

Better security levels for ‘Broken Arrows’

Fuchun Xie¹, Teddy Furon², and Caroline Fontaine³

¹ INRIA-Rennes research center, Campus de Beaulieu, 35042 Rennes, France

² Thomson Security Labs, 35511 Cesson-Sévigné, France

³ CNRS/Lab-STICC/CID and Télécom Bretagne/ITI, Technopôle Brest-Iroise, CS 83818 29238 Brest, France

ABSTRACT

This paper considers the security aspect of the robust zero-bit watermarking technique ‘Broken Arrows’(BA),¹ which was invented and tested for the international challenge BOWS-2. The results of the first episode of the challenge showed that BA is very robust. Last year, we proposed an enhancement so-called AWC,² which further strengthens the robustness against the worst attack disclosed during the challenge. However, in the second and third episodes of the challenge, when the pirate observes plenty of watermarked pictures with the same secret key, some security flaws have been discovered. They clearly prevent the use of BA in multimedia fingerprinting application, as suggested in.³ Our contributions focus on finding some counter-attacks. We carefully investigate BA and its variant AWC, and take two recently published security attacks⁴ as the potential threats. Based on this, we propose three countermeasures: benefiting from the improved embedding technique AWC; regulating the system parameters to lighten the watermarking embedding footprint; and extending the zero bit watermarking to multi-bits for further increase the security level. With this design, experimental results show that these security attacks do not work any more, and the security level is further increased.

Keywords: Watermarking, Security, Robustness, Key estimation attacks, Multimedia fingerprinting application

1. INTRODUCTION

BA has been designed especially for the international challenge BOWS-2. Its performance in terms of robustness and imperceptibility are state of the art. Recently, we proposed a complete scheme of multimedia fingerprinting,³ using BA as a practical watermarking layer, coupled with Tardos symmetric fingerprinting code^{5,6}. However, some security flaws have been discovered during the second and third episodes of the challenge, when the pirate observes plenty of watermarked pictures with the same secret key. They clearly prevent the use of BA in multimedia fingerprinting application, since the fingerprinted content provides a huge number of contents marked with the same key at the same time. This gives a great potential for attackers to estimate the watermark or the secret key used by the system, and thereby to remove the fingerprint.

The goal of this paper is to find some countermeasures to fill the security flaws. As pointed out by^{7,8,9} security and robustness issues are different: watermarking robustness concerns the common signal processing actions; while watermarking security involves certain intentional attacks. In this paper, we review the embedding strategies BA and its variant AWC; and carefully analyze its security issues. Furthermore, we deeply study the two watermarking security attacks proposed by Bas and Westfeld.⁴ Based on this, we introduce some efficient solutions to enhance the security performance of BA: reactivate the AWC embedding technique; regulation the embedding parameters; and extension from zero-bit to multi-bits. Experimental results show that, with these solutions, these two watermarking security attacks no longer work.

The outline of this paper is as follows. We review the embedding strategies BA and its variant AWC in Section 2. We describe two watermarking security attacks in Section 3. Some proposed solutions to increase the security performance and their experimental evaluations are provided in Section 4. Finally Section 5 concludes the paper and points out some future directions.

Further author information: (Send correspondence to Fuchun Xie: Fuchun.Xie@inria.fr)

2. EMBEDDING STRATEGIES

BA is used as a practical watermarking solution in our multimedia fingerprinting scheme. Before presenting the watermarking strategy, we make the link between watermarking and fingerprinting clear in using some notation here. Let n be the number of users, m be the fingerprint code length. The fingerprinting code is a set of n different m symbol sequences $\{\mathbf{X}_j\}_{j=1}^n$, where the j -th row corresponds to the j -th user. The symbols $X_{j,i}$ belong to a q -ary discrete alphabet \mathcal{X} , whose size is $|\mathcal{X}| = q$. Each sequence $\mathbf{X}_j = \{X_{j,i}\}_{i=1}^m$ identifying user j has to be hidden in his personal copy via a watermarking technique. The embedding is block based, and we assume that the content is long enough to be divided into at least m blocks. For example, a video can be split into a set of images. We hide one symbol $X_{j,i}$ per block, according to the secret key $K(X_{j,i})$. Here we define q different secret keys to embed symbols of a q -ary alphabet. We now focus on the embedding technique.

