, pcomesan@gts.tsc.uvigo.es

Received 11 May 2007; Accepted 31 August 2007

Recommended by A. Piva

From December 15, 2005 to June 15, 2006, the watermarking community was challenged to remove the watermark from 3 different 512×512 watermarked images while maximizing the *peak signal-to-noise ratio* (PSNR) measured by comparing the watermarked signals with their attacked counterparts. This challenge, which bore the inviting name of *Break Our Watermarking System* (BOWS), had as its main objective to enlarge the current knowledge on attacks to watermarking systems. In this paper, the main results obtained by the authors when attacking the BOWS system are presented and compared with strategies followed by other groups. Essentially, two different approaches have been followed: *exhaustive search of the secret key* and *blind sensitivity attacks*.

1. INTRODUCTION

In the last decade, with the spreading of the Internet as an impressive communication tool and the appearance of advanced editing tools which can be used by almost any non-trained user, the need of new technical solutions to the problems of copyright protection, authentication, fingerprinting, or annotation of digital contents has soared. Digital watermarking has been widely recognized as a potentially powerful instrument against piracy, illegal modification, or improper use of contents. Nevertheless, experience has shown that the challenge of designing a watermarking method robust against an active attacker is extremely difficult. Even without considering geometrical attacks (which can be regarded as some of the most harmful attacks against watermarking techniques), the range of strategies an attacker could envisage to remove the watermark from a watermarked content is virtually as diverse as the attackers themselves.

We believe that challenging the watermarking community (and the public in general) to break a certain watermarking system is valuable for a number of reasons: (1) the contest serves to pinpoint the weaknesses of state-of-the-art methods, and likely, promote new research aimed at improving those methods; (2) the inherent applicability of the attacks serves as a benchmark to test results developed under more theoretical conditions; (3) the existence of independent attackers acts in a way as a "Monte Carlo" testing of the algorithms.

The design of *good (and new) attacks* is one of the main motivations of the *Break Our Watermarking System* (BOWS) challenge. This *contest* consisted in removing the watermark from three watermarked signals, trying to maximize the peak signal-to-noise ratio (PSNR), a squared error distortion measure. (Remember that for 8-bits signals, $\mathrm{PSNR}(\mathbf{x}, \mathbf{y}) \triangleq 10 \log_{10}[L \cdot 255^2 / \|\mathbf{x} - \mathbf{y}\|^2]$, where L is the length of the compared signals.) The fact of choosing a mean square error (MSE) measure could be criticized, as it does not really reflect the impact of the attack on the semantics of the signal, but the lack of a universally recognized perceptual distortion measure also makes difficult the choice of a non-MSE-based measure. In addition, the use of an MSE measure is supposed to leave out geometrical attacks for removing the watermark, as the resulting MSE is typically believed to be quite high; nevertheless, Andreas Westfeld showed in [1] that the watermark can be removed using geometrical attacks (in that case, rotation) achieving PSNRs as high as 28.94, 22.9, and 24.99 dBs, respectively. Geometrical attacks, whose perceptual impact may be quite reduced, are known to be extremely harmful to the performance of most watermarking methods, as they often cause a loss in the synchronization of the watermark.

Given the interest of the organizers of BOWS in investigating whether the knowledge of the watermarking scheme could be useful for devising a better attacking strategy, BOWS was divided into two different stages: for the first three months, just three 512×512 watermarked images (see Figure 1) and three binary detectors (one per image) were provided; later, and for the next three months, the watermarking method was made public. Moreover, at the very beginning the number of calls to the detector per day was limited, trying to avoid oracle attacks.

Within this framework, we tried to remove the watermark from the provided images in two different circumstances.

(1) The attacker completely lacks any knowledge of the used watermarking method and only has access to a detector, that he feeds with an image, and provides a binary output. This situation corresponds to the first stage of the BOWS challenge.
(2) The attacker knows all the details about the watermarking scheme, except for a secret parameter, the *secret key*, which is only shared by embedder and detector.

For the first case, we used the blind sensitivity attack previously published in [6], whereas for the second one, we followed a strategy based on a exhaustive search on secret key space.

Nevertheless, due to the large number of calls to the detector that would be needed for performing a successful blind sensitivity attack, we decided to attack the system in the reverse order of that assumed by the organization. Thus, we first attacked BOWS trying to learn the secret key to later perform the blind sensitivity attack. Such blindness implied that information regarding the actual watermarking scheme, such as the embedding domain, the used DCT coefficients, or the perceptual shaping of the watermark, was totally disregarded. Our own objective here was to show the feasibility of attacking a watermarking system without knowing any a priori information about it. Needless to say, this assumption could be considered as quite pessimistic for the attacker, for most of this information could be actually learned by a smart attacker, as it was practically shown by other participants in the BOWS contest [10–12].

