Watermarking attack: security of WSS techniques

François Cayre, Caroline Fontaine, Teddy Furon

Watermarking Attack: Security of WSS Techniques

François Cayre[1], Caroline Fontaine[2], and Teddy Furon[1]

[1] INRIA projet TEMICS
{francois.cayre,teddy.furon}@irisa.fr
[2] CNRS LIFL, Université des Sciences et Technologies de Lille
Caroline.Fontaine@lifl.fr *

Abstract. Most of watermarking techniques are based on Wide Spread Spectrum (WSS). Security of such schemes is studied here in adopting a cryptanalysis point of view. The security is proportional to the difficulty the opponent has to recover the secret parameters, which are, in WSS watermarking scheme, the private carriers. Both theoretical and practical points of view are investigated when several pieces of content are watermarked with the same secret key. The opponent's difficulty is measured by the amount of data necessary to estimate accurately the private carriers, and also by the complexity of the estimation algorithms. Actually, Blind Source Separation algorithms really help the opponent exploiting the information leakage to disclose the secret carriers. The article ends with experiments comparing blind attacks to these new hacks. The main goal of the article is to warn watermarkers that embedding hidden messages with the same secret key might is a dangerous security flaws.

1 Introduction, Context and Notation

A lot of digital watermarking techniques have been designed those last years. They mainly aim at embedding an invisible watermark into the document in a robust manner. Several kinds of schemes have been proposed, but this article only deals with blind robust watermarking. The reliability is usually evaluated through benchmark tests aiming at removing the watermark [1].

Benchmarking is not really a *security* evaluation, but mainly a *robustness* evaluation. In [2], Kalker defines *robust* watermarking as a communication channel multiplexed into original content in a non-perceptible way, and whose "*capacity [...] degrades as a smooth function of the degradation of the marked content*", and *security* as "*the inability by unauthorized users to access the communication channel*" established by a robust watermark. Accessing the communication channel means to remove, read, or write the hidden message. Hence, *security* deals with intentional attacks, excluding those already encompassed in the robustness category since the watermark is assumed to be robust.

This paper adopts a cryptanalytic approach, in the sense that the attacker first recovers the secret that has been used for the generation of the watermark. This approach is certainly not the only one but secret disclosure is a very powerful hack: it gives the access of the communication channel at the lowest distortion price to hack content. The key idea of this security analysis is that information about the secret key might leak from the observations. Hence, the *a posteriori* ignorance of the opponent decreases as he makes more and more observations. As suggested by Diffie and Hellman [3], different contexts of attack are investigated according to the type of observations available to the opponent.

* This work is supported by the french ACI Fabriano and the european Network of Excellence ECRYPT.

1. In *Known Original Attack – KOA* – the opponent observes N_o pairs of (watermarked / original contents).
2. In *Known Message Attack – KMA* – the opponent has access to N_o (watermarked contents / hidden messages) pairs.
3. In *Watermark Only Attack – WOA* – the opponent has only access to N_o watermarked contents.

As Shannon did [4], it is worth distinguishing what can be stated as a theoretical fact, and the practical tools making the attack really work. Hence, for each of the above-mentioned contexts of attack, the security analysis aims at evaluating two criteria: the *security level*, that is, the theoretical number of observations needed to disclose the secret key, and the *work*, that is, the complexity of the algorithm extracting information about the secret key from observations.

Such a security analysis can only be assessed for a given watermarking algorithm. Here, we decided to focus on spread spectrum based techniques, which are widely used for still images watermarking. Theoretical studies [5] and practical implementations [6] focus on the optimization of operational capacity-robustness functions for a given embedding distortion.

The novelty of this paper resides in the practical implementation of the new watermarking security paradigm whose theoretical background is exposed in [7]. The algorithms we found to hack wide spread spectrum (WSS) techniques come from the Blind Source Separation (BSS), community like Principal Component Analysis (PCA) and Independent Component Analysis (ICA). This use of PCA and ICA in watermarking security analysis is new, as the only other papers mentioning PCA/ICA in the watermarking community have different purposes. González-Serrano *et al* [8] and Bounkong *et al* [9] used ICA to design a watermarking embedder. Du *et al* [10] presented a technique for estimating the watermark by observing only one image. Their purpose is the simple erasure of the whole watermark signal and not the disclosure of the secret parameters. Our approach allows a complete access to the watermarking communication channel to remove, read or write hidden data[3].

The paper is organized as follows. Section 2 summarizes the theoretical discussions about the measurement of the secret information leakages from the observations and the *security level*. Section 3 focuses on the *work* as the complexity of the tools. In particular, extremely high complexity renders the attacks hardly possible. We discuss some possible strategies to decrease the work to an acceptable amount. In both sections, the three contexts (KOA, KMA, WOA) are investigated. Section 4 finally presents some practical results on watermarked images, where adaptation of the tools to real signals was necessary.

2 Theoretical Results

We present a model for WSS watermarking and the methodology applied in the rest of this section. Details in proofs are omitted but can be found in [7].

2.1 Watermarking Model

Let us denote by $\boldsymbol{x}$ a vector of N_v samples extracted from original content. The embedding is the addition of the watermark signal, giving $\boldsymbol{y} = \boldsymbol{x}+\boldsymbol{w}$. The watermark

[3] We have discovered after submission a similar approach uniquely devoted to watermark removal and only based on PCA in [11].

signal $\boldsymbol{w}$ is the modulation of N_c private carriers $\boldsymbol{u}_i$:

$$\boldsymbol{w} = \frac{\gamma}{\sqrt{N_c}} \sum_{k=1}^{N_c} a(k)\boldsymbol{u}_k, \tag{1}$$

where $\gamma > 0$ is a small gain fixing the embedding strength and $\|\boldsymbol{u}_k\| = 1$, $1 \leq k \leq N_c$. An inverse extraction function puts back the watermarked vector $\boldsymbol{y}$ into the media to produce the watermarked piece of content.

The symbols a_k represent the message to be hidden/transmitted through content. In the case of a BPSK [12], symbols $a(i)$ take one of the following values $\{-1, +1\}$. Note that this model also covers some side-informed watermarking techniques called spread transform [13, 14], where $a(k)$ are real values uniformly distributed in $[-\Delta/2, \Delta/2]$. In all cases, the WSS aims at increasing the signal to noise ratio by projecting signals on a smaller subspace of dimension N_c. This implies that $N_v > N_c$. Moreover, to cancel inter-symbol interferences at the decoding side, the carriers are two-by-two orthogonal. For security reason, they are private and issued by a pseudo-random generator fed by the secret key.

In the sequel, the security analysis considers several watermarked vectors $\boldsymbol{y}_j$ $(1 \leq j \leq N_o)$, with different embedded symbols $\boldsymbol{a}_j = (a_j(1) \dots a_j(N_c))^T$ being linearly mixed by the $N_v \times N_c$ matrix $\mathcal{U} = (\boldsymbol{u}_1 \dots \boldsymbol{u}_{N_c})$. Index i denotes the i^{th} samples of a given signal, whereas j indices the different signals. Thus, there are N_o watermarked vectors or, equivalently, with the $N_v \times N_o$ matrix $\mathcal{Y} = (\boldsymbol{y}_1 \dots \boldsymbol{y}_{N_o})$, and the $N_c \times N_o$ matrix $\mathcal{A} = (\boldsymbol{a}_1 \dots \boldsymbol{a}_{N_o})$:

$$\boldsymbol{y}_j = \boldsymbol{x}_j + \frac{\gamma}{\sqrt{N_c}}\mathcal{U}\boldsymbol{a}_j \quad \text{is equivalent to} \quad \mathcal{Y} = \mathcal{X} + \frac{\gamma}{\sqrt{N_c}}\mathcal{U}\mathcal{A}\,. \tag{2}$$

2.2 Methodology

Some preliminary works have already adapted the classical guidelines of cryptanalysis to watermarking [15, 16]. Their first assumption is given by the Kerckhoffs' principle [17] stating that the encryption/watermarking algorithm is public, but parametrized by a secret key. In watermarking, it means that the attacker knows the extraction and inverse extraction functions. Thus, if he has access to a watermarked piece of content, he can observe its extracted vector $\boldsymbol{y}_j$.

Now, as Shannon did in [4], we consider that several pieces of content have been watermarked with the same key K. The opponent's goal is to disclose K by observing these pieces of content (i.e. their extracted vectors $\{\boldsymbol{y}_j\}$). Shannon named the *equivocation* $e(N_o)$ the entropy of K knowing N_o observations. It measures the ignorance of the attacker after having observed N_o pieces of content, as the following equation holds:

$$e(N_o) = H(K) - I(K; \mathcal{Y}), \tag{3}$$

where $H(K)$ is the entropy of K (i.e. the ignorance of the attacker before observing any content) and $I(K; \mathcal{Y})$ is the mutual information (i.e., a measure of the information about K that leaks from signal set $\{\boldsymbol{y}_j\}_{j=1}^{N_o}$). A physical interpretation readily comes: when $e({N_o}^\star) = 0$, the attacker has enough observations to disclose the secret key. The security level of the system is of ${N_o}^\star$ observations.

Does Information Leak? However, the knowledge of the carriers is sufficient to hack a WSS watermarking scheme: these private parameters allow the decoding, the embedding and the removal of the watermark. It is not necessary to disclose the secret key K that fed the pseudo-random generator issuing the carriers [2]. The real issue then concerns the information leakage about $\boldsymbol{w}$ from watermarked

signal $\boldsymbol{y}$. For instance, suppose that host signal $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \mathcal{R}_{\boldsymbol{X}})$ and $\boldsymbol{w}$ is picked up randomly among sequences distributed as $\mathcal{N}(\boldsymbol{0}, \mathcal{R}_{\boldsymbol{W}})$. Then, $p_{\boldsymbol{Y}} = \mathcal{N}(\boldsymbol{0}, \mathcal{R}_{\boldsymbol{X}} + \mathcal{R}_{\boldsymbol{W}})$ and $p_{\boldsymbol{Y}|\boldsymbol{W}=\boldsymbol{w}} = \mathcal{N}(\boldsymbol{w}, \mathcal{R}_{\boldsymbol{X}})$. This gives:

$$I(\boldsymbol{W}; \boldsymbol{Y}) = \frac{1}{2} \log \frac{\det \mathcal{R}_{\boldsymbol{X}} + \mathcal{R}_{\boldsymbol{W}}}{\det \mathcal{R}_{\boldsymbol{X}}} \geq 0\,. \tag{4}$$

This equation is extremely important as it shows that there is a leak of information about $\boldsymbol{W}$ from $\boldsymbol{Y}$.