2.1 Broken Arrows

The embedding and detection of BA involve four nested spaces: the pixel space, the wavelet subspace, the secret subspace and the MCB (Miller, Cox and Bloom) plane. The main process of watermark generation in BA can be summarized here: 1) Taking the $H_i \times W_i$ matrix \mathbf{i}_X of 8-bit luminance values as the original image in the pixel space; 2) Performing the 2D wavelet transform (Daubechies 9/7) on three levels of decomposition of \mathbf{i}_X , then selecting the coefficients from all the bands in the wavelet subspace except the low-frequency LL band. These $N_s = H_i \times W_i(1 - 1/64)$ wavelet coefficients are then stored as \mathbf{s}_X . 3) They use N_v secret binary antipodal carriers signals of size N_s : $\mathbf{s}_{C,j} \in \{-1/\sqrt{N_s}, 1/\sqrt{N_s}\}^{N_s}$, $\forall j \in \{1, \dots, N_v\}$, produced by a pseudorandom generator seeded by the secret key K' . The host signal is projected onto these carrier signals: $v_X(t) = \mathbf{s}_{C,t}^T \mathbf{s}_X$, these N_v correlations being stored as $\mathbf{v}_X = (v_X(1), \dots, v_X(N_v))^T$. This means that \mathbf{v}_X represents the host signal in the secret subspace. We can write this projection with the $N_s \times N_v$ matrix \mathbf{S}_C whose columns are the carrier signals: $\mathbf{v}_X = \mathbf{S}_C^T \mathbf{s}_X$. The norm is conserved because the secret carriers are assumed to constitute a basis of the secret subspace: $\|\mathbf{v}_X\|^2 = \mathbf{s}_X^T \mathbf{S}_C \mathbf{S}_C^T \mathbf{s}_X \approx \|\mathbf{s}_X\|^2$. 4) Then the host signal \mathbf{v}_X is transferred to the MCB plane. Denote $\mathbf{v}_C^* \in \mathbb{R}^{N_v}$ as the secret vector in the secret subspace. The basis of the MCB plane is given by $(\mathbf{v}_1, \mathbf{v}_2)$ as in [1, Eq.(3)]:

$$\mathbf{v}_1 = \mathbf{v}_C^*, \quad \mathbf{v}_2 = \frac{\mathbf{v}_X - (\mathbf{v}_X^T \mathbf{v}_1) \mathbf{v}_1}{\|\mathbf{v}_X - (\mathbf{v}_X^T \mathbf{v}_1) \mathbf{v}_1\|} \quad (1)$$

Hence, the MCB plane contains \mathbf{v}_C^* and \mathbf{v}_X . The coordinates representing the host are $\mathbf{c}_X = (c_X(1), c_X(2))^T$ with $c_X(1) = \mathbf{v}_X^T \mathbf{v}_1$ and $c_X(2) = \mathbf{v}_X^T \mathbf{v}_2$. According to a certain criterion for maximizing the robustness (see [1, Sec 3.1.2]), the watermarked coordinates $\mathbf{c}_Y = (c_Y(1), c_Y(2))^T$ are presented as:

$$\mathbf{c}_Y = \begin{cases} (c_X(1) + \sqrt{\rho^2 - c_X(2)^2}, 0)^T & \text{for } c_X(2) \leq \rho \cos(\theta) \\ \mathbf{c}_X + \rho(\sin(\theta), -\cos(\theta))^T & \text{for } c_X(2) > \rho \cos(\theta) \end{cases} \quad (2)$$

Here the parameter ρ is related to the embedding distortion constraint, and θ is an angle defining the cone of the detection region. Therefore, the watermark signal in the MCB plane can be represented by $\mathbf{c}_W = \mathbf{c}_Y - \mathbf{c}_X$.

In order to go back to the wavelet subspace, \mathbf{c}_W is firstly projected in the secret space as $\mathbf{v}_W = (\mathbf{v}_1, \mathbf{v}_2)\mathbf{c}_W$; thereby, the watermarked signal in the secret space is $\mathbf{v}_Y = \mathbf{v}_X + \mathbf{v}_W$. Then, \mathbf{v}_W is projected back in the wavelet subspace to get the watermark signal in the wavelet domain \mathbf{s}_W , which can be written as $\mathbf{s}_W = \mathbf{S}_C \mathbf{v}_W$. The watermarking step can then be written as: $\mathbf{s}_Y = \mathbf{s}_X + \mathbf{mask} \cdot \mathbf{s}_W$, where \mathbf{mask} denotes the perceptual mask that modulates the watermark signal \mathbf{s}_W . In BA, we have: $\mathbf{mask}_{BA} = |\mathbf{s}_X|$, where $|\mathbf{s}_X|$ denotes the absolute value of the wavelet coefficients of the host \mathbf{s}_X . This scheme provides perceptually acceptable watermarked pictures.

With PSNR greater than 40 dB, it appears that the amplitude of the samples of \mathbf{s}_W are almost all lower than 1. Therefore, this embedding technique conserves the sign of the wavelet coefficients. This, in our opinion, is a first security flaw in the technique: if the attack strongly modifies the amplitude of the coefficients while preserving their signs, it will be successful with a big probability. This is precisely the case with the attack mounted by A. Westfeld¹⁰ in the first episode of BOWS-2.

2.2 AWC proportional embedding

In order to resist Westfeld’s denoising, some improvements of the embedding scheme were proposed in,² which take into account the dependency between the neighboring coefficients. AWC (Averaging Wavelet Coefficients) proportional embedding is one of the best solution. It replaces the coefficient amplitude in the mask by an average of five coefficients: itself and four neighbors,

$$\mathbf{mask}_{\text{AWC}}(m, n) = \frac{1}{5} \left| \sum_{l=m-1}^{m+1} \sum_{s=n-1}^{n+1} \mathbf{s}_X(l, s) \right| \quad (3)$$

where $\mathbf{s}_X(m, n)$ denotes the wavelet coefficient of the position (m, n) in \mathbf{s}_X for any band except the low frequency LL band. $\mathbf{s}_X(m-1, n)$, $\mathbf{s}_X(m, n-1)$, $\mathbf{s}_X(m+1, n)$, and $\mathbf{s}_X(m, n+1)$ are its four neighbors. In this way, the watermark signal modifies the signs of some host coefficients. This solution is sufficient to cope with Westfeld’s denoising,¹⁰ and some experimental works confirm our argument.²

3. WATERMARKING SECURITY ATTACKS

It is now well known that robustness and security are different issues in the watermarking area^{789, 4} watermarking robustness usually considers classical content processing, while watermarking security is dedicated to malicious attacks aiming at disclosing the secrets of a watermarking technique. In the multimedia fingerprinting scenario, the watermarking technique embeds the fingerprinting code in a video block by block. For a given user, all these blocks are watermarked with a few number of secret keys according to the fingerprinting code. The threat is that a pirate extracts enough information for estimating the secret keys, and thereby removing the watermark while preserving an excellent perceptual quality. We focus here on security attacks specifically dedicated to BA watermarking scheme.