Detailed information on our two approaches to breaking BOWS is given in Sections 2 and 3, respectively; approaches followed by other groups are outlined in Section 4, and the obtained results are compared in Section 5.

2. GUESSING THE SECRET KEY

Given that one of us was member of the steering committee, we did not directly participate in the contest. It is important to remark, however, that our knowledge about the watermarking scheme was exactly the same as that publicly available during the second stage of the contest: the used algorithm was the well-known data-hiding method by Miller et al. [2]. This means that we did not use any additional information on the algorithm implementation (including the version number). Interestingly, other groups were able to get access to the same details through their social engineering (or coercive) strategies [12]. Miller et al.'s side-informed method is based on the use of trellis codes for performing the source and channel coding, being parameterized by some of the properties of the used trellis (number of states and number of arcs per state), as well as the spreading factor.

Still one open question for the attacker (and for us) was to figure out how the original data-hiding scheme had been adapted to this particular application, since Miller et al.'s algorithm was intended to work in decoding (i.e., multiple bit watermarking) scenarios instead of detection (i.e., one-bit watermarking) ones. We thought that probably the most straightforward way to carry out this adaptation would be to compare the decoded message with a secret reference (which is the actual embedded message); if they were identical, the watermark could be said to be present in the received signal, and absent otherwise. In order to test our conjecture on the BOWS system available at [3] and decide if the three provided images were watermarked with both the same key and the same reference message, we input those images to the detectors corresponding to the other two images. As a result, although very small PSNRs were obtained, the watermark could still be detected, thus confirming our intuition. Notice that at this stage, the three available images were the only inputs to the webpage detectors.

After the contest had started, we became aware of the existence of a new version of Miller et al.'s algorithm's code [2], publicly available at the webpage of one of the authors [4]. In this implementation, some values for the aforementioned parameters are taken by default, in such a way that the decoder is only parameterized by the image to be checked $\mathbf{z}$, the secret key θ, and the length (in bytes) M of the message to be decoded; this last point represents an additional difficulty to the attackers' task, as they must also consider the different choices for that parameter. Despite these obstacles, and without certainly knowing if that was the version of the algorithm used in BOWS, we decided to peruse this implementation with the two-fold objective of better learning how it actually worked, and performing a security attack (here meaning an attack trying to gain knowledge about the chosen secret key).

Therefore, if we were able to find a pair (θ, M), such that the outputs of the decoding function provided by the authors of that trellis-based algorithm [2] were the same for the three given images, we could be fairly confident that the used secret key was θ, and that the algorithm implementation was the publicly available one [4]. Given that no a priori information about the value of the secret key was available, we decided to try the exhaustive search mechanism. Furthermore, we had to establish a range of possible values for M, in order to try an exhaustive search for each possible value of the secret key. Taking into account that the studied scheme was really robust against most signal processing operations, it was clear that its rate (i.e., the inverse of the number of used coefficients per embedded bit) should be small enough. Considering that the number of coefficients of the provided images used by the algorithm for embedding information is $12 \cdot 512 \cdot 512/64 = 49152$, we decided that it was unlikely that the number of hidden bits was larger than 200, that is, $M < 25$ bytes, since otherwise the number of coefficients per

(a) (b) (c)

Figure 1: The three watermarked signals provided by the organization.

embedded bit should be smaller than 240, a choice which would not afford much robustness against conventional signal processing attacks. In view of these considerations, we decided to implement the exhaustive search represented in Figure 2. After 9 days of computation of a PERL script verifying if the decoded messages for different secret keys and message lengths were the same for all the three images, run on a shared computer with 2 Intel Pentium Xeon processors at 3.06 GHz and 2 GB RAM (be aware that the time requirement could be lightened if this process was paralleled), the authors found a pair of values $(\theta, M) = (7168, 5)$ verifying the condition introduced above, that is, the decoded message for all the three images was the same ('BOWS*').