Information Measurement Yet, Shannon's definition of equivocation based on conditional entropy, is inappropriate in the watermarking field. As we now deal with continuous random vectors, $H(\boldsymbol{W})$ and $H(\boldsymbol{W}|\{\boldsymbol{y}_i\}_{i=1}^{N_o})$ do not measure a quantity of information. The physical interpretation of (3) does not hold anymore. This is the reason why we change the information measurement tools.

In statistics, Fisher was the first to introduce the measure of the amount of information supplied by the observations about unknown parameters. In our case, FIM (Fisher Information Matrix) is defined as:

$$\mathrm{FIM}(\theta) = E\psi\psi^T \quad \text{with} \quad \psi = \nabla_\theta \log p_{\mathcal{X}}(\mathcal{Y} - \mathcal{W}(\theta)). \tag{5}$$

θ denotes the unknown parameter vector. In the KMA case, $\theta = (\boldsymbol{u}_1^T \dots \boldsymbol{u}_{N_c}^T)^T$, whereas, in the WOA context, $\theta = (\boldsymbol{u}_1^T \dots \boldsymbol{u}_{N_c}^T \boldsymbol{a}_1^T \dots \boldsymbol{a}_{N_o}^T)^T$.

The Cramér-Rao theorem gives a lower bound of the covariance matrix of an unbiased estimator whenever the FIM is inversible:

$$\mathcal{R}_{\hat{\theta}} \geq \mathrm{FIM}(\theta)^{-1}, \tag{6}$$

where $\mathcal{B} \geq \mathcal{C}$ means that $\mathcal{B} - \mathcal{C}$ is definite non-negative. Equation (6) provides a physical interpretation: the bigger the information leakage, the more accurate the estimation of the secret parameters. In the sequel, we will see that the trace of $\mathrm{FIM}(\theta)^{-1}$ is proportional to ${N_o}^{-1}$ if the N_o observations are statistically independent. We define the security level ${N_o}^\star$ by the slope of the line such that $\mathrm{tr}(\mathrm{FIM}(\theta)^{-1}) = {N_o}^\star / N_o$. The accuracy of the estimation of θ increases significantly when the number of observations increases of ${N_o}^\star$.

2.3 Security Levels

We apply in this section the methodology to the three contexts of attack.

Known Original Attack (KOA) The reader might be surprised that this context deserves any attention. Seemingly, there is no need to attack watermarked content when one has the original version. The pirate does not hack these contents, but his goal is to gain information about the secret key, in order, later on, to hack different pieces of content watermarked with the same key.

Only one carrier: In this case, the opponent has access to $\boldsymbol{x}$ and $\boldsymbol{y} = \boldsymbol{x} + \gamma a(1)\boldsymbol{u}_1$. The game is over with just one observation as a good estimation of the secret carrier is $\widehat{\boldsymbol{u}_1} = (\boldsymbol{y} - \boldsymbol{x})/\|\boldsymbol{y} - \boldsymbol{x}\|$. However, note that it is impossible to disclose $\boldsymbol{u}_1$ up to a sign, as the estimation depends on the sign of $a(1)$.

Several carriers: In this case, the situation is more complicated because the knowledge of $\boldsymbol{w}$ does not directly give the opponent the carriers. Indeed, he observes several instances of $\boldsymbol{d}_j = \boldsymbol{y}_j - \boldsymbol{x}_j = \gamma \sum_{k=1}^{N_c} a_j(k)\boldsymbol{u_k}/\sqrt{N_c}$. And he is interested in guessing the N_c secret carriers $\boldsymbol{u}_k$. However, note that it is impossible to disclose them up to a sign and a permutation of the order.

Theorem 1. *The security level of WSS watermarking schemes against the Known Original Attack is in the order of N_c pairs $\{(\boldsymbol{x}_j, \boldsymbol{y}_j)\}$. However, this attack reveals the secret carriers up to sign and permutation.*

If the goal of the pirate is to remove the watermark signal, then, he has to render whatever watermarked vector $\boldsymbol{y}$ orthogonal to all estimated $\{\hat{\boldsymbol{u}}_k\}$. If his goal is to decode or encode without authorization, he has not enough information. The ambiguity about the sign and order prevents him to decode the hidden symbols. Yet, he notes whether hidden symbols change from a watermarked content to another. Moreover, the accidental knowledge of hidden symbols in few watermarked pieces of content may fix this ambiguity.

