Watermarking security part II: practice

François Cayre, Caroline Fontaine, Teddy Furon

Watermarking Security Part Two: Practice

François Cayre[a], Caroline Fontaine[b], and Teddy Furon[a] *

[a] INRIA, TEMICS project, Rennes, France
[b] CNRS, LIFL, Université des sciences et des technologies de Lille, France

ABSTRACT

This second part focuses on estimation of secret parameters of some practical watermarking techniques. The first part reveals some theoretical bounds of information leakage about secret keys from observations. However, as usual in information theory, nothing has been said about practical algorithms which pirates use in real life application. Whereas Part One deals with the necessary number of observations to disclose secret keys (see definitions of security levels), this part focuses on the complexity or the computing power of practical estimators. Again, we are inspired here by the work of Shannon as in his famous article [15], he has already made a clear cut between the *unicity distance* and the *work* of opponents' algorithm. Our experimental work also illustrates how Blind Source Separation (especially Independent Component Analysis) algorithms help the opponent exploiting this information leakage to disclose the secret carriers in the spread spectrum case. Simulations assess the security levels theoretically derived in Part One.

Keywords: W atermarking, Security, Blind source separation.

1. ATTACK ALGORITHMS FOR SPREAD SPECTRUM BASED TECHNIQUES

Part one not only gives security levels of the substitutive method, but also contains almost practical implementations of workable algorithms. On the contrary, it only presents theoretical assessment of security levels for spread spectrum based techniques. Hence, this section deals with practical algorithms useful to hack these schemes. For each attack, an algorithm is presented, and tested on synthetic data as supposed by the model of (I-13), with BPSK symbols and gaussian host vectors. At the end of the section, these algorithms are applied on spread transform side information methods and one still image technique.

This section has an intensive use of PCA and ICA algorithms, which is completely new in watermarking security analysis, as the only other papers mentioning PCA/ICA in the watermarking community have different purposes. [32] and [33] used ICA to design a watermarking embedder. [34] 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, whereas the approach here allows a complete access to the watermarking communication channel to remove, read or write hidden data †.

The following average normalized correlation measures the efficiency of our attack:

$$\eta = \frac{1}{N_c} \sum_{\ell=1}^{N_c} \frac{\hat{\mathbf{u}}_\ell^T \mathbf{u}_\ell}{\|\hat{\mathbf{u}}_\ell\|}. \tag{1}$$

Although the normalization renders estimators $\hat{\mathbf{u}}_j / \|\hat{\mathbf{u}}_j\|$ biased [36], the normalized correlation is preferred because it is an extremely popular measure in the watermarking community. $\eta \lesssim 1$ means that the opponent discloses vectors almost collinear with the secret carriers. When existing, we manually removed the ambiguity of the signed permutation. Measures of η are done averaging $N_t = 128$ experimental results.

* Author names appear in alphabetical order. Contact Information: teddy.furon@irisa.fr
The work described in this paper has been supported in part by the French Government through the ACI Fabriano, and by the European Commission through the IST Programme under Contract IST-2002-507932 ECRYPT.

† We discovered after submission a similar approach uniquely devoted to watermark removal and only based on PCA in [35].

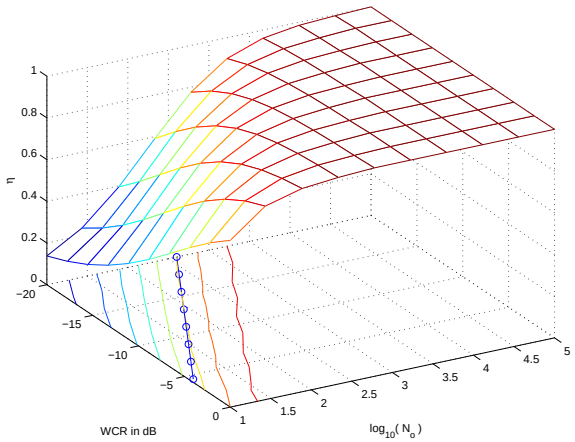

Figure 1. KMA for DSSS ($N_c = 4$, $N_v = 512$). η against $\log_{10}(N_o)$ and WCR in dB. The curve $N_o = N_c\sigma_x^2/\gamma^2$ is plotted with small circles.

The relation with the theoretical security levels is not difficult to find out. (1) is in expectation the cosine of the angle between $\mathbf{u}_\ell$ and $\hat{\mathbf{u}}_\ell = \mathbf{u}_\ell + \mathbf{n}$, $\mathbf{n}$ being the estimation noise (orthogonal to $\mathbf{u}_\ell$ and whose norm is $\sqrt{\mathrm{tr}(\mathrm{CRB}(\mathrm{Vect}(\mathcal{U})))/N_c}$, with $\mathrm{tr}(A)$ the trace of matrix A.) The following relation holds:

$$\eta \approx \frac{\|\mathbf{u}_\ell\|}{\sqrt{\|\mathbf{u}_\ell\|^2 + \mathrm{tr}(\mathrm{CRB}(\mathrm{Vect}(\mathcal{U})))/N_c}}. \tag{2}$$

1.1. Known Message Attack

Observing $(\mathbf{y}, \mathbf{a})^{N_o}$, the opponent can use the Maximum Likelihood Estimator (MLE) related to (I-14). This estimator is also defined by $\frac{\partial \log L}{\partial \mathbf{u}_\ell} = \mathbf{0} \quad \forall \ell \in \{1, \ldots, N_c\}$. With the following log-likelihood

$$\log L(\mathcal{Y}) = const - \frac{1}{2\sigma_x^2} \sum_{j=1}^{N_o} \|\vec{y}_j - \frac{\gamma}{\sqrt{N_c}} \mathcal{U}\vec{a}_j)\|^2 , \tag{3}$$

this gives:

$$\hat{\mathcal{U}} = \frac{\sqrt{N_c}}{\gamma} (\mathcal{Y}\mathcal{A}^T)(\mathcal{A}\mathcal{A}^T)^{-1}. \tag{4}$$

The MLE is known to be unbiased and consistent, *i.e.* it asymptotically achieves the CRB derived in Part One. Fig. (1) shows experimental values of η against N_o and WCR $= \gamma^2/\sigma_x^2$ for the DSSS case. The locus of points such that $\eta = const$ are projected on the plane $\eta = 0$. They appear to be parallel with the curve $N_o = N_c\sigma_x^2/\gamma^2$. Tests done with different N_v confirm that the efficiency of the attack does not depend on the vector length. This asserts the theoretical security level given Part One.

The complexity of this estimator is quite small. A rough approximation gives an order of $O(N_v N_c N_o)$ for the matrix multiplications, plus $O({N_c}^3)$ for the inversion of $\mathcal{A}\mathcal{A}^T$, which is extremely easy to do on nowadays computer for N_c of several hundreds.

1.2. Known Original Attack

In this case, the opponent observes several instances of $\mathbf{d}_j = (\mathbf{y}_j - \mathbf{x}_j) \propto \mathcal{U}\mathbf{a}_j$. As mentioned in Part One, this 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 just recall here its most common algorithms. Note that spread spectrum corresponds to the BSS over-determined case (*i.e.*, $N_v \geq N_c$).

The most classical algorithm in BSS is the Principal Component Analysis (PCA). Denote $\mathcal{D} = \mathcal{Y} - \mathcal{X}$. This technique makes an eigendecomposition of the matrix $\mathcal{D}\mathcal{D}^T = \gamma^2\mathcal{U}\mathcal{A}\mathcal{A}^T\mathcal{U}^T/N_c$. This corresponds to a Gram-Schmidt orthogonalization of vectors $\mathbf{d}^{N_o}$. Please, note that $\rho \triangleq \mathrm{Rank}\,(\mathcal{A})$ is also the rank of $\mathcal{D}\mathcal{D}^T$.

Figure 2. PCA *vs.* ICA. PCA finds the secret carriers up to a rotation, whereas ICA succeeds to align the estimated carriers $\hat{\mathbf{u}}^{N_c}$ with $\mathbf{u}^{N_c}$ (Here, $N_c = 2$). An ambiguity remains in their order (permutation) and orientation (sign).

Hence, the decomposition outputs ρ orthonormal vectors lying in Span $(\mathcal{U})$. In the best case, the opponent has $\rho = \min(N_o, N_c)$. Nevertheless, in reality, he may have $\rho \leq \min(N_o, N_c)$ if the N_o symbol vectors are linearly dependent.

When successful (*i.e.*, when $\rho = N_c$), the PCA technique yields a orthonormal basis of the secret subspace Span $(\mathcal{U})$. The possibilities to hack watermarked pieces of content when Span $(\mathcal{U})$ is disclosed are summarized in subsection 1.4. Yet, the vectors of this basis are not necessary collinear with the private carriers. This is due to the unitary matrix $\mathcal{P}$ mentioned in subsection 4.4 of Part One. The opponent cannot decode, as projection of watermarked signals onto this basis gives a mixture of the hidden symbols. This is illustrated by Fig. 2. The same reason prevents him transmitting information in the hidden channel. 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, constraining the output estimated symbol vectors to be independent [25]. Good tutorials on ICA and on its links with BSS are [27, 37]. A very general ICA algorithm named FastICA [38] has been preferred to algorithms dedicated to specific symbol distribution as mentioned in Part One [28, 29].

In short, ICA algorithms usually work in the basis recovered by a PCA. This basis describes exactly the secret subspace (provided that $\rho = N_c$). The problem is now reduced to the estimation of the $N_c \times N_c$ matrix $\mathcal{P}$. Hence, