Towards Robust and Secure Watermarking

Fuchun Xie, Teddy Furon, Caroline Fontaine

Towards Robust and Secure Watermarking*

Fuchun Xie
INRIA-Rennes Bretagne
Atlantique Research Center
Campus de Beaulieu
35042 Rennes, France
fuchun.xie@inria.fr

Teddy Furon
INRIA-Rennes Bretagne
Atlantique Research Center
Campus de Beaulieu
35042 Rennes, France
teddy.furon@inria.fr

Caroline Fontaine
CNRS/Lab-STICC/CID
and Télécom Bretagne/ITI
CS 83818-29238 Brest,France
caroline.fontaine@telecom-bretagne.eu

ABSTRACT

This paper presents yet another attempt towards robust and secure watermarking. Some recent works have looked at this issue first designing new watermarking schemes with a security oriented point of view, and then evaluating their robustness compared to state-of-the-art but unsecure techniques. Our approach is, on contrary, to start from a very robust watermarking technique and to propose changes in order to strengthen its security levels. These changes include the introduction of a security criterion, an embedding process implemented as a maximization of a robustness metric under the perceptual and the security constraints, and a watermarking detection seen as a contrario decision test.

Our experimentations lead to, once again, a trade-off between security and robustness. The technique is now perfectly secure against attacks mounted during the second edition of the BOWS challenge, but the price to pay is either a lower robustness against common image processing, either a bigger probability of false alarm.

Categories and Subject Descriptors

I.4.9 [**Computing Methodologies**]: Image processing and computer vision, Applications

General Terms

Design, Experimentation

Keywords

Watermarking, robustness, security, attacks, optimization under constraints.

1. INTRODUCTION

Security of digital watermarking is now widely considered as very different from robustness since [1] and [4]. This last reference provides a framework for assessing the security levels of a watermarking technique, and it has been used to analyze some very popular schemes, for instance quantization index modulation in [13] or dirty paper trellis in [2]. These analyses revealed the relatively weak security levels of state-of-the-art robust techniques.

More recently, this subject turned into a more fundamental question: is it possible to create a watermarking technique which is both robust and secure? Some articles designed schemes with security as a top priority, then evaluated the robustness [10, 11]. The gap is quite big compared to past robust techniques. But even if there is no fundamental proof (as far as we know), a trade-off between robustness and security seems to exist in real-life watermarking techniques.

Here our approach is the reverse, as we start with a very robust zero-bit watermarking technique[1], *Broken Arrows*, and try to increase its security levels. *Broken Arrows* has been designed for the second contest of Break Our Watermarking System (BOWS-2) [9]. The embedding is based on spread spectrum in the wavelet domain, but with side information enforced through a maximization under an unique constraint: the robustness, defined as the minimum noise power to go outside the detection region, is maximized under the perceptual constraint, which is here accounted by an Euclidean distance. Exploiting the fact that the embedding did not modify the signs of the wavelet coefficients, A. Westfeld found the more powerful attack of the first episode of the contest. This weakness has been patched in [5]: an attack removing the watermark with probability 1/2 will now degrade the image quality down to 26dB whereas the embedding quality is set to 43dB.

This increases the robustness, but the security levels of *Broken Arrows* were to be improved as well. The embedding is considerably modifying the power distribution of the signal in the secret subspace. With OPAST, an available implementation of the Principal Component Analysis (PCA), the authors of [3] succeeded to disclose the secret space of the original version of *Broken Arrows*, pointing out a lack of security. In a recent paper, we proposed a counter-measure to this attack [15]. The distribution of the signal power is much more uniform than before, and this is sufficient to ruin the above mentioned attacks. However, this is not perfect. It might still be possible for the pirate to use more powerful implementations of PCA than OPAST, and collect

*Partly funded by ANR-06-SETI-009 Nebbiano and ANR-07-AM-005 Medievals French projects.

MM&Sec'10, September 9–10, 2010, Roma, Italy.

[1]There is no capacity since this is zero-bit watermarking, see [8].

Representatives	Original	Watermarked	Watermark	Dimension
subscript	X	Y	W	-
"image" in pixel space	$\mathbf{i}_X$	$\mathbf{i}_Y$	$\mathbf{i}_W$	$W_\mathbf{i} \times H_\mathbf{i}$
"signal" in wavelet space	$\mathbf{s}_X$	$\mathbf{s}_Y$	$\mathbf{s}_W$	N_s
"vector" in secret space	$\mathbf{v}_X$	$\mathbf{v}_Y$	$\mathbf{v}_W$	N_v

Table 1: The table of terminology and notations.

much more watermarked images to take advantage of this still slightly uneven power distribution.

In this paper, we strenghen this scheme against attacks based on second order statistics such as PCA, by striking a perfectly even distribution of the power. This security oriented focus is implemented in practice by a maximization under two constraints: the embedding aims at maximizing robustness under a perceptual constraint *and* a security constraint. To the best of our knowledge, this is the first time that watermarking security is enforced right into the embedding algorithm.

The remaining of this paper is organized as follows. In Section 2, we introduce the motivations and relevant inspiration to design this robust and secure watermarking scheme. Section 3 describes the implementation of the proposed watermarking algorithm. Section 4 depicts the watermark embedding core process. Experimental results are illustrated in Section 5, and the conclusion is given in Section 6. The terminology and notations are summarized in Table 1.