3.1 Westfeld clustering attack

A clustering attack was introduced by A. Westfeld in the third episode of BOWS-2.⁴ Its main steps can be summarized as follows: 1) He applies Westfeld’s denoising¹⁰ to the (10,000 in BOWS-2) watermarked images. 2) As these attacked images look like estimated original images, he remove them from the watermarked images to estimate the watermark signals. 3) These estimated watermarks are sorted into several bins ($N_c = 30$ in BOWS-2), by using a clustering method. 4)For a given bin, he averages all the estimated watermarks of the bin to estimate the secret carrier of this bin, and subtracts it from images watermarked with this carrier. Finally he obtains the attacked images.

3.2 Bas subspace estimation attack

Another watermarking security attack is the subspace estimation attack recently proposed by P. Bas.⁴ We sum up its main process here: 1) Through a huge number of observations, the fast and efficient OPAST algorithm estimates the projection matrix \mathbf{W} , whose size is $N_s \times N_p$, here N_s is the number of the wavelet coefficients to be watermarked and N_p represents the number of principal components. In order to assess the performance of his subspace estimation algorithm, he used the Square Chordal Distance (SCD) between the secret subspace $\text{span}(\mathbf{S}_C)$ and the estimated subspace $\text{span}(\hat{\mathbf{W}})$ during this step. The smaller the SCD , the better the subspace estimation. 2) With this projection matrix $\hat{\mathbf{W}}$, he uses Independent Component Analysis (ICA) technique to estimate each axis direction and thereby the whole secret matrix $\hat{\mathbf{C}}$. 3) Finally, he pushes the watermarked content outside the detection region by making use of the estimated secret matrix $\hat{\mathbf{C}}$. In this way, the watermark is removed with a high PSNR. Our experimentation confirmed its good performance. OPAST (Orthogonal Projection Approximation Subspace Tracking)is the key ingredient of this attack. The usual PCA algorithm based on eigenvalue decomposition cannot operate a so big data set. The designers of BA thought that therefore PCA was no longer a threat. However, the discovery of the inline and iterative OPAST proved that they were wrong.

4. SECURITY IMPROVEMENTS AND EVALUATIONS

In order to prevent these two attacks pulling down the watermarking scheme and thereby our fingerprinting system, we have to find some ways to enhance the watermarking security. In this part, several efficient solutions will be introduced.

4.1 Security of AWC and Westfeld clustering attack

As we mentioned in Section 2.2, AWC proportional embedding is introduced as an efficient solution to prevent the Westfeld's denoising while maintaining a good robustness against a lot of usual attacks.² But its impact on the security performance have never been examined before.

In other words, Westfeld's clustering attack was the best security attack during the third episode of BOWS-2,¹¹ and Westfeld's denoising is a core step in this clustering attack (See Section 3.1). Therefore, since AWC proportional embedding prevents Westfeld's denoising to estimate the watermarks correctly, Westfeld clustering attack may no longer do a good classification of the watermarks.

In order to confirm this idea, we implemented Westfeld's clustering attack in Matlab and found two slight improvements of Algorithm 1.⁴ In this algorithm, all the bin leaders are updated at each iteration (step 4 to step 6). This operation does a lot of repetitive work, and wastes a lot of computing power. Experimentally, for a given cluster, it usually needs 3 or 4 iterations to find a stable bin leader, while the following iterations output the last 2 leaders alternatively. The observation of this phenomenon is used as a stopping condition. This greatly reduces the computing time by 25%. In Westfeld's clustering algorithm (Algorithm 1⁴), the observation which has the smallest correlation with all the existing bin leaders, is selected as a leader of a new bin. But, with this initialization of the new bin, it is possible to select as the bin leader a vector which was already a leader of another bin. This tends to split a 'correct' bin into several small clusters. So in our simulation, we pay attention in truly finding new leaders. In this way, the probability of splitting bins is reduced, and this improves the accuracy of the classification.

We keep the same test condition as A. Westfeld in order to obtain comparable results. Firstly, we take the M images of the BOWS-2 database ($M = 10,000$), then during the watermark embedding, we save the cone index information for every image. This allows to build a ground truth classification. Denote $\mathcal{B}_{ref}(i)$ the subset of all images which have been watermarked with the i -th secret cone. As there are N_c secret cones, the M watermarked images are classified into a partition \mathcal{B}_{ref} of N_c subsets: $\mathcal{B}_{ref} = \bigcup_{i=1}^{N_c} \mathcal{B}_{ref}(i)$. This partition is the ground truth and it will be used to evaluate the accuracy of the clustering attack.