Once the secrets of the system have been completely learned, the attacker has to devise how this information can be used to produce signals as close as possible to the provided watermarked ones, but where the watermark has been removed, that is, where the decoded message is no longer the reference one. The optimal way of doing so is to perform a search over the trellis looking for the boundary to the codeword related to a message different to the embedded one (BOWS*') which is closer to the provided image. Partial results on this optimization were presented by Andreas Westfeld in the *Security, Steganography and Watermarking of Multimedia Contents IX Conference* celebrated in January 2007 in San Jose (Calif, USA).[1] In our case, we followed a simpler but nonoptimal strategy: we watermarked the provided signals with the same key θ, but with a different message of length M. In fact, taking into account the trellis nature of the used code, we only changed the last bit of the reference message to 'BOWS+,' assuming that in that way the distance between the originally embedded codeword and the newly obtained one would be close to the minimum; this reasoning is based on the trellis structure of the codebook. Nevertheless, be aware that this strategy is not necessarily the optimal one, due to both the heuristic nature of the embedding algorithm, and the fact that there could be a codeword associated to a message different from BOWS+' that is closer to the original codeword. In any case, computing the new signal in the described way, and considering the linear convex combinations of this signal and the provided one, we were able to produce signals that were really close to the latter, but where the watermark had been removed; this procedure is illustrated in Figure 3. The PSNRs obtained for the three proposed images are, respectively, 53.5051 dB, 56.1106 dB, and 56.9275 dB.

[1] Andreas Westfeld obtained the secret key from the organizers after the two stages of the contest ended.

3. BLIND SENSITIVITY ATTACKS

Once both the secret key and the embedded message were correctly guessed, we were able to perform in our local computers as many calls to the detector as we wished, skipping the initial constraint on the number of calls imposed by BOWS rules, and circumventing the communication delays with the server where the detectors were hosted.

As we previously mentioned, by performing the sensitivity attack we tried to show that an attacker without any knowledge of the watermarking scheme (not even any intention to gain such knowledge) would be able to obtain very good results, by just applying a method already presented by the authors in [6] and termed *Blind Newton Sensitivity Attack* (BNSA). For this reason, although this method could be modified to reduce the number of calls to the detector [5], we decided to use the basic version described in [6]. Furthermore, note that given that our algorithm assumes no prior knowledge of the watermarking technique, it complies with the rules established for the first stage of BOWS.

The chosen algorithm, which has shown to be effective against a wide range of watermarking methods, has the following characteristics that make it suitable for the BOWS setup:

(i) it does not require any knowledge about the detection function;

(ii) it just needs to know the binary output of the detection function for a given input (not the actual value of the detection function);

(iii) it can be highly paralleled, in such a way that several attackers can work together, each of them using a different detector (or a set of them). In any case, be aware that this characteristic could not be applied in the BOWS setup, as only one detector was available;

(iv) a full iteration of the algorithm is not necessary for obtaining relatively good results.

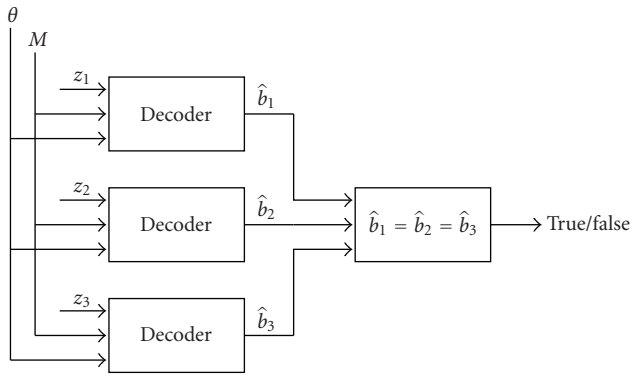

Figure 2: Block diagram of the exhaustive search of the secret key approach. We checked $1 \leq M \leq 25$ for increasing values of $\theta \in \mathbb{N}$, until $\hat{b}_1 = \hat{b}_2 = \hat{b}_3$. The source code of the decoder is publicly available in [4].

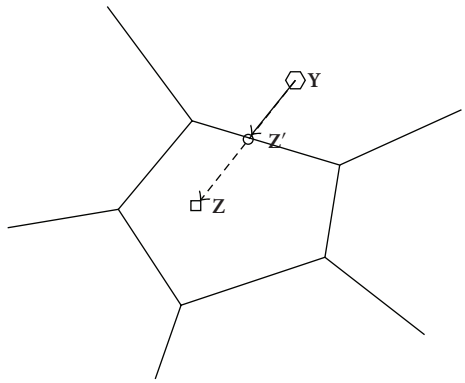

Figure 3: Diagram describing the generation of the attacked signal $\mathbf{Z}'$ from the watermarked signal $\mathbf{Y}$, and the signal $\mathbf{Z}$ obtained by embedding a message different of the reference one into the watermarked content.

Furthermore, as we will explain later, it is not required to compute exactly the Hessian matrix nor the gradient in order to be able to produce a descent direction of the target function. In fact, we will show that some gradient components are enough to obtain high-quality signals where the watermark is already removed. A basic block diagram of BNSA, depicting its iterative nature, is plotted in Figure 5.