2. A CONTRARIO DECISION

As far as we know, there is no known optimum zero-bit watermarking technique for multimedia contents. This is mostly due to the lack of stationarity and to the wide variety of distribution from a content to another. From the theoretical viewpoint, [8] proposes a unifying theory of zero-bit watermarking, but its main drawback is its lack of universality: the embedder and detector must know the statistical distribution of the host content. More recently, Comesaña et al. have found the optimum scheme under the restrictive assumption that the host distribution belongs to the white and Gaussian family [6], whatever its variance. To apply this theoretical result into a real application, *Broken Arrows* [9] projects many wavelet coefficients $\mathbf{s}_X \in \mathbb{R}^{N_s}$ of the image (with $N_s > 200,000$) into a secret (i.e. pseudo-randomly generated) subspace $\mathcal{S} = \text{Span}(\mathbf{S}_C)$ of very low dimension ($N_v = 256$). This makes the projection vector $\mathbf{v}_X = \mathbf{S}_C^T \mathbf{s}_X \in \mathbb{R}^{N_v}$ almost white and Gaussian distributed. At the end, we can write that $\mathbf{v}_X \sim \mathcal{N}(\mathbf{0}, \sigma_X^2 \mathbf{I}_{N_v})$, where $\sigma_X^2 = ||\mathbf{s}_X||^2$, i.e. the variance varies from a content to another. Several tweaks are also deployed to tackle a perceptual model and to improve the robustness against common image processing (see [9]).

However, as already mentioned in the introduction, *Broken Arrows* is robust but not secure. The embedding pushes the vector $\mathbf{v}_X$ deep into the acceptance region which is an hypercone, focusing most of the embedding energy along its directions. This brings in an uneven distribution of the watermarked signals power in the space, and a PCA algorithm can disclose the directions of the hypercone.

We propose here a very different paradigm, inspired by works in computer vision proposing a partial gestalt system based on the Helmholtz principle [7]. This principle groups a set of observed objects into one class if it is very unlikely that randomness could have generated such configuration. This is also known in statistics as *a contrario detection*. The detection decides one of the two hypotheses H_0 or H_1 based on some observations $\mathbf{x}$. A statistical model is assumed under H_0 as a distribution p_{H_0}, but there is no such model for the alternative H_1, because this hypothesis is far too broad, and/or what happens under H_1 is not well known. Therefore, the Neyman-Pearson theorem doesn't apply and no maximum likelihood test is possible. Instead, the detector evaluates the probability to observe $\mathbf{x}$ under p_{H_0}, and if this event is unlikely, it decides for H_1.

Here we exactly consider the same idea: for original natural images, we assume that the projection vector $\mathbf{v}_R$ of the received image is white Gaussian distributed: $p_{H_0} \propto \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{N_v})$. Watermark detection is triggered on when this vector is not typical from that distribution. Therefore, embedding amounts to render $\mathbf{v}_X$ as much as possible (i.e. while fulfilling the constraints of embedding distortion and security) not typical with regard to p_{H_0}. This method is *different* from usual approaches where embedding transforms host vector $\mathbf{v}_X \sim p_{H_0}$ into a watermarked vector $\mathbf{v}_Y \sim p_{H_1}$ and where the detector measures the likelihood ratio $p_{H_1}(\mathbf{v}_R)/p_{H_0}(\mathbf{v}_R)$ (or its derivative for a LMP test) to decide whether $\mathbf{v}_R$ is watermarked or not.

3. THE EMBEDDING CORE PROCESS

We detail in this section how the concepts introduced in the above sections are implemented. We assume here that we have extracted a Gaussian distributed vector $\mathbf{v}_X$ from the host content, and that a targeted PSNR controls the watermark embedding. This scheme applies to any spread spectrum based watermarking on any type of content. The next session gives details when the scheme is applied to a particular still image watermarking for illustration purpose.

3.1 Three functions

As mentioned above, we have to manage three criteria during the embedding: the perceptual quality of the watermarked image, the global security of the scheme, and the Gaussianity of the signals.

3.1.1 Quality

We assume the targeted PSNR imposes a constraint on the Euclidean distance of the watermark signal $\mathbf{v}_W = \mathbf{v}_Y - \mathbf{v}_X$, such that $\|\mathbf{v}_W\| < \rho$. We will see later that parameter ρ is not only a function of the targeted PSNR but also on some statistic of the host image. For this reason, it might vary from a content to another, even if the required PSNR is fixed. We define the following function:

$$c_Q(\mathbf{v}_W) = \frac{\|\mathbf{v}_W\| - \rho^2}{\rho^2}. \quad (1)$$

3.1.2 Security

As mentioned in the introduction, a counter-attack to the OPAST threat (and any algorithm implementing a PCA) is to make sure that, on average, the power of the watermarked signal in the secret space $\mathcal{S}$ is equal to the one in the complementary space $\mathcal{S}^\perp$: $P_Y(\mathcal{S}) = P_Y(\mathcal{S}^\perp)$. We assume that for natural images, the power of the host signal is evenly distributed: $P_X(\mathcal{S}) = P_X(\mathcal{S}^\perp)$. Since the embedding only modifies the signal in $\mathcal{S}$, we need to enforce a conservation of the power in this subspace: $P_Y(\mathcal{S}) = P_X(\mathcal{S})$.