Secondly, we apply Westfeld's denoising on all the watermarked images \mathbf{s}_Y to get an estimation of the original images. This yields an estimation of the watermark signal: $\hat{\mathbf{s}}_W = \mathbf{s}_Y - \hat{\mathbf{s}}_X$. Thirdly, we run our improved Westfeld's clustering attack with a targeted bin number N_t in the range $\{1, \dots, N_c\}$. This yields a partition \mathcal{B}_{est} of N_t subsets: $\mathcal{B}_{est} = \bigcup_{i=1}^{N_t} \mathcal{B}_{est}(i)$.

The question is now how to evaluate the accuracy of this clustering attack. Note that N_t might not be equal to N_c . For that purpose, the confusion matrix \mathbf{P}_{conf} is first computed:

$$P_{conf}(k, l) = \frac{|\mathcal{B}_{est}(k) \cap \mathcal{B}_{ref}(l)|}{M}, \quad \forall (k, l) \in \{1, \dots, N_t\} \times \{1, \dots, N_c\}. \quad (4)$$

This confusion matrix can be considered as the probability transition of a noisy Discrete Memoryless Channel. The subset indices of ground truth partition are the symbols of the source to be broadcast through this channel. Their probabilities are given by $P_{ref}(l) = |\mathcal{B}_{ref}(l)|/M$, $l \in \{1, \dots, N_c\}$. The indices of the partition induced by the clustering attack are the received symbols. Denote $P_{est}(k) = |\mathcal{B}_{est}(k)|/M$, $k \in \{1, \dots, N_t\}$. Then, the accuracy of the attack is measured as the quantity of information its clustering carries about the ground truth partition, ie. the mutual information between the index of the clustering (the 'received symbols') and the index of the ground truth partition (the 'emitted symbols'):

$$MI(\mathcal{B}_{est}, \mathcal{B}_{ref}) = \sum_{k=1}^{N_t} \sum_{l=1}^{N_c} P_{conf}(k, l) \log \frac{P_{conf}(k, l)}{P_{ref}(l) \cdot P_{est}(k)} \quad (5)$$

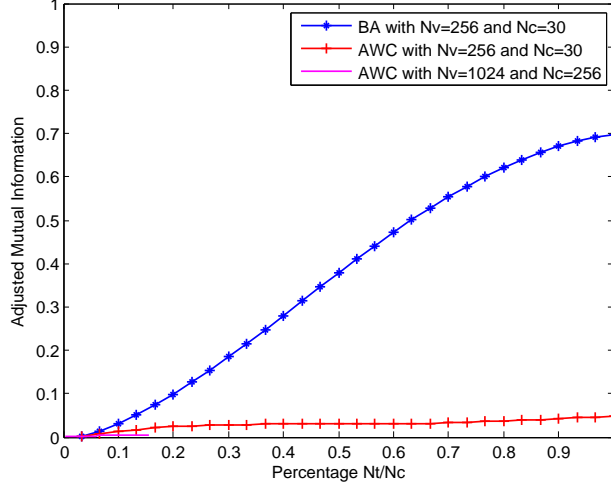


Figure 1. Probability of the good classification for Westfeld classifier against BA and AWC, with different N_v and N_c

Now, the problem is that this measure is very well calibrated. For instance, if the attack doesn't work at all, producing a clustering which is purely random and independent of the ground truth, the expected value of the mutual information will depend on N_t and N_c . This prevents us from comparing clustering accuracy for different values of N_t . The adjusted mutual information (AMI) has been recently proposed by Vinh et al.¹² as a calibrated measure:

$$AMI(\mathcal{B}_{est}, \mathcal{B}_{ref}) = \frac{MI(\mathcal{B}_{est}, \mathcal{B}_{ref}) - \mathbb{E}\{MI(\mathcal{R}(N_t), \mathcal{R}(N_c))\}}{\max_{\mathcal{C}}\{MI(\mathcal{C}, \mathcal{B}_{ref})\} - \mathbb{E}\{MI(\mathcal{R}(N_t), \mathcal{R}(N_c))\}} \quad (6)$$

where $\mathbb{E}\{MI(\mathcal{R}(N_t), \mathcal{R}(N_c))\}$ is the expected mutual information between two random clusterings $\mathcal{R}(N_t)$ of size N_t and $\mathcal{R}(N_c)$ of size N_c , under a given statistical model; and $\max_{\mathcal{C}}\{MI(\mathcal{C}, \mathcal{B}_{ref})\}$ is maximum value of the mutual information for this particular ground truth partition (indeed the maximum of the entropies of the 'emitted' and 'received' symbols).

In our simulation we measure the accuracy of the clustering attack by the adjusted mutual information $AMI(\mathcal{B}_{est}, \mathcal{B}_{ref})$ and then we plot it with respect to the ratio N_t/N_c in Figure 1. For the original BA embedding technique, with $N_t/N_c = 1$, the adjusted mutual information AMI is 0.7, this means that the estimated clustering \mathcal{B}_{est} is very similar to the ground truth partition, and the Westfeld clustering attack succeeds in estimating the secret cones with a decent accuracy. However, this classifier does not work with our improved embedding method AWC: for almost all ratio N_t/N_c from 0 to 1, the adjusted mutual information is smaller than 0.05. So AWC proportional embedding is an efficient solution to block Westfeld's clustering attack.