Known Message Attack (KMA) In this subsection, the opponent has access to (watermarked signals/hidden messages) pairs: $\{\boldsymbol{y}_j, \boldsymbol{a}_j\}_{j=1}^{N_o}$. For simplicity reason, we assume that each occurrence of random vector $\boldsymbol{X}$ is independently drawn from $\mathcal{N}(\boldsymbol{0}, \sigma_X^2 \mathcal{I}_{N_v})$. The following theoretical derivations can be easily adapted to colored original signals.

The Fisher Information Matrix is, here, equal to

$$\text{FIM} = \frac{\gamma^2}{N_c \sigma_x^2} \mathcal{A}\mathcal{A}^T \otimes \mathcal{I}_{N_v} \overset{N_o \to +\infty}{\longrightarrow} N_o \frac{\gamma^2 \sigma_a^2}{N_c \sigma_x^2} \mathcal{I}_{N_v N_c}\,, \tag{7}$$

where $\otimes$ denotes the Kronecker product and $\mathcal{I}_N$ the identity matrix of size N. The information leakage is linear with the number of observations, and the slope is given by the watermark to original power ratio per carrier $\gamma^2\sigma_a^2/N_c\sigma_x^2$.

Theorem 2. *The security level of WSS watermarking schemes against the Known Message Attack is $N_o{}^\star = N_c\sigma_x^2/\gamma^2\sigma_a^2$ of $\{(\boldsymbol{y}_j, \boldsymbol{a}_j)\}_j$ pairs.*

Watermarked Only Attack (WOA) In this subsection, messages are unknown so that they must be regarded as nuisance parameters. It is well-known that these nuisance parameters usually render estimation less accurate. Moreover, constraints must be added to the estimation problem to remove unidentifiability and singularity of the Fisher Information Matrix. The main rationale of this presentation was used in [18] to give an alternative expression for the bound in the case where the unconstrained problem is unidentifiable. We add $N_c(N_c-1)/2$ constraints: the estimated carriers must be orthonormal.

The Fisher Information Matrix is then equal to

$$\text{FIM} = \frac{N_o \sigma_a^2 \gamma^2}{N_c \sigma_x^2} (\mathcal{U}^\perp \mathcal{U}^{\perp T})^{-1}\,, \tag{8}$$

where $\mathcal{U}^\perp$ is a basis of the complementary space of $\text{Span}(\mathcal{U})$. The information leakage is linear with the number of observations, and the slope is given by the watermark to original power ratio per carrier $\gamma^2\sigma_a^2/N_c\sigma_x^2$.

Theorem 3. *The security level of WSS watermarking schemes against the Watermarked Only Attack is in $N_o{}^\star = N_c\sigma_x^2/\gamma^2\sigma_a^2$ watermarked vectors $\{\boldsymbol{y}_j\}$. However, the secret carriers are revealed up to sign and permutation.*

3 Practical Tools

The main tools come from Blind Source Separation like Principal Component Analysis and Independent Component Analysis. Their principles will be recalled when needed, through the analysis of the three cases KOA, KMA and WOA.

3.1 Known Original Attack (KOA)

Actually, this case is related to the well known problem of signal processing called Blind Source Separation (BSS) with no noise. A lot of papers have already been written on BSS, and we will just recall here its goals and well-known algorithms. The main idea of BSS is that several source signals are linearly mixed, and that only the mixed signals are available. The goals are the reconstruction of the source signals and the identification of the mixing matrix.

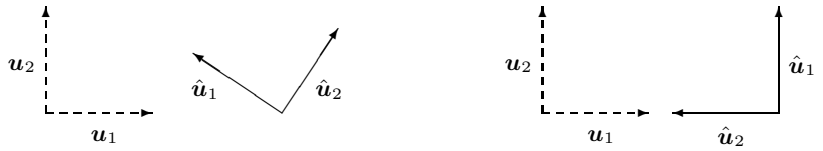
For each observation j, the source signal $\boldsymbol{a}_j$ is linearly mixed by matrix $\mathcal{U}$:

$$\boldsymbol{d}_j = \frac{\gamma}{\sqrt{N_c}}\,\mathcal{U}\boldsymbol{a}_j\,. \tag{9}$$

This system is not unique as for whatever $N_c \times N_c$ invertible matrix $\mathcal{P}$, we have $\boldsymbol{d}_j = \gamma\widetilde{\mathcal{U}}\widetilde{\boldsymbol{a}_j}/\sqrt{N_c}$ with $\widetilde{\mathcal{U}} = \mathcal{U}\mathcal{P}$ and $\widetilde{\boldsymbol{a}_j} = \mathcal{P}^{-1}\boldsymbol{a}_j$. However, the mixing matrix is composed of orthonormal vectors: $\mathcal{U}^T\mathcal{U} = \mathcal{I}_{N_c}$. Thus, the system is now be determined, up to a unitary matrix $\mathcal{P}$ (i.e. a rotation).