This argument holds on average: $P_X(\mathcal{S})$ is the average power present in $\mathcal{S}$ over a very large collection of natural images. However, we have noticed that the energy of the projection onto $\mathcal{S}$, namely $\|\mathbf{v}_X\|$, greatly varies from one image to another, and it might be hazardous to bet on any average value. Therefore, we enforce a stricter rule: $\|\mathbf{v}_Y\| = \|\mathbf{v}_X\|$. If the energy is conserved for any host image, then it must be true on average. We define the following function:

$$c_S(\mathbf{v}_W, \mathbf{v}_X) = \frac{\|\mathbf{v}_X + \mathbf{v}_W\| - \|\mathbf{v}_X\|}{\|\mathbf{v}_X\|}. \quad (2)$$

3.1.3 Gaussianity

There are plenty of tests to decide whether a collection of i.i.d. data has been drawn from a given probability distribution. For continuous distributions and especially the Gaussian distribution, the Anderson-Darling test is one of the most famous. In brief, the test computes the statistic $\hat{A}^2$ from the samples $\{v(i)\}_{i=1}^{N_v}$ as follows:

1. Mean μ and the standard deviation σ are estimated from the samples.
2. Samples are normalized: $v^n(i) = (v(i) - \mu)/\sigma$.
3. Samples are then sorted in increasing order to get $(v^s(1), \dots, v^s(N_v))$.
4. Compute $\hat{A}^2 = (-N_v - T)(1 + \frac{4}{N} - \frac{25}{N^2})$.
5. If $\hat{A}^2 > \alpha$, then the samples are not Gaussian distributed.

with

$$T = N_v^{-1} \sum_{i=1}^{N_v} \quad [(2i-1)\ln\Phi(v^s(i)) + (2(N-i)+1)\ln(1-\Phi(v^s(i)))].$$

The value of α depends on the level of the test (see [14, Table 1, part (a), page 239]). For instance, if $\hat{A}^2 > 1.029$, the data are deemed non Gaussian, and the probability of being wrong is 0.01. We take $\hat{A}^2$ as the detection score $f(\mathbf{v}_R)$ for the vector $\mathbf{v}_R$ extracted from the received image. The image is declared as watermarked if $f(\mathbf{v}_R) > \alpha$ and the probability of false alarm is indeed the level of the test corresponding to this threshold.

3.2 Constrained optimization

We consider the watermark embedding as a maximization under constraints. The embedding looks for the watermark vector $\mathbf{v}_W^\star$ which maximizes the objective function $f(\mathbf{v}_X + \mathbf{v}_W)$ under the constraints that $c_Q(\mathbf{v}_W) \leq 0$ and $c_S(\mathbf{v}_W, \mathbf{v}_X) = 0$:

$$\mathbf{v}_W^\star = \arg \max_{\mathbf{v}_W \in \mathbb{R}^{N_v} : c_Q(\mathbf{v}_W) \leq 0, c_S(\mathbf{v}_W, \mathbf{v}_X) = 0} f(\mathbf{v}_X + \mathbf{v}_W) \quad (3)$$

3.2.1 Necessary conditions

So far, the constraints can be trivially fulfilled by setting $\mathbf{v}_W = \mathbf{0}$. Therefore, we are sure to maximize an objective function over a non empty set. However, since watermarking is always a matter of trade-off between distortion and robustness, we would like to consume all the allowed distortion to maximize our chance of being robust. In other words, we would like to replace the inequality by the equality $c_Q(\mathbf{v}_W) = 0$. There is a necessary condition so that both equality constraints can be satisfied. The constraint on quality describes a hypersphere of radius ρ centered on $\mathbf{v}_X$, whereas the constraint on security defines a hypersphere of radius $\|\mathbf{v}_X\|$ centered on $\mathbf{0}$. Both constraints can be fulfilled if the intersection of those two regions is not empty. This holds if the necessary condition is true:

$$\rho/2 \leq \|\mathbf{v}_X\|. \quad (4)$$

The equality holds in this equation when the hyperspheres are tangent in a point $\mathbf{v}_W = -2\mathbf{v}_X$ so that $\mathbf{v}_Y = -\mathbf{v}_X$. Since $\mathbf{v}_X$ is assumed to be Gaussian distributed, so is $\mathbf{v}_Y$, and consequently an embedding restricted to consume all the distortion budget fails in this case.

To avoid this situation, we need to properly design the technique so that over a vast majority of images, Inequality (4) holds. However, it is clear that some pictures will not be watermarked, such as a uniform image. This is indeed quite sound. Watermarking content when the host power is too weak raises a security flaw as the host is not properly hiding the watermarking signal. Since $v_X(i) \sim \mathcal{N}(0, \overline{\Sigma^2})$, with $\overline{\Sigma^2}$ the average power of the wavelet coefficients (see [9, Sec. 3.3]), then $\mathbb{E}[\|\mathbf{v}_X\|^2] = N_v\overline{\Sigma^2}$. This shows that the dimension reduction operated by the projection from $\mathbb{R}^{N_s}$ to $\mathbb{R}^{N_v}$ must not be too strong. This is the reason why we increase N_v from 256 (as set in the original *Broken Arrows* technique) to 1024. There is clearly a trade-off with the complexity of the embedder and detector.