3.1. Blind Newton Sensitivity Attack

The BNSA [6] is based on formalizing the target of the attacker as

$$\arg \min_{\mathbf{t}: g(\mathbf{y}+\mathbf{t}) \leq \eta} d_{\mathbf{y}}(\mathbf{t}), \tag{1}$$

where $d_{\mathbf{y}}(\mathbf{t})$ quantifies the distortion introduced by the attacking vector $\mathbf{t}$ on the watermarked signal $\mathbf{y}$, $g(\cdot)$ is the detection function, and η is a threshold which determines the detection region, in such a way that the detector will decide that the watermark is present if and only if $g(\mathbf{y}) > \eta$. Given that in our problem the objective is to maximize the PSNR, or equivalently to minimize the MSE of the distorting vector, we can see that in this particular scenario $d_{\mathbf{y}}(\mathbf{t}) = \|\mathbf{t}\|^2$.

Therefore, the BNSA tries to iteratively solve this problem by using a surjection $h_{\mathbf{y}}$, that projects the attacking vectors on the decision boundary and should verify some specific characteristics (see [6] for further information about these particulars), yielding an update of the algorithm with the generic form

$$\mathbf{s}_{k+1} = \mathbf{s}_k - \xi_k \cdot \mathbf{B}^{-1} \cdot \hat{\nabla} \left(d_{\mathbf{y}}^{\star} \circ h_{\mathbf{y}}\right)(\mathbf{s}_k), \tag{2}$$

where ξ_k is the stepsize of the update, $d_{\mathbf{y}}^{\star}(\cdot)$ is the constraint of $d_{\mathbf{y}}(\cdot)$ to the points on the boundary of the decision region, $\hat{\nabla}(d_{\mathbf{y}}^{\star} \circ h_{\mathbf{y}})(\mathbf{s}_k)$ is the estimate of the gradient of $(d_{\mathbf{y}}^{\star} \circ h_{\mathbf{y}})(\mathbf{s}_k)$, and its ith component is computed as

$$\left[\hat{\nabla}\left(d_{\mathbf{y}}^{\star} \circ h_{\mathbf{y}}\right)(\mathbf{s}_k)\right]_i = \frac{\left(d_{\mathbf{y}}^{\star} \circ h_{\mathbf{y}}\right)(\mathbf{s}_k + \delta \mathbf{e}_i) - \left(d_{\mathbf{y}}^{\star} \circ h_{\mathbf{y}}\right)(\mathbf{s}_k)}{\delta}, \tag{3}$$

with $\delta > 0$ an arbitrarily small positive number. This computation requires also to estimate $h_{\mathbf{y}}(\cdot)$, which is not known by the attacker. The proposed strategy is to use $h_{\mathbf{y}}(\mathbf{s}) = \alpha^* \cdot \mathbf{s}$, where α^* is a scaling factor computed using a bisection algorithm.

Concerning the matrix $\mathbf{B}$, different possibilities can be considered; probably the best choice would be to use an approximation of the Hessian, but due to computing limitations, we have preferred to use $\mathbf{B} = \mathbf{I}_{L \times L}$, that is, the identity matrix of size L. For a small-enough ξ_k, this choice of $\mathbf{B}$ guarantees a decrease of the target function as long as $\left(\hat{\nabla}(d_{\mathbf{y}}^{\star} \circ h_{\mathbf{y}})(\mathbf{s}_k)\right)^T \cdot \nabla(d_{\mathbf{y}}^{\star} \circ h_{\mathbf{y}})(\mathbf{s}_k) > 0$, where the last condition is based on the Taylor series expansion of the objective function. In order to verify this condition, the attacker does not need to compute all the components of the estimate of the gradient; instead it is enough to calculate a small number of them and set the remaining to 0. Obviously, better results will be obtained when all the components are available, but the former strategy allows the attacker to stop the algorithm whenever he has obtained a suboptimal solution which yields a satisfactory (following his quality criterion) attacked image; of course, adopting such strategy will reduce the computational cost of his attack.

Finally, the computation of the step length [7, 8] of the kth iteration ξ_k was performed following Armijo's rule, due to its simplicity.

3.2. Pseudocode

For the sake of clarity, next we give a pseudocode description of the used implementation of the BNSA; this implementation slightly differs with the later one presented in [9], and that was debugged after close interaction with Prof. Barni's group in the University of Siena. In the following description we assume that the provided watermarked signal $\mathbf{y}$ is arranged as a vector.