4.2 A counter attack to Bas' subspace estimation

Patrick Bas' subspace estimation attack doesn't use Andreas Westfeld's denoizing. Therefore, AWC is likely not to provide any hint. This subspace estimation is based on the fact that the embedding deeply changes the power of the signal in the secret space. This leaves a clue for disclosing this space.

4.2.1 Security measurement

The main idea of BA is to project the signal composed of the wavelet coefficients onto a secret subspace of dimension $N_v = 256$. From 2000 images of the BA databases,¹ the results of the power distributions of the original projected vector \mathbf{v}_X and the watermarked projected vector \mathbf{v}_Y are shown in Figure 2. The power P_X in the secret space is uniformly distributed: there is no particular reason why this vector could have more power in any given direction of the secret space. We can measure the power for the original host signal \mathbf{v}_X in the secret space by:

$$P_X = \frac{1}{N_v} \mathbb{E}(c_X(1)^2 + c_X(2)^2) \quad (7)$$

Yet, the power P_Y of \mathbf{v}_Y is very different: The embedding process has changed the power distributions. In order to maximize the robustness, we push the watermarked vector inside the detection region as deep as possible. This operation inevitably increases the power along the secret cone direction, and decreases the power of the other directions.

We model the power distribution as follows: The embedder selects one secret cone among N_c ones with a uniform probability $p_s = 1/N_c$. Once a given secret cone is selected, the power is $P_s = \mathbb{E}(c_Y(1)^2)$; otherwise, the power is $P_n = \mathbb{E}(\frac{c_Y(2)^2}{N_v-1})$, because the $N_v - 1$ elements share the energy of $c_Y(2)^2$. See notations in the paper of Furon and Bas.¹ Thus, for the first N_c components of \mathbf{v}_Y , the power (expectation of the power per component) $P(\mathbf{v}_Y(k))$ is:

$$\begin{aligned} P_Y(k) &= p_s \cdot P_s + (1 - p_s) \cdot P_n \\ &= \frac{1}{N_c} \mathbb{E}(c_Y(1)^2) + \frac{N_c - 1}{N_c(N_v - 1)} \mathbb{E}(c_Y(2)^2) \end{aligned} \quad (8)$$

The $(N_v - N_c)$ remaining directions of the secret space are not secret cone support. The expectation of the power $P_Y(k)$ for $N_c < k \leq N_v$ is:

$$P_Y(k) = \frac{1}{N_v - 1} \mathbb{E}(c_Y(2)^2) \quad (9)$$

The difference of power between a direction selected as a secret cone support and a direction not selected as secret cone support in the watermarked correlation \mathbf{v}_Y can be written as:

$$d_1 = |P_Y(1) - P_Y(N_v)| \quad (10)$$

$$= \frac{1}{N_c} \mathbb{E}(c_Y(1)^2) - \frac{1}{N_v - 1} \mathbb{E}(c_Y(2)^2) \quad (11)$$

The difference of power between a direction of the secret space (not a secret cone support) and a direction not in the secret space is denoted d_2 :

$$d_2 = |P_X - P_Y(N_v)| \quad (12)$$

$$= \frac{1}{N_v} \mathbb{E}((c_X(1)^2 + c_X(2)^2)) - \frac{1}{N_v - 1} \mathbb{E}(c_Y(2)^2) \quad (13)$$

Bigger values of (d_1, d_2) ease the pirate job in disclosing the secret subspace, and in this subspace, the directions used as secret cone directions. An embedding technique lower this two values provides a better level against Bas' subspace estimation attack, but it is impossible to achieve the ideal case: $d_1 = 0$ and $d_2 = 0$.

We compare the distances for these two embedding techniques: original BA with proportional embedding and BA with AWC embedding. They have almost the same values for d_2 ; AWC has a distance d_1 just a little bit smaller than the one of the original BA (Figure 2). Whereas AWC embedding is a good counter-measure against Westfeld's denoising and clustering attack, it does not help against Bas' subspace estimation attack.

4.2.2 Regulated parameters

Inserting (2) in (10) and (12), we have:

$$d_1 = \pi \frac{\mathbb{E}((c_X(1) + \sqrt{\rho^2 - c_X(2)^2})^2)}{N_c} + (1 - \pi) \left(\frac{\mathbb{E}((c_X(1) + \rho \sin(\theta))^2)}{N_c} - \frac{\mathbb{E}((c_X(2) - \rho \cos(\theta))^2)}{N_v - 1} \right), \quad (14)$$

and

$$d_2 = \pi \frac{\mathbb{E}(c_X(1)^2 + c_X(2)^2)}{N_v} + (1 - \pi) \left(\frac{\mathbb{E}(c_X(1)^2 + c_X(2)^2)}{N_v} - \frac{\mathbb{E}((c_X(2) - \rho \cos(\theta))^2)}{N_v - 1} \right), \quad (15)$$

where π is the probability that $c_X(2) \leq \rho \cos(\theta)$. In these two equations, ρ is a parameter related to the embedding distortion, we can not modify it arbitrarily since we need a high quality watermarked content and an acceptable PSNR. Parameter θ is the angle of the detection cone region; it cannot be modified because it fixes the false detection probability.