For some rare images, this precaution is not enough and (4) does not hold. We then reduce the embedding distortion to 90% of the maximum $2\|\mathbf{v}_X\|$ and we stay with an equality quality constraint. Therefore, we hope that the maximization is done over a large enough set, and with a large enough embedding distortion budget in order to find a big and robust extremum of f.

3.2.2 Numerical algorithm

We use the Matlab implementation of the 'Interior-Point Algorithm' to solve this maximization under constraints. This program can tackle large scale problems and is robust, as it can recover from 'NaN' or 'Inf' results. It also takes benefit from 'user-supplied' derivatives, and Hessian of the objective and the constraints functions. An important point is that these functions are not convex; therefore, there are *a priori* local maxima. When starting from different initial points, the algorithm might end at different local maxima. Here is a way to find a suitable initial point:

1. Define the constants
$$\alpha = 1 - \frac{\rho^2}{2\|\mathbf{v}_X\|^2}, \ \beta = \|\mathbf{v}_X\|\sqrt{(1-\alpha^2)}$$
2. Randomly draw a vector $\mathbf{n} \in \mathbb{R}^{N_v}$.
3. Compute $\mathbf{n}' = \mathbf{n} - \frac{\mathbf{n}^T\mathbf{v}_X}{\|\mathbf{v}_X\|^2}\mathbf{v}_X$

4. Set $\mathbf{v}_W^{(0)} = (\alpha - 1)\mathbf{v}_X + \beta \frac{\mathbf{n}'}{\|\mathbf{n}'\|}$.

It is easy to see that $c_Q(\mathbf{v}_W^{(0)}) = c_S(\mathbf{v}_W^{(0)}, \mathbf{v}_X) = 0$. Since the creation of such an initial vector is not a computational burden, we generate plenty of them, we compute their scores with the function f, and we give Matlab the one with the biggest score as an initial vector.

4. PLUGGING *BROKEN ARROWS*

This section focuses on the embedding process and how the above algorithm is plugged into *Broken Arrows* still image watermarking technique. The discrete wavelet transform of the original image is computed. Some wavelet coefficients are stored in a vector $\mathbf{s}_X$. Its projection onto the secret subspace is $\mathbf{v}_X = \mathbf{S}_C^T \mathbf{s}_X$. This matrix is composed of N_v binary carriers of $\{+1, -1\}^{N_s}$, normalized by a factor of $1/\sqrt{N_s}$. From the input $\mathbf{v}_X$, the above algorithm outputs a watermark vector $\mathbf{v}_W$ of norm ρ, which is mapped back into the wavelet domain by $\mathbf{s}_W = \mathbf{S}_C \mathbf{v}_W$. The norm of $\mathbf{s}_W$ equals $\mathbf{v}_W^T \mathbf{S}_C^T \mathbf{S}_C \mathbf{v}_W$. If $\mathbf{S}_C$ was orthonormal, then we would have $\|\mathbf{s}_W\| = \rho$. Yet, this is not the case. We prefer to have a very fast generator of $\mathbf{S}_C$ (pseudo-random binary generator) than a true orthonormal matrix. However, since it is a very long random matrix, we assume that it is almost the case:

$$\|\mathbf{s}_W\| \approx \|\mathbf{v}_W\| = \rho. \tag{5}$$

4.1 Impacts of the perceptual mask

The watermark signal is added to $\mathbf{s}_X$ with some perceptual mask m:

$$s_Y(i) = s_X(i) + m(i).s_W(i). \tag{6}$$

In the original technique, the mask was 'proportional' in the sense that $m(i) = |s_X(i)|$. However, a security flaw stemmed from this mask, and we use the more secure AWC mask proposed in [5].

The mask has two important impacts we will take into account in the following subsections.

4.1.1 Effect on the embedding distortion

We model the masking weights by random variables statistically independent of $\mathbf{s}_W$, and with second order moment empirically measured as $\overline{M^2} = N_s^{-1} \sum_{i=1}^{N_s} m(i)^2$. This assumption allows us to write that $\|\mathbf{s}_Y - \mathbf{s}_X\|^2 \approx \overline{M^2}.\|\mathbf{s}_W\|^2 \approx \overline{M^2}\rho^2$. This squared norm is also equal to the Mean Squared Error over the image, times the number of pixels (because the wavelet transform conserves the Euclidean norm). Hence the following relationship between ρ and the required PSNR:

$$\rho \approx \frac{255.\sqrt{W_\mathbf{i} H_\mathbf{i}}}{\sqrt{\overline{M^2}}} 10^{-\mathrm{PSNR}/20}, \tag{7}$$

where $(W_\mathbf{i}, H_\mathbf{i})$ is the width and height of the original image.