(1) Generate an i.i.d. zero-mean Gaussian random vector $\mathbf{t}$, with the same size as the watermarked signal $\mathbf{y}$ and variance $\sigma_T^2 = 10^{-4}$.

(2) Compute a scaling factor β such that $\mathbf{y} + \beta \cdot \mathbf{t}$ is on the detection boundary. The squared Euclidean norm of the vector $\beta \cdot \mathbf{t}$ is denoted by γ_{start}.
(3) For each component of the vector $\mathbf{t}$,
  (a) slightly modify the vector $\mathbf{t}$, obtaining $\mathbf{t}_i = \mathbf{t} + \epsilon_1 \cdot \mathbf{e}_i$, where $\mathbf{e}_i$ is the ith vector of the canonical basis and $\epsilon_1 = 10^{-3}$.
  (b) compute a scaling factor β such that $\mathbf{y} + \beta \cdot \mathbf{t}_i$ is on the detection boundary. The squared Euclidean norm of the vector $\beta \cdot \mathbf{t}_i$ is denoted by γ_i.
(4) Estimate the gradient of the squared Euclidean norm of the vector necessary for obtaining a nonwatermarked signal, when the vector $\mathbf{t}$ is considered. The ith component of the gradient is estimated as $\hat{\nabla}[i] = (\gamma_i - \gamma_{\text{start}})/\epsilon_1$.
(5) Look for a step size providing a decrease in the objective function as follows
  (a) $\xi = 10$.
  (b) $\mathbf{t}_{\text{new}} = \mathbf{t} - \xi \cdot \hat{\nabla}$.
  (c) Compute a scaling factor β such that $\mathbf{y} + \beta \cdot \mathbf{t}_{\text{new}}$ is on the detection boundary. The squared Euclidean norm of the vector $\beta \cdot \mathbf{t}_{\text{new}}$ is denoted by γ_{after}.
  (d) While $\gamma_{\text{start}} < \gamma_{\text{after}}$,
    (i) $\xi = 0.7 \cdot \xi$;
    (ii) $\mathbf{t}_{\text{new}} = \mathbf{t} - \xi \cdot \hat{\nabla}$;
    (iii) Compute a scaling factor β such that $\mathbf{y} + \beta \cdot \mathbf{t}_{\text{new}}$ is on the detection boundary. The squared Euclidean norm of the vector $\beta \cdot \mathbf{t}_{\text{new}}$ is denoted by γ_{after}.
(6) If the resulting signal $\mathbf{y}_2 = \mathbf{y} + \beta \cdot \mathbf{t}_{\text{new}}$ verifies the quality criteria established by the attacker, then $\mathbf{y}_2$ is the solution. Otherwise, the algorithm is iterated again from point 2 with $\mathbf{t} = \mathbf{t}_{\text{new}}$.

3.2.1. *Computation of a scaling factor β such that $\mathbf{y} + \beta \cdot \mathbf{t}_0$ is on the detection boundary*

(1) $\mathbf{t} = \mathbf{t}_0$.
(2) If $\mathbf{y} + \mathbf{t}$ is out of the detection region, then $\mathbf{v}_{\text{out}} = \mathbf{t}$ and $\mathbf{v}_{\text{in}} = \mathbf{0}$. Otherwise, if $\mathbf{y} - \mathbf{t}$ is out of the detection region, then $\mathbf{v}_{\text{out}} = -\mathbf{t}$ and $\mathbf{v}_{\text{in}} = \mathbf{0}$.
(3) If both $\mathbf{y} + \mathbf{t}$ and $\mathbf{y} - \mathbf{t}$ are in the detection region,
  (a) while both $\mathbf{y} + \mathbf{t}$ and $\mathbf{y} - \mathbf{t}$ are in the detection region, $\mathbf{t} = 2 \cdot \mathbf{t}$.
  (b) If $\mathbf{y} + \mathbf{t}$ is out of the detection region, then $\mathbf{v}_{\text{out}} = \mathbf{t}$ and $\mathbf{v}_{\text{in}} = \mathbf{t}/2$. Otherwise, if $\mathbf{y} - \mathbf{t}$ is out of the detection region, then $\mathbf{v}_{\text{out}} = -\mathbf{t}$ and $\mathbf{v}_{\text{in}} = -\mathbf{t}/2$.
(4) While $\|\mathbf{v}_{\text{out}} - \mathbf{v}_{\text{in}}\| > \epsilon_2 (= 10^{-3})$,
  (a) $\mathbf{v}_{\text{middle}} = (\mathbf{v}_{\text{out}} + \mathbf{v}_{\text{in}})/2$.
  (b) If $\mathbf{y} + \mathbf{v}_{\text{middle}}$ is in the detection region, then $\mathbf{v}_{\text{in}} = \mathbf{v}_{\text{middle}}$; otherwise, $\mathbf{v}_{\text{out}} = \mathbf{v}_{\text{middle}}$.
(5) $\mathbf{v} = \mathbf{v}_{\text{out}}$.
(6) The scaling factor β such that $\mathbf{y} + \beta \cdot \mathbf{t}_0$ is nonwatermarked is given by the ratio between the value of any component of vectors $\mathbf{v}$ and $\mathbf{t}_0$, that is, $\beta = \mathbf{v}[i]/\mathbf{t}_0[i]$, which is the same for any component.