To show how the accumulation of observations reveals the carriers, denote $\mathcal{D} = (\boldsymbol{d}_1 \dots \boldsymbol{d}_{N_o})$. A Gram-Schmidt orthogonalization of vectors $\{\boldsymbol{d}_j\}_{j=1}^{N_o}$ yields ρ orthonormal vectors lying in Span($\mathcal{U}$) (this can also be done through a SVD of $\mathcal{D}\mathcal{D}^T$, as PCA does - see Sect. 3.3), with $\rho \stackrel{\Delta}{=} \text{Rank}\ (\mathcal{A})$. Hence, the decomposition outputs a basis of Span($\mathcal{U}$) if the opponent has observed N_c pairs with linearly independent symbols $\{\boldsymbol{a}_j\}_{j=1}^{N_c}$.

Once a basis of Span($\mathcal{U}$) found, the opponent can focus the attack's noise in this subspace to far more efficiently jam the communication, or to nullify the watermarked signals projection in this subspace to remove the watermark. Yet, the vectors of this basis are not necessarily collinear with the private carriers. This is due to the rotation matrix $\mathcal{P}$ ambiguity mentioned above. The opponent cannot decode as projection of watermarked signals onto his basis gives a mixture of the hidden symbols as illustrated in Fig. 1. The same reason prevents him transmitting information in the hidden channel.



(PCA) (ICA)

Fig. 1. PCA *v.s.* ICA. PCA finds the secret carriers up to a rotation, whereas ICA succeeds to align the estimated carriers $\{\hat{\boldsymbol{u}}_k\}_{k=1}^{\hat{N}_c}$ with $\{\boldsymbol{u}_l\}_{k=1}^{N_c}$ (Here, $N_c = \hat{N}_c = 2$). An ambiguity remains about their order (permutation) and their orientation (sign).

Nevertheless, under the assumption that the symbol vectors are *statistically* independent, the opponent can resort to a more powerful tool: the Independent

Component Analysis (ICA). It is an extension of PCA that 'rotates' the basis until the estimated symbols are independent. This happens when the estimated carriers are collinear with the secret carriers, as illustrated in Fig. 1. For the opponent, ICA reduces the ambiguity from the set of rotation matrices $\mathcal{P}$ to the one of permutations with possible changes of sign matrices. In practice, an ICA algorithm needs $N_o > N_c$ observations to converge. A good tutorial on ICA is [19], and a welcome reference on the links between BSS and ICA is [20].

3.2 Known Message Attack (KMA)

The Maximum Likelihood Estimator (MLE) has been chosen because it converges to the Cramér-Rao bound. The log-likelihood is the logarithm of the probability of observing the data $\{\boldsymbol{y}_j\}_1^{N_o}$ knowing the model:

$$\log L(\mathcal{Y}) = cst - \frac{1}{2\sigma_x^2} \sum_{j=1}^{N_o} \|\boldsymbol{y}_j - \frac{\gamma}{\sqrt{N_c}} \mathcal{U} \boldsymbol{a}_j)\|^2 \,. \tag{10}$$

The MLE can be defined by $\frac{\partial \log L}{\partial \boldsymbol{u}_j} = \mathbf{0}$ for all $j \in \{1, \ldots, N_c\}$ giving $\widehat{\mathcal{U}} = \gamma^{-1} \mathcal{Y} \mathcal{A} (\mathcal{A}\mathcal{A}^T)^{-1}$. The complexity of this estimator is quite small. Assuming that $N_c \ll N_o < N_v$, a rough approximation gives an order of $O({N_v}{N_o}^2 N_c)$ for the matrix multiplications, plus $O({N_c}^3)$ for the inversion of $\mathcal{A}\mathcal{A}^T$.

3.3 Watermark Only Attack (WOA)

This case is similar to BSS with noise which is really harder than the previous ones. The covariance matrix of the observed signals is the following:

$$\mathcal{R}_y = \mathcal{R}_x + \frac{\gamma^2}{N_c} \mathcal{U} E\{\mathcal{A}\mathcal{A}^T\} \mathcal{U}^T = \sigma_X^2 \mathcal{I}_{N_v} + \frac{\sigma_a^2 \gamma^2}{N_c} \mathcal{U}\mathcal{U}^T \,. \tag{11}$$

The PCA algorithm first estimates $\mathcal{R}_y$ by $\mathcal{Y}\mathcal{Y}^T$, and performs its SVD decomposition. A (noisy) estimation of a basis of Span($\mathcal{U}$) is given by the eigenvectors of $\mathcal{R}_y$ related to the N_c biggest eigenvalues. Then, ICA rotates this basis until the decoded symbols look-like statistically independent.

From a complexity point of view, the bottleneck is the SVD of the covariance matrix whose size is $N_v \times N_v$. In practical cases, schemes spread the watermark on very long extracted signals. This prevents the feasibility of the attack, as is.

A first idea, to make the attack work, is to split the extracted vectors in order to process smaller vectors of size ${N_v}' = N_v/p$. Yet, the problem then is to put them back together because the ambiguity about the sign and the order completely messes the pieces. The idea shall be given up.