4.1.2 Effect on the projection vector

A difficulty stems from the fact that the mask disturbs the retro-projection. When we mix the generated watermark signal $\mathbf{s}_W$ in the wavelet space, and then retro-project the watermarked signal $\mathbf{s}_Y$ back onto the secret space, it is not located where we expect, that is, not in $\mathbf{v}_Y = \mathbf{v}_X + \mathbf{v}_W$. Actually, the retro-projection denoted by $\mathbf{v}_Y^{(1)}$ works as follows:

$$v_Y^{(1)}(k) = v_X(k) + \sum_{j=1}^{N_v} v_W(j) \sum_{i=1}^{N_s} m(i) s_{C,j}(i) s_{C,k}(i). \tag{8}$$

We need to assume that i) the involved variables can be treated as independent r.v., ii) the second sum over N_s coefficients can be seen as the empirical average equaling the expectation, iii) $\mathbf{S}_C^T \mathbf{S}_C \approx \mathbf{I}_{N_v}$, to derive this simplification:

$$v_Y^{(1)}(k) \approx v_X(k) + v_W(k)\overline{M} \tag{9}$$

with $\overline{M} = N_s^{-1} \sum_{i=1}^{N_s} m(i)$. Therefore, from a vector $\mathbf{v}_W$ of norm ρ created at the embedding side, we end up with a vector $\mathbf{v}_W^{(1)} = \mathbf{v}_Y^{(1)} - \mathbf{v}_X \approx \overline{M}\mathbf{v}_W$, at the detection side.

We must take this 'amplification' into account when looking for the best watermark vector. For this, we modify ρ by a factor $\overline{M}$, and the maximization under constraint yields a vector $\mathbf{v}_W$ that we scale by a factor $1/\overline{M}$, before embedding in signal:

$$\rho = \frac{255.\overline{M}\sqrt{W_\mathbf{i} H_\mathbf{i}}}{\sqrt{\overline{M^2}}} 10^{-\mathrm{PSNR}/20} \tag{10}$$

$$s_Y(k) = s_X(k) + m(k).s_W(k)/\overline{M}. \tag{11}$$

In this way, we are sure to maximize the score at the detection side, while yielding the required PSNR at the embedding side.

4.2 Improving the accuracy

A big difference with *Broken Arrow* is that the score is calculated from a vector of big dimension N_v ($f : \mathbb{R}^{N_v} \to \mathbb{R}$), whereas in *Broken Arrows* this vector was projected again on a 2D space before the score was computed ($\mathbb{R}^2 \to \mathbb{R}^+$). Figure 8 in [9] shows that there is some inaccuracy in the embedding: in this 2D space, the embedding targets a given location, and at the detection side, the watermarked image is projected on a different position nearby. This inaccuracy is due to the approximations we made so far, and we noticed that its impact is even bigger with our proposed scheme, certainly because the score is now computed on a much higher dimension space. We propose here to reduce this inaccuracy.

4.2.1 Orthonormal Matrix

As discussed in Section 4.1.2, Relation (9) is obtained thanks to the three listed assumptions, and it is quite easy to get rid off the last one. As already mentioned, the generation of the secret matrix $\mathbf{S}_C$ is very fast, but this matrix is not exactly orthonormal. $\mathbf{S}_C^T \mathbf{S}_C$ is positive definite, and is thus diagonalizable as $\mathbf{V}^T \Lambda \mathbf{V}$ with $\mathbf{V}^T \mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}_{N_v}$ and Λ a diagonal matrix. Let us denote by $\mathbf{C}$ the square root of the inverse matrix: $\mathbf{C} = \mathbf{V}^T \Lambda^{-1/2} \mathbf{V}$. For a given secret key, we compute in advance this $N_v \times N_v$ matrix. Finally, we modify the projection steps as follows:

$$\mathbf{v} = \mathbf{C}\mathbf{S}_C^T \mathbf{s}, \tag{12}$$

$$\mathbf{s} = \mathbf{S}_C \mathbf{C} \mathbf{v}. \tag{13}$$

This renders our scheme more accurate for two reasons:

- Now we have exactly $\|\mathbf{s}_W\|^2 = \mathbf{v}_W^T \mathbf{C}^T \mathbf{S}_C^T \mathbf{S}_C \mathbf{C} \mathbf{v}_W = \mathbf{v}_W^T \Lambda^{-1/2} \Lambda \Lambda^{-1/2} \mathbf{v}_W = \|\mathbf{v}_W\|^2$ instead of the Approximation (5).
- If Assumptions i) and ii) of 4.1.2 hold, then at the detection side we would get $\mathbf{v}_Y = \mathbf{v}_X + \overline{M}\mathbf{v}_W$.

The storage of the pre-computed matrix $\mathbf{C}$ however needs 4 MB for $N_v = 1024$.

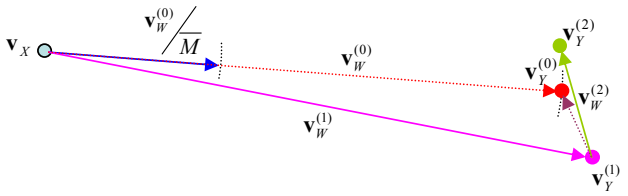

Figure 1: The iterative projection scheme

4.2.2 *Iterative embedding*

The remaining assumptions i) and ii) of 4.1.2 do not exactly hold in practice, and even with (10) and (11), we get $\mathbf{v}_Y^{(1)} = \mathbf{v}_Y + \epsilon$ with $\|\epsilon\| \ll \|\mathbf{v}_Y\|$. We propose the following iterative embedding sketched in Figure 1 to combat this effect.