3.3. Results

After one iteration of the BNSA *performed in the pixel domain* of the three provided images, the PSNR obtained for the first image is 56.3410 dB, for the second 56.9559 dB, and for the third one is 58.1586 dB, and the resulting images are plotted in Figure 6. Nevertheless, as pointed out by a reviewer, when those images are captured from the document in pdf format and the resulting image is fed to the detector, the watermark is still present (in the tests we performed, this was the case for images 1 and 3). Although the obtained result will obviously depend on the sequence of operations performed for capturing the images (e.g., type of the file images inserted in the paper, conversion of file types, etc.), the fact that we had been able to reproduce these striking results obtained by the reviewer led us to try to find a plausible explanation. In trying to do this, we considered the energy of the three attacking signals for each 8×8 block-DCT frequency, as well as the ratio between the energy of the captured image and the energy of the watermarked one for each frequency. In that way we realized that whereas the power of the attacking signal is approximately constant for the frequencies considered by the detector algorithm, the capturing process can be modeled as low-pass. This is not rare, if one considers the small size of the images in the paper, which made one suspicious about the fact that the image could have undergone a process of down-sampling and interpolation. Notice that this could imply that the filtering undergone by the image is affecting more the attacking signal than the watermarked one.

This reasoning would also explain why this phenomenon is not observed for the images attacked after the exhaustive search of the secret key (Figure 4). In order to provide a further argument supporting our hypothesis, we performed the Wiener filtering of the three images attacked by BNSA, and in all three cases, the watermark was recovered.

Furthermore, we think that is particularly interesting the fact that acceptably good results are achieved even when a full iteration of the BNSA is not completed. To illustrate, Figure 7 shows the PSNR achieved versus the number of actually computed gradient components. One can see that the PSNR obtained for 4096 computed coefficients of the gradient is already larger than 40 dB for all the three images, meaning that a quite reduced amount of computations would be required for obtaining high-quality contents where the watermark is removed. Furthermore, be aware that when the number of considered gradient components is large enough (for values larger than approximately 2048), the PSNR grows almost linearly with the number of components, meaning that the reduction in the attacking distortion (in dB) grows linearly with the logarithm of the number of available components of the gradient.

Finally, although we have performed the BNSA in the pixel domain, a smart attacker could suspect that the images

(a) (b) (c)

Figure 4: The three signals obtained using the attack based on the exhaustive search of the secret key.

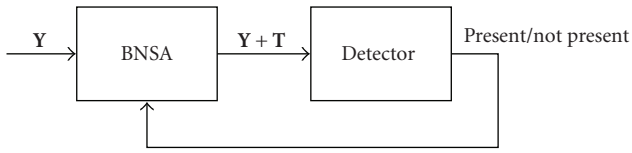

Figure 5: Block diagram of the BNSA.

were not watermarked in all their frequency components, but just in a subset of them, as most of the algorithms in the literature use that kind of strategy. Indeed, after performing some tests, the attacker could realize that the watermark was only embedded in 12 out of the 64 coefficients of an 8×8 block-DCT [10–12]. Using that information, the attacker could speed up the attack more than 5 times, by just focusing on estimating the gradient of those block DCT coefficients that he knows (or assumes) that they are being used by the watermarking method. Even though we did not pursue this line, we have determined the PSNR needed for removing the watermark from the provided images using the attacking vectors already computed with the BNSA in the spatial domain, and setting to 0 those 52 8×8-DCT components which are not used in each block. Remarkably, the results only show a small improvement on the obtained PSNR, yielding values of 57.5496, 57.8056, and 60.0081 dB, respectively, indicating that BNSA already succeeded in allocating almost all the power in those DCT coefficients that were really used, even when it was performed in the pixel domain.

4. APPROACHES FOLLOWED BY OTHER GROUPS

In this section, we summarize the strategies followed by some other groups involved in BOWS contest, discussing the similarities and differences with our approach. The first three strategies were applied during the first stage, whereas the fourth one corresponds to the second stage, where the watermarking algorithm was disclosed.