We design an hybrid strategy, mixing this idea of splitting with the MLE algorithm used in the KMA case. The principle of the attack is resumed in Fig. 2. When the ICA algorithm process one block, it outputs N_c estimated carrier blocks and the estimated symbols. Taking ${N_v}'$ as the biggest size the ICA algorithm can manage (this depends on the available computing power), one has a chance to receive reliable hidden symbols. The pirate can now switch to the KMA context to estimate the whole carriers at a low complexity. Thanks to the Kerkhoff's principle, the decoding process is public. The pirate estimates again the symbols with the estimated carriers. It is likely that this produces a better result than the ICA on small vectors. The iteration of the two last operations is indeed the transcription to our case of the Expectation Maximization algorithm invented by Dempster *et al* [21]. Let us summarize the algorithm:

– **Initialization: ICA algorithm.** Split the extracted vectors by chunks of size $N_v{}'$, so that the ICA algorithm works on pieces. It estimates not only pieces of carriers but also hidden symbols $\hat{\mathcal{A}}(0)$.
– **Iteration: EM algorithm.**
  - Maximization step. From the estimated symbols $\hat{\mathcal{A}}(k)$, the MLE algorithm estimates the carriers: $\hat{\mathcal{U}}(k) = \text{MLE}(\mathcal{Y}, \hat{\mathcal{A}}(k))$.
  - Expectation step. The decoding algorithm gives a new estimation of the hidden symbols: $\hat{\mathcal{A}}(k+1) = \text{Decoder}(\mathcal{Y}, \hat{\mathcal{U}}(k))$.

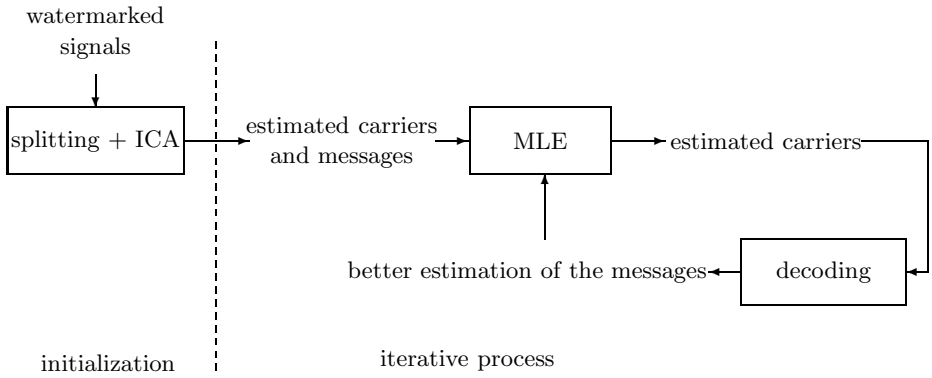


Fig. 2. Final attack for the WOA case.

4 Experimental Works

This section shows experiments about the estimation of the secret carriers with KMA and WOA, and the exploitation of this knowledge to forge pirated images.

4.1 Robust Watermarking

We have chosen a robust watermarking technique [12] embedding $N_c = 8$ bits in still images of size 512×512. It spreads the watermark signal on $N_v = 205008$ coefficients in the wavelet domain. Wavelet coefficients are modeled as independent random variables having their own distribution $\mathcal{N}(0, \sigma^2_{X_i})$. The watermark amplitude factor is proportional to this variance: $\gamma_j(i) = G_j \sigma_{X_{i,j}}$. G_j is set for each image in order to fulfill a distortion constraint expressed by PSNR in dB (set to 38 dB in the experiments).

4.2 Adaptation to Real Images

We need to adapt the estimators that are based on the too simple model of Sect. 2.1. Note that normalized coefficient $y'_j(i) = y_j(i)/\sigma_{X_{i,j}}$ is distributed as $\mathcal{N}(G_j w_j(i), 1)$. The rewriting of the likelihood of $\mathcal{Y}'$ shows that $\boldsymbol{y}_j$ must be weighted by $G_j/1+G_j^2$. The opponent does not know G_j, but he estimates it with the variances $\hat{\sigma}_{X_{i,j}}$. Algorithms are run with these weighted vectors.

4.3 Secret Carriers Estimation

We think that it is more natural for watermarkers to measure the efficiency of the attack by a normalized correlation of estimations with the secret carriers, rather than by a mean square error power (as the Cramér-Rao theorem would recommend). Hence, the criteria is defined as $\eta = \mathrm{tr}\ (\mathcal{U}^T\hat{\mathcal{U}})/N_c$. For this purpose, the estimated carriers are normalized. Moreover, the sign and order ambiguity is automatically removed before measuring the efficiency (we know the secret carriers during the simulations but, of course, a pirate can not do this in real life). Fig. 3 shows the experimental results.