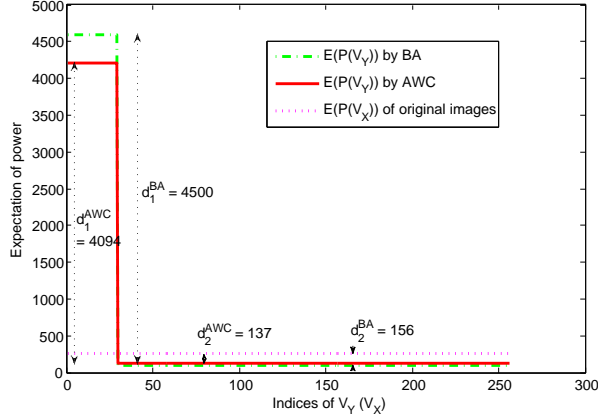


Figure 2. Power distribution of the correlation vectors $\mathbf{v}_Y(\mathbf{v}_X)$ with BA and AWC proportional embeddings (with $N_v=256$ and $N_c=30$).

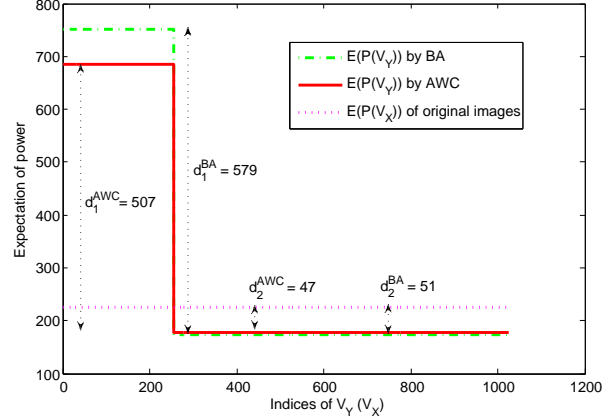


Figure 3. Power distribution of the correlation vectors $\mathbf{v}_Y(\mathbf{v}_X)$ with BA and AWC proportional embeddings (with $N_v=1024$ and $N_c=256$).

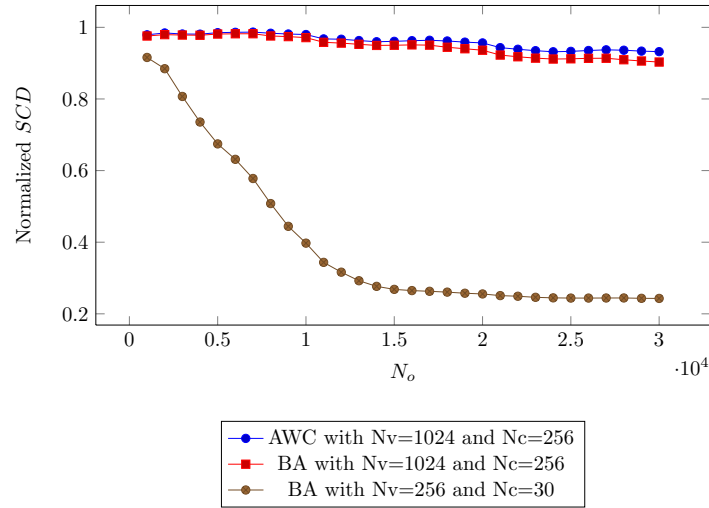


Figure 4. Normalized SCD for the embedding techniques BA and AWC with different parameters N_v and N_c

Therefore, we can only tune parameters N_v and N_c . The analysis is quite involved since the statistics of \mathbf{c}_X , π and θ also depends on these parameters. Firstly, we fix N_c and analyze the impact of N_v . Changing N_v has almost no effect on d_1 , whereas d_2 is clearly decreasing with N_v . Thereby, after carefully considering that the complexity of the embedding and detection algorithms are in $O(N_v)$, we choose $N_v = 1024$.

Now we study the last parameter N_c : d_1 is decreasing function wrt N_c . In this regard, we should choose N_c as big as possible. But, a bigger N_c lowers the value of θ giving birth to a small detection region, and this significantly degrades the robustness of the system. We made a trade-off with $N_c = 256$.

4.2.3 Security evaluations against attacks

With the new parameters $N_v = 1024$ and $N_c = 256$, we check the security level of the embedding techniques against these attacks with the same database as in the third episode of BOWS-2 (10,000 images). First of all, we test Westfeld's clustering attack. Figure 1 shows the performance of Westfeld's classifier against AWC with $N_v=1024$ and $N_c=256$, up to $N_t \leq 40$: the adjusted mutual information AMI is around 0.006, ie. much smaller than for the case of AWC embedding with $N_v = 256$ and $N_c = 30$. So AWC with the regulated parameters has a better security level than before.