4.1. Le Guelvouit et al. [12]

In this paper, the authors claim to have used their social engineering skills (very remarkable indeed) to learn the watermarking scheme that was actually used. Once in possession of such information, several attacks are explored as follows:

(i) using a different version of the third image which was previously published, they managed to remove the watermark from that image with a PSNR of 37.35 dB;
(ii) a new attack, based on setting to zero some DCT coefficients and randomly shuffling other ones, was proposed, yielding respective PSNRs of 30.39 dB, 31.21 dB, and 30.22 dB;
(iii) using an adaptation of a previous work [13] on data hiding game theory, they compute the SAWGN (Scaling and Additive White Gaussian Noise) attack minimizing the capacity of a spread spectrum-based system, and apply it to the provided watermarked images to remove the watermark. After taking into account the perceptual considerations made in [2], they reached respective PSNRs of 34.95 dB and 34.04 for the first two images.

4.2. Earl [5]

Earl proposed a sensitivity analysis algorithm similar to that in [6], with some modifications aimed at speeding-up convergence. According to Earl, the main differences between this new scheme and BNSA lie in

(i) how and when the surjection function is computed;
(ii) the approximate nature of the minimization;
(iii) the selection of a perceptually scaled basis on which to search.

In fact, in view of the presented results, it seems that the perceptual considerations have a major role in the reduction of the number of calls to the detector needed by the proposed method. Following this strategy, the author achieved respective PSNRs of around 38 dB, 35 dB, and 36 dB with a number of detector calls slightly larger than 1500.

4.3. Craver et al. [11]

Craver et al. followed a different approach for getting information about the used watermarking scheme. They determined the frequency transform and the subbands by decreasing the PSNR as much as possible while keeping the watermark still detectable. This information allowed them to develop an attack which removes the watermark from the given image by amplifying just a few AC coefficients, obtaining an average PSNR of 39.22 dB. An additional interesting concept

(a) (b) (c)

Figure 6: The three signals obtained attacking the system with BNSA.

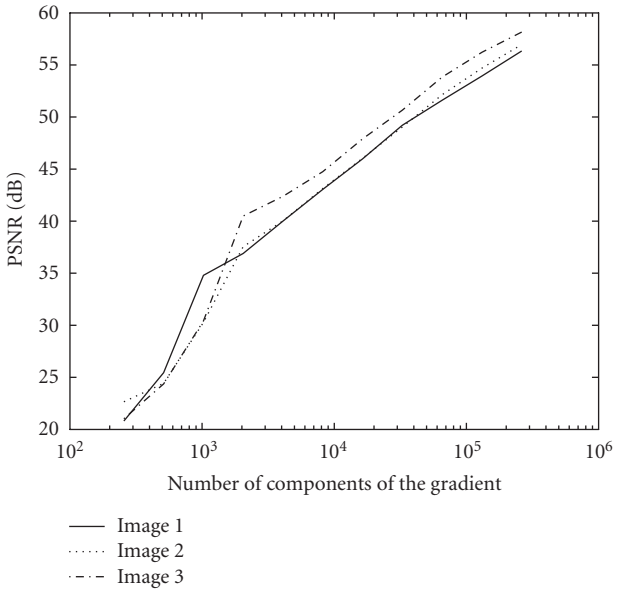

Figure 7: PSNR obtained as a function of the number of computed components of the gradient.

introduced by the authors is *super-robustness*, which they define as *the property of a watermarking algorithm to survive select types of quality degradation far beyond what any reasonable person would expect*, constituting *in fact a security weakness*.

4.4. Westfeld [1, 10]

Westfeld also managed to determine that the watermark was embedded in 12 low-frequency coefficients per block of 8×8 DCT. The author then developed a sensitivity attack and a postprocessing step, which *modifies the pixel values in the spatial domain before they are rounded to integer levels of grey*. The impressive PSNRs obtained with this method are, respectively, 60.74 dB, 57.05 dB, and 57.29 dB.

5. CONCLUSIONS

In this paper, we have presented two approaches for breaking the watermarking scheme employed in the BOWS contest: a key guessing attack and a sensitivity attack. Even though both approaches lead to similar results, only the former relies on certain a priori knowledge about the watermarking system, while the latter is totally blind. This entails a much larger computational load for the sensitivity-based attack.