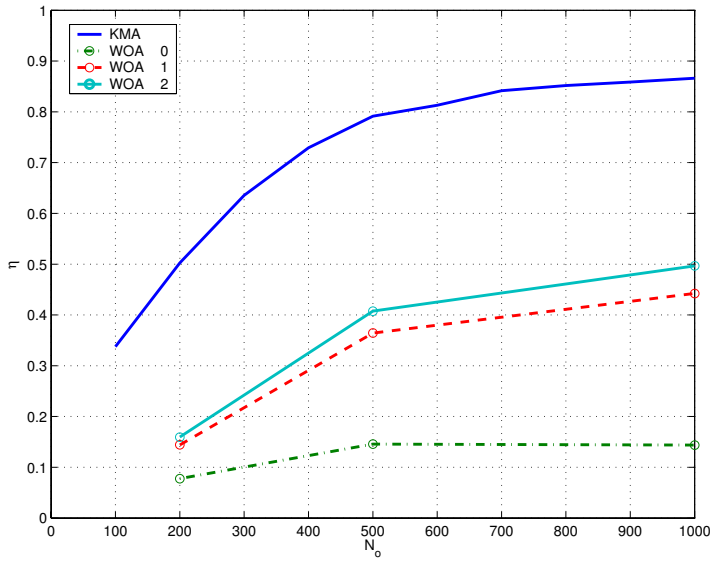


Fig. 3. Mean normalized correlation η between the estimated carriers and the secret ones as the number of observations increases. With circles, correlations with $\hat{\mathcal{U}}(0)$, $\hat{\mathcal{U}}(1)$, and $\hat{\mathcal{U}}(2)$ (see EM algorithm in Sect. 3.3). The WOA EM-algorithm is initialized with the FastICA algorithm [22] on $N_v{}' = 2048$.

4.4 Hacking Content

Fifty other 512×512 images were watermarked. Two opponents try to pirate them. They succeed if the decoded message is not equal to the hidden one (even if just one bit is different). Pirate A uses a blind attack (i.e. pertaining to *robustness*). He scales the size of the images by 1/4, compresses with JPEG at quality factor Q, and he scales them back to the original size. Pirate B uses the following hack (i.e. pertaining to *security*). He has estimated the private carriers (KMA or WOA contexts). For each image, he estimates the hidden message and he tries to flip one bit. The first step is to find the carrier leading to the lowest correlation in absolute value:

$$k^\star = \underset{1 \le k \le N_c}{\arg\min} \left| \hat{\boldsymbol{u}}_k^T \boldsymbol{y} \right| . \tag{12}$$

This maximizes the chance of flipping the corresponding bit at the lowest distortion. The second step is the alteration of the corresponding bit. The attacked vector $\boldsymbol{z}$ is

formed as follows:

$$z(i) = y(i) - G_{hack}.\sigma_{X_i}.\text{sign}(\hat{\boldsymbol{u}}_{k^\star}^T \boldsymbol{y}).\hat{u}_{k^\star}(i) \quad \forall i \in \{1 \dots N_v\}\,, \tag{13}$$

and the inversion extraction function concludes the hack.

Three contexts have been tested: KMA with $N_o = 100$ image/message pairs ($\eta \sim 0.3$), WOA with $N_o = 1000$ images ($\eta \sim 0.5$), and KMA with $N_o = 4500$ image/message pairs ($\eta \sim 0.9$). To compare the two strategies, we measure the probability of success (i.e. the Message Error Rate - MER) against the attack distortion between original and pirated content. For this purpose, pirate A decreases quality factor Q of the JPEG compression and pirate B increases parameter G_{hack}. Figure 4 clearly shows the power of smart attacks. They need a far smaller dis-

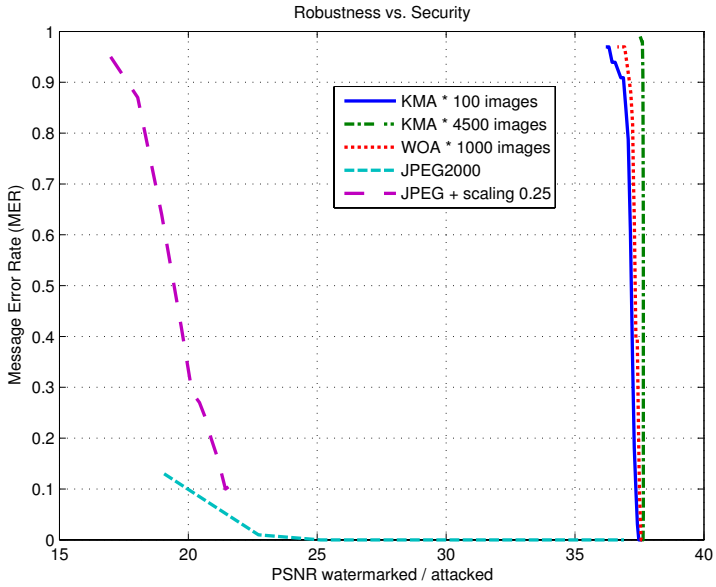


Fig. 4. MER against the attack distortion - PSNR in dB.

tortion budget than the blind attack (a difference of 15 dB!). In our experiment, pirate A's images are so damaged that any exploitation is impossible, as illustrated by Fig. 5. Indeed, we selected in purpose such a robust technique to better illustrate the danger of information leakages. Moreover, the slope of the MER/distortion characteristics of smart attacks is very high. It means that pirate B can really trust in his attack, whereas pirate A is never sure he succeeded until the decoding process happens.