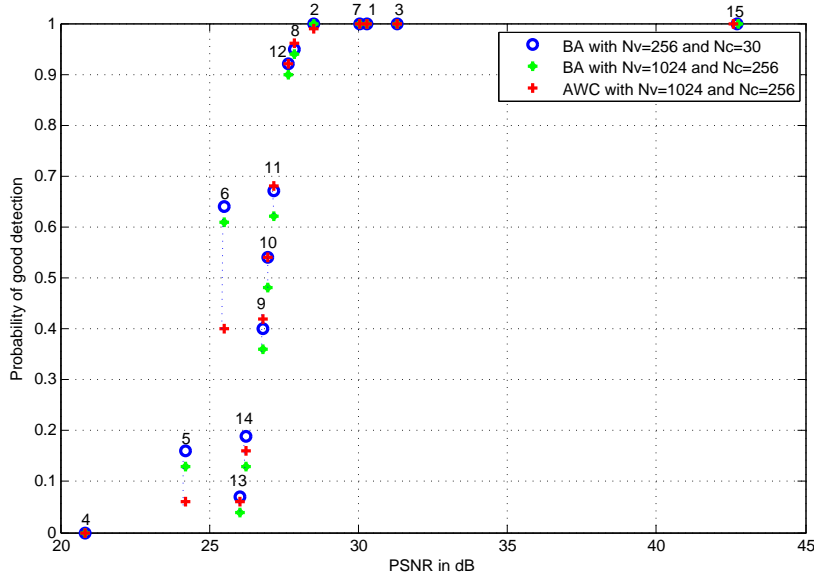


Figure 5. Probability of good detection versus average PSNR of the attacked images for the three watermark embedding techniques: ‘BA’ proportional embedding with $N_v=256$ and $N_c=30$ ‘o’, ‘BA’ proportional embedding with $N_v=1024$ and $N_c=256$ ‘*’, ‘AWC’ proportional embedding with $N_v=1024$ and $N_c=256$ ‘+’. Selection of attacks: 1) denoise threshold 20; 2) denoise threshold 30; 3) JPEG Q = 20; 4) JPEG2000 r = 0.001; 5) JPEG2000 r = 0.003; 6) JPEG2000 r = 0.005; 7) scale 1/2; 8) scale 1/3; 9) scale 1/3 + JPEG Q = 50; 10) scale 1/3 + JPEG Q = 60; 11) scale 1/3 + JPEG Q = 70; 12) scale 1/3 + JPEG Q = 90; 13) scale 1/4 + JPEG Q = 70; 14) scale 1/4 + JPEG Q = 80; 15) no attack.

We also evaluate its security performance against Bas’ subspace estimation attack. In order to compare the performances of this attack, we define a normalized square chordal distance: $SCD_{norm} = SCD/N_c$. Note that here SCD is the Square Chordal Distance between the secret space and the estimated subspace, and N_c is the number of the secret cone directions in the embedding process. $SCD_{norm} = 0$ means that the estimated space is equal to the secret space; $SCD_{norm} = 1$ means the subspaces are orthogonal and, therefore, the attack has failed. Figure 4 shows the results of the OPASt algorithm applied against the original BA and AWC embedding with different parameters. N_o is the number of observations. We can see that, for BA embedding technique with $N_v = 256$ and $N_c = 30$, SCD_{norm} is decreasing with the number of observations, and the estimation keeps on improving very quickly. This confirms Patrick Bas’ results.⁴ However, for BA and AWC embedding techniques with the regulated parameters $N_v = 1024$ and $N_c = 256$, SCD_{norm} is always close to 1, this shows that the OPASt algorithm cannot estimate the secret subspace any longer.

4.2.4 Robustness evaluations

To examine the robustness impact brought by the proposed solution in the watermarking layer, we apply the same benchmark as in BA’s original paper: a number of attacks mainly composed of combinations of JPEG and JPEG 2000 compressions at different quality factors, low-pass filtering, wavelet subband erasure, and a simple de-noising algorithm. Figure 5 reveals the impact of 15 most significant attacks on the two embedding techniques. The probability of detecting the watermark (i.e. number of good detections divided by 2,000) is plotted with respect to the average PSNR of the attacked images. Because these classical attacks produce almost the same average PSNR, the three points for a given attack are almost vertically aligned.

The probability of detection is slightly decreased when N_v (resp. N_c) increases from 256 (resp. 30) to 1024 (resp. 256) for BA embedding. The proposed counter-attacks trade a great improvement of security levels against a little bit of robustness. Comparing to the original BA embedding, the AWC embedding is more robust against attacks 9-14, but less robust against attacks 5 and 6; and comparable for attacks 1-4, 7, and 15.

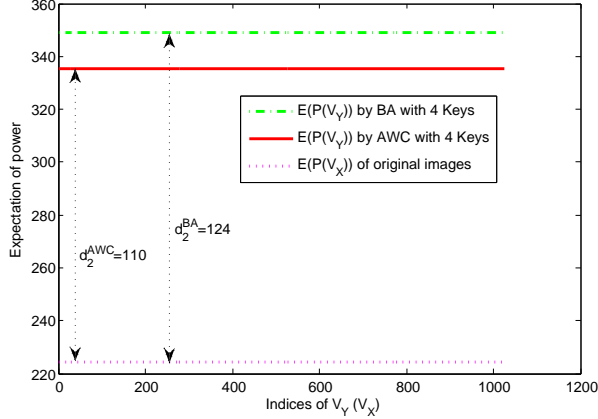


Figure 6. Power distribution of the correlation vectors $\mathbf{v}_Y(\mathbf{v}_X)$ with BA and AWC proportional embeddings extended to multi-bits (with $N_v=1024$ and $N_c=256$).

4.3 Extension to On-Off Keying

So far, we have discussed about BA as a zero-bit watermarking technique and independently of any particular scenario. In a previous paper³, we used BA in a traitor tracing application in conjunction with an anti-collision q -ary Tardos code. The secret subspace is decomposed of q complementary spaces: each one of them gathers the secret cone directions associated to one symbol. Therefore, we force $N_v = q \cdot N_c$. Since the symbol to be embedded are uniformly distributed, all the directions of the secret subspace have the same probability to serve as a secret cone support. This cancels the use of distance d_1 (or in other words, this automatically sets $d_1 = 0$). Another advantage of this solution is that it also reduces d_2 .