- Initialization: Once the wavelet transform completed and the extracted coefficients stored in $\mathbf{s}_X$, compute $\mathbf{v}_X$ by (12) and ρ by (10). Denote $\mathbf{v}_W^{(0)}$ the result of the maximization under constraints (3) and $\mathbf{v}_Y^{(0)} = \mathbf{v}_X + \mathbf{v}_W^{(0)}$. Compute $\mathbf{s}_W^{(0)}$ (13) and embed it in the host signal (6) to get $\mathbf{s}_Y^{(1)}$.
- Iteration K: From $\mathbf{s}_Y^{(K)}$, project back onto the secret space (12). This gives $\mathbf{v}_Y^{(K)}$. Compute $\mathbf{v}_W^{(K)} = \mathbf{v}_W^{(K-1)} + (\mathbf{v}_Y^{(K)} - \mathbf{v}_Y^{(0)})$. Compute $\mathbf{s}_W^{(K)}$ (13) and embed it the host signal (6) to get $\mathbf{s}_Y^{(K+1)}$.
- Stop: We stop iterating when we are close to the desired point $\mathbf{v}_Y^{(0)}$, i.e. $\|\mathbf{v}_Y^{(K)} - \mathbf{v}_Y^{(0)}\| < \eta$. The watermarked coefficients are copied back in the wavelet domain and the inverse wavelet transform gives the watermarked image $\mathbf{i}_Y$.

This iterative process does take time but, in practice, just one iteration is enough, as it already greatly reduces the inaccuracy of the embedding.

5. EXPERIMENTAL RESULTS

The perceptual distortion and the security performance of the proposed watermarking system are ensured in the watermark signal generation as we have shown above. So, in this section, we just assess the robustness of the proposed watermarking scheme. First of all, we give the experimental setup.

5.1 Setup

We test 2000 luminance images of size 512×512. These pictures represent natural and urban landscapes, people, or objects, taken with many different cameras from 2 to 5 millions of pixels. A three-level wavelet decomposition is performed for each image, using a Daubechies 9/7 biorthogonal wavelet. Then, the selected wavelet coefficients are projected to the secret matrix, which is generated by the Mersenne Twister pseudorandom number generator seeded by a secret key. The dimension of the secret space is $N_v = 1024$. The embedding distortion is set by a targeted PSNR of 43dB (except in Section 5.2). The number of the tested starting vectors is set to 100. The critical value α is set to 1.029 for a 1% level. The options of the Matlab Interior Point algorithm are set as follows: TolFun=10^{-6}, TolCon=10^{-6}, MaxIter=120, gradConstr = 'on', gradObj = 'on', DerivativeCheck = 'off', FunValCheck = 'on', Hessian = 'user-supplied'.

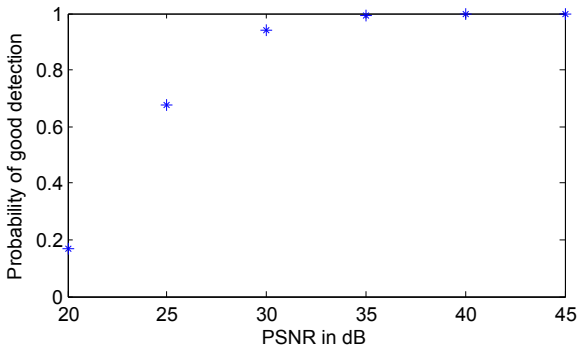

Figure 2: The good detection probability with the Gaussian noise attack.

5.2 Noise attacks

To evaluate the robustness against attack noise, we add the noise directly to the watermarked vector $\mathbf{v}_Y$:

$$\mathbf{v}'_Y = \mathbf{v}_Y + \sigma_N \mathbf{n} \tag{14}$$

where $\mathbf{n}$ is drawn from a normal distribution, and σ_N is the power of the attack in relation with a PSNR_a between the attacked image and the watermarked image:

$$\sigma_N = 255\sqrt{\frac{W_\text{i} H_\text{i}}{N_s}}.10^{-\frac{\text{PSNR}_\text{a}}{20}}. \tag{15}$$

This artificial attack allows to benchmark the key ideas of this paper: i) include the security criterion right into the embedding process, ii) *a contrario decision* test, while decoupling it from this image processing implementation.

To get a better simulation, for each image, we test it with 30 different noise patterns, and compute the average acceptance rate. Figure 2 plots the good detection probability against PSNR_a in dB.

5.3 Common attacks

The benchmark of the real still image watermarking technique is the same as in [9]: the attacks are mainly composed of combinations of JPEG and JPEG 2000 compressions at different quality factors, low-pass filtering, wavelet subband erasure, and a simple denoising algorithm. Figure 3 shows the average PSNR of the attacked images and the average probability of good detection for the 14 most efficient attacks on the proposed watermarking technique. The result of the benchmark is quite similar to the ones shown in [9, Figure 11] or [15, Figure 5]. However, this technique is much weaker than the previous ones because the probability of false alarm here is set to 1.10^{-2} whereas the previous levels were at 3.10^{-6}. This is the price to pay for a good security level.