An important remark is that complete key disclosure was possible in this case due to its small length. This only ratifies the well-known requirement of a large enough key space for preventing exhaustive search attacks. In this sense, the size of the space of the watermaked signal is not so important, as the attacker can focus his attack on the usually smaller key space. This is especially important, as the disclosure of the secret key does not only allow to obtain signals where the watermark is not detected with an excellent quality, but also allows the attacker to generate falsely watermarked signals (also known as *forgeries*) in all cases with almost no extra cost for successive contents, as the attack needs to be performed once per secret key (not per content).

Another important conclusion regarding the BNSA is that it is possible to tradeoff the final PSNR and the computational load; this compromise is achieved by reducing the number of gradient components that are actually computed.

Compared with the results obtained by other groups and recorded in the BOWS webpage [3], we can see that the results achieved by our two attacks are only comparable with those obtained by Westfeld [1] in the stage of the contest where the watermarking method was publicly known. On the first stage, when that method was still undisclosed, the winner was the team led by Scott Craver, who achieved much smaller PSNRs. These results clearly show the effectiveness of the proposed attacks for "black-box" detectors.

Finally, we would like to congratulate the organizers of the BOWS contest, Drs. Barni and Piva, for the success of this challenge. We definitely believe that these initiatives dramatically contribute to the advancement of our discipline. Trying to break BOWS was not only a lot of fun but also allowed us to perfect our blind attack.

ACKNOWLEDGMENTS

The authors would like to thank Dr. M. Miller, Dr. G. Doërr, Prof. I. Cox, and their Ph.D. students for the availability of their code implementing [2]. This work was partially funded by Xunta de Galicia under projects PGIDT04 TIC322013PR, PGIDT04 PXIC32202PM, and Competitive research units program Ref. 150/2006; MEC project DIPSTICK, reference

TEC2004-02551/TCM and European Comission through the IST Programme under Contract IST-2002-507932 ECRYPT. ECRYPT disclaimer: the information in this paper is provided as is, and no guarantee or warranty is given or implied that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.

REFERENCES

[1] A. Westfeld., "Lessons from the BOWS contest," in *Proceedings of the 8th Workshop on Multimedia and Security*, pp. 208–213, Geneva, Switzerland, September 2006.

[2] M. L. Miller, G. J. Doërr, and I. J. Cox, "Applying informed coding and embedding to design a robust high-capacity watermark," *IEEE Transactions on Image Processing*, vol. 13, no. 6, pp. 792–807, 2004.

[3] http://lci.det.unifi.it/BOWS.

[4] http://www.adastral.ucl.ac.uk/~gwendoer/dptWatermarking.

[5] J. W. Earl, "Tangential sensitivity analysis of watermarks using prior information," in *Security, Steganography and Watermarking of Multimedia Contents IX*, E. J. Delp III and P. W. Wong, Eds., vol. 6505 of *Proceedings of SPIE*, pp. 12 pages, San Jose, Calif, USA, January 2007.

[6] P. Comesaña, L. Pérez-Freire, and F. Pérez-González, "Blind newton sensitivity attack," *IEE Proceedings on Information Security*, vol. 153, no. 3, pp. 115–125, 2006.

[7] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, New York, NY, USA, 1999.

[8] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, Nashua, NH, USA, 1995.

[9] M. Barni, F. Pérez-González, P. Comesaña, and G. Bartoli, "Putting reproducible signal processing into practice: a case study in watermarking," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '07)*, vol. 4, pp. 1261–1264, Honolulu, Hawaii, USA, April 2007.

[10] A. Westfeld, "Tackling bows with the sensitivity attack," in *Security, Steganography and Watermarking of Multimedia Contents IX*, E. J. Delp III and P. W. Wong, Eds., vol. 6505 of *Proceedings of SPIE*, pp. 11 pages, San Jose, Calif, USA, January 2007.

[11] S. Craver, I. Atakli, and J. Yu, "How we broke the bows watermark," in *Security, Steganography and Watermarking of Multimedia Contents IX*, E. J. Delp III and P. W. Wong, Eds., vol. 6505 of *Proceedings of SPIE*, pp. 8 pages, San Jose, Calif, USA, January 2007.

[12] G. Le Guelvouit, T. Furon, and F. Cayre, "The good, the bad, and the ugly: three different approaches to break their watermarking system," in *Security, Steganography and Watermarking of Multimedia Contents IX*, E. J. Delp III and P. W. Wong, Eds., vol. 6505 of *Proceedings of SPIE*, pp. 8 pages, San Jose, Calif, USA, January 2007.

[13] S. Pateux and G. Le Guelvouit, "Practical watermarking scheme based on wide spread spectrum and game theory," *Signal Processing: Image Communication*, vol. 18, no. 4, pp. 283–296, 2003.