In order to verify our arguments, we use 2000 images (as in BA¹) to evaluate distance d_2 in the traitor tracing scenario. The PSNR of the watermarked images is controlled around 43 dB. The alphabet size for the fingerprinting codewords is $q = 4$. The reason has been given in a former paper:³ If we watermark an image with q different symbols, then an averaging attack followed by a JPEG compression deludes the detection more than half of the time. So there is no advantage in having q higher than 4. We also fix the parameters $N_v=1024$ and $N_c=256$ for both BA and AWC proportional embeddings, and thereby $N_v = q * N_c$.

In this experiment, we assume that the fingerprinting symbols are uniformly distributed over all the 2000 images. Figure 6 shows that, with our proposed solution, d_1 is artificially reduced to 0 in this application. This is a significant progress, since in a pure zero-bit scenario d_1 has a huge value for both embedding methods ($d_1^{AWC} \approx 4100$ and $d_1^{BA} \approx 4500$, see Figure 2). On the other hand, d_2 is also slightly decreased: $d_2^{AWC} = 110$ and $d_2^{BA} = 124$ (before $d_2^{AWC} = 137$ and $d_2^{BA} = 156$, see Figure 2).

5. CONCLUSION

We proposed counter-attacks to known attacks against the Broken Arrow watermarking technique. Thanks to a conjunction of the AWC embedding, the regulated parameters N_v and N_c plus the conditions of use in the traitor tracing scenario, these attacks are no longer threats.

The cost of a better security levels is a small loss in robustness compared to the original BA technique, and slower embedding and detection algorithms (by a factor of 4).

However, the assessment of a higher security level is not completed: we addressed only some known attacks, worse threats certainly still exist. Moreover, the main counter-attack simply suggests to use a ‘bigger’ secret (a ‘longer secret key’ would say a cryptographer), which is not a new idea. Our future work will try to find more universal evidences of better security levels for Broken Arrows.

6. ACKNOWLEDGEMENT

The authors thank the French national programme ‘Audiovisuel et Multimedia’ under project MEDIEVALS 2007-AM-005-04 for their financial support.

The authors also would like to thank P. Bas for his constructive discussions and the source code of the subspace estimation attack, and A. Westfeld for his useful explanations about the clustering attack.

REFERENCES

- [1] Furon, T. and Bas, P., “Broken arrows,” *EURASIP Journal on Information Security* (2008).
- [2] Charpentier, A., Xie, F., Furon, T., and Fontaine, C., “Expectation maximisation decoding of tardos probabilistic fingerprinting code,” in [*Proc. of SPIE on Security, Steganography and Watermarking of Multimedia Contents XI, San Jose, California, USA*], *Proc. of SPIE on Media Forensics and Security XI, San Jose, California, USA* (January 2009).
- [3] Xie, F., Furon, T., and Fontaine, C., “On-off keying modulation and tardos fingerprinting,” in [*Proc. of 10th ACM Multimedia and Security Workshop, Oxford, UK*], *Proc. of 10th ACM Multimedia and Security Workshop, Oxford, UK* (September 2008).
- [4] Bas, P. and Westfeld, A., “Two key estimation techniques for the broken-arrows watermarking scheme,” in [*Proc. of 11th ACM Multimedia and Security Workshop, Princeton, USA*], *Proc. of 11th ACM Multimedia and Security Workshop, Princeton, USA* (September 2009).
- [5] Tardos, G., “Optimal probabilistic fingerprint codes,” in [*Proc. of the 35th annual ACM symposium on theory of computing*], 116–125, ACM (2003).
- [6] Skoric, B., Katzenbeisser, S., and Celik, M., “Symmetric Tardos fingerprinting codes for arbitrary alphabet sizes,” *Designs, Codes and Cryptography* **46**, 137–166 (February 2008).
- [7] Cayre, F., Fontaine, C., and Furon, T., “Watermarking security: Theory and practice,” *IEEE Trans. Signal Processing* **53**, 3976 – 3987 (october 2005).
- [8] Furon, T., “A survey of watermarking security,” in [*Proc. of Int. Work. on Digital Watermarking*], Barni, M., ed., *Lecture Notes on Computer Science* **3710**, 201–215, Springer-Verlag, Sienna, Italy (september 2005).
- [9] Pérez-Freire, L., Comesaña, P., Troncoso-Pastoriza, J. R., and Pérez-González, F., “Watermarking security: a survey,” *Transactions on Data Hiding and Multimedia Security I* **4300**, 41–72 (October 2006).
- [10] Westfeld, A., “A regression-based restoration technique for automated watermark removal,” in [*Proc. of 10th ACM Multimedia and Security Workshop, Oxford, UK*], *Proc. of 10th ACM Multimedia and Security Workshop, Oxford, UK* (September 2008).
- [11] BOWS-2, “<http://bows2.gipsa-lab.inpg.fr/>,” (2007).
- [12] Vinh, N., Epps, J., and Bailey, J., “Information theoretic measures for clusterings comparison: Is a correction for chance necessary?,” in [*the 26th International Conference on Machine Learning (ICML’09), Montreal, Canada*], (2009).