5.4 Collusion attacks

We speak of *collusion attacks* when several copies of the same piece of content, watermarked with different secret

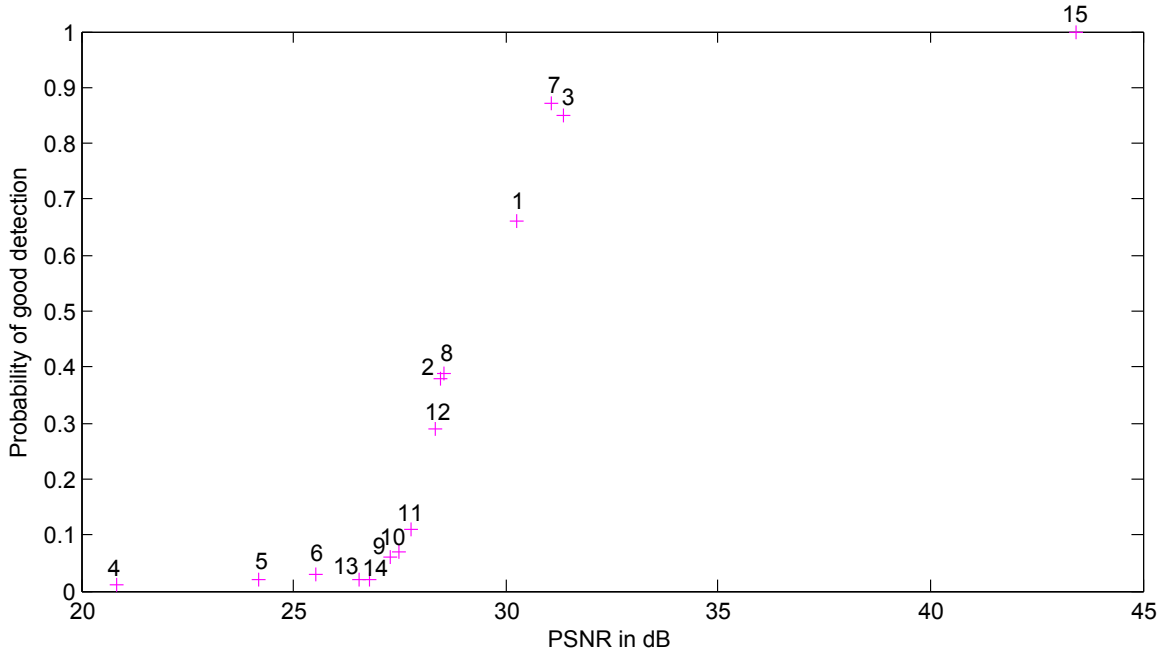

Figure 3: Probability of good detection versus average PSNR of the attacked images for the proposed watermark embedding techniques. Selection of attacks: 1) denoise threshold 20; 2) denoise threshold 30; 3) JPEG Q = 20; 4) JPEG2000 r = 0.001; 5) JPEG2000 r = 0.003; 6) JPEG2000 r = 0.005; 7) scale 1/2; 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.

keys are mixed to forge an illegal copy. This is the typical scenario of multimedia fingerprinting or traitor tracing. As already discussed in a previous paper [16], a zero-bit watermarking technique as *Broken Arrows* is a good candidate as the embedding layer in such a scenario, different message symbols being related to different keys. If we focus on binary anti-collusion codes, as Tardos codes, one bit is embedded in each block of the piece of content. So, each block i exists only in two versions: $\mathbf{i}_{Y1}$ and $\mathbf{i}_{Y2}$. We have tested the following fusion processes at the mixing step of the attack:

1. Average: $i'_Y(i,j) = (i_{Y1}(i,j) + i_{Y2}(i,j))/2$;
2. Interleaving: $i'_Y(i,j) \in \{i_{Y1}(i,j), i_{Y2}(i,j)\}$ with probability $\text{Prob}(i'_Y(i,j) = i_{Y1}(i,j)) = 1/2$;
3. Maximum: $i'_Y(i,j) = \max\{i_{Y1}(i,j), i_{Y2}(i,j)\}$;
4. Minimum: $i'_Y(i,j) = \min\{i_{Y1}(i,j), i_{Y2}(i,j)\}$;
5. Uniform: $i'_Y(i,j)$ is a r.v. $\sim \mathcal{U}([i_{Y1}(i,j), i_{Y2}(i,j)])$.

Over a set of 2,000 images, Table 2 shows the estimated probability $\text{Prob}(K)$ of detecting $K \in \{0,1,2\}$ watermarks from the attacked images $\mathbf{i}'_Y$. The collusion succeeds in erasing both watermarks only with a maximum rate not larger than 3%. All attacks yield double detection with a big probability (more than 94%), which greatly improves the performances of the tracing algorithm [12]. Note that the high probability of false alarm is not a problem in this application because we are only dealing with the attacked watermarked images and because the anti-collusion fingerprinting code can deal with a small amount of detection errors.