5 Conclusion

This article is an illustration of the recent theory about watermark security. Practical tools from the BSS community help us creating estimators for the KOA, KMA and WOA contexts. However, a double adaptation was necessary. First, real images require a more complex statistical model than the one used in BSS and in the theoretical study of the security levels. Secondly, an ICA algorithm cannot be used as is because it is too complex for such long signals. This is the reason why we

(a) Pirate A. Best quality for a successful attack: PSNR=21.8 dB.

(b) Pirate B. Best quality for a successful hack: PSNR=35.8 dB.

Fig. 5. Comparison between the two pirated Lena images.

develop an EM-like algorithm. However, ICA, working on small pieces of extracted vectors, is necessary to initialize the process. Figure 4 is key fact of the paper. It shows that a robust WSS watermarking technique might be secure iff the embedder changes the secret key for each image. As soon as one secret key is used to watermark several images, there exist information leakages imperilling the security of the watermarking primitive.

References

1. Petitcolas, F., Steinebach, M., Raynal, F., Dittmann, J., Fontaine, C., Fates, N.: Public automated web-based evaluation service for watermarking schemes: Stirmark benchmark. In: IS&T/SPIE International Symposium on Electronic Imaging 2001. Volume 4314. (2001) 575–584 Security and Watermarking of Multimedia Contents III.
2. Kalker, T.: Considerations on watermarking seccurity. In: Proc of the IEEE Multimedia Signal Processing workshop, Cannes, France (2001) 201–206
3. Diffie, W., Hellman, M.: New directions in cryptography. IEEE Trans. on information theory **22** (1976) 644–54
4. Shannon, C.: Communication theory of secrecy systems. Bell system technical journal **28** (1949) 656–715
5. Moulin, P.: The role of information theory in watermarking and its application to image watermarking. Signal Processing **81** (2001) 1121–1139
6. Cox, I., Miller, M., Bloom, J.: Principles and Practice. Morgan Kaufmann Publisher (2001)
7. Cayre, F., Fontaine, C., Furon, T.: Watermarking security: Theory and Practice (2004) accepted to IEEE transactions of Signal Processing.
8. González-Serrano, F., Murillo-Fuentes, J.: Independent component analysis applied to image watermarking. In: ICASSP'01. (2001)
9. Bounkong, S., Toch, B., Saad, D., Lowe, D.: ICA for watermarking digital images. Journal of Machine Learning Research **1** (2002) 1–25
10. Du, J., Lee, C.H., Lee, H.K., Suh, Y.: Watermark attack based on blind estimation without priors. In: IWDW 2002. Lecture Notes in Computer Science, Springer-Verlag (2002)
11. Doërr, G., Dugelay, J.L.: Danger of low-dimensional watermarking subspaces. In: Proc. ICASSP. Volume 3., Montrea, Canada, IEEE (2004)

12. Pateux, S., Guelvouit, G.L.: Practical watermarking scheme based on wide spread spectrum and game theory. Signal Processing: Image Communication **18** (2003) 283–296
13. J.Eggers, Baüml, R., Tzschoppe, R., B.Girod: Scalar costa scheme for information embedding. IEEE Trans. on Signal Processing **51** (2003) 1003–1019
14. Chen, B., Wornell, G.: Quantization index modulation: A class of provably good methods for digital watermarking and information embedding. IEEE Trans. on Information Theory **47** (2001) 1423–1443
15. Barni, M., Bartolini, F., Furon, T.: A general framework for robust watermarking security. Signal Processing **83** (2003) 2069–2084 Special issue on Security of Data Hiding Technologies, invited paper.
16. Furon, T., Duhamel, P.: An asymmetric watermarking method. IEEE Trans. on Signal Processing **51** (2003) 981–995 special issue on signal processing for data hiding in digital media & secure content delivery.
17. Kerckhoffs, A.: La cryptographie militaire. Journal des sciences militaires **9** (1883) 5–38
18. Stoica, P., Ng, B.: On the cramer-rao bound under parametric constraints. IEEE Signal Processing Letters **5** (1998) 177–179
19. Hyvärinen, A., Oja, E.: Independent component analysis: a tutorial. Neural Networks **13** (2000) 411–430
20. Douglas, S.: Blind source separation and independent component analysis: a crossroad of tools and ideas. In: Proceedings of Fourth International Symposium on Independent Component Analysis and Blind Signal Separation, ICA2003. (2003)
21. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the em algorithm. J. Roy. Stat. Soc. (1977) 1–38
22. Hyvärinen, A.: Fast and robust fixed-point algorithms for independent component analysis. IEEE Transactions on Neural Networks **10** (1999) 626–634