Attacks	$K=0$	$K=1$	$K=2$
Average	0.0195	0.0255	0.9550
Interleaving	0.0295	0.0275	0.9430
Maximum	0.0215	0.0270	0.9515
Minimum	0.0195	0.0300	0.9505
Uniform	0.0210	0.0285	0.9505

Table 2: The probabilities Prob(K) of detecting K watermarks from the attacked images $\mathbf{i}'_Y$.

6. CONCLUSION

Despite the introduction of an arsenal of new ideas (maximization under a security constraint, *a contrario* decision test ...), we face once again the same trade-off between security and robustness: To keep the same performances in terms of good detection against common image processing, we had to increase the probability of false alarm.

The proposed scheme is not perfectly secure. Taking into account this particular security constraint in the embedding guarantees that the scheme is only secure against second order statistics analysis tools. However, some high order statistics might leak information on the secret space. Therefore, the issues now turn to be how many contents and computing power a high order analysis requires to work accurately. We believe that it is significantly more demanding, we will try to quantify this gap in the future.

Despite its poor trade-off probability of false alarm vs. robustness, we believe that this scheme has some serious po-

tential in the images (or video) watermarking applications, especially in multimedia fingerprinting (a.k.a. traitor tracing), since the contents are all watermarked in this scenario the probability of false alarm is no longer a problem. The question is more about the symbols likely to be hidden in the pirated copy. As far as we know, any watermark detector outputs binary decision about the presence or absence of the watermarks (this includes potential multiple detections). The *a contrario* decision test can indeed provide a probability of the presence of a given symbol; ie. a soft output bringing more information for the Tardos accusation step. This idea is also the subject of our next work.

7. REFERENCES

[1] M. Barni, F. Bartolini, and T. Furon. A general framework for robust watermarking security. *Signal Processing*, 83(10):2069–2084, October 2003.

[2] P. Bas and G. Doërr. Practical security analysis of dirty paper trellis watermarking. In *9th Information Hiding workshop*, volume 4567 of *LNCS*, Saint-Malo, June 2007. Springer Verlag.

[3] P. Bas and A. Westfeld. Two key estimation techniques for the broken arrows watermarking scheme. In *Proc. of 11th ACM Multimedia and Security Workshop, Princeton, USA*, September 2009.

[4] F. Cayre, C. Fontaine, and T. Furon. Watermarking security: Theory and practice. *IEEE Trans. Signal Processing*, 53(10):3976 – 3987, october 2005.

[5] A. Charpentier, F. Xie, T. Furon, and C. Fontaine. Expectation Maximisation decoding of Tardos probabilistic fingerprinting code. In *Proc. of SPIE Electronic Imaging on Media Forensics and Security XI, San Jose, California, USA*, January 2009.

[6] P. Comesaña, M. Barni, and N. Merhav. Asymptotically optimum embedding strategy for one-bit watermarking under gaussian attacks. In *Security, Steganography and Watermarking of Multimedia contents VIII*, volume 6819 of *Proc. of SPIE-IS&T Electronic Imaging, SPIE*, San Jose, CA, USA, jan 2008.

[7] A. Désolneux, L. Moisan, and J.-M. Morel. A grouping principle and four applications. *IEEE Trans. on Pattern Analysis and Machine Learning*, 25(4):508–513, April 2003.

[8] T. Furon. A constructive and unifying framework for zero-bit watermarking. *IEEE Trans. Information Forensics and Security*, 2(2):149–163, jun 2007.

[9] T. Furon and P. Bas. Broken arrows. *EURASIP Journal on Information Security*, 2008.

[10] B. Mathon, P. Bas, and F. Cayre. Practical performance analysis of secure modulations for WOA spread-spectrum based image watermarking. In *Proceedings of the 9th ACM workshop on Multimedia and Security*, pages 237–244, New York, NY, USA, 2007.

[11] B. Mathon, P. Bas, F. Cayre, and B. Macq. Optimization of natural watermarking using transportation theory. In *Proceedings of the 9th ACM workshop on Multimedia and Security*, Princeton , NJ, USA, Jun 2009.

[12] L. Pérez-Freire and T. Furon. Blind decoder for binary probabilistic traitor tracing codes. In *Proceedings of First IEEE International Workshop on Information Forensics and Security*, pages 56–60, London, UK, December 2009. IEEE WIFS'09.

[13] L. Pérez-Freire, F. Pérez-González, T. Furon, and P. Comesaña. Security of lattice-based data hiding against the known message attack. *IEEE Trans. on Information Forensics and Security*, 1((4)):421–439, dec 2006.

[14] G. Shorack and J. Wellner. *Empirical Processes With Applications to Statistics*. Wiley, 1986.

[15] F. Xie, T. Furon, and C. Fontaine. Better security levels for Broken Arrows. In *Proc. of SPIE Electronic Imaging on Media Forensics and Security XII, San Jose, California, USA*, January, 2010.

[16] F. Xie, T. Furon, and C. Fontaine. On-off keying modulation and tardos fingerprinting. In *Proc. of 10th ACM Multimedia and Security Workshop, Oxford, UK*, September 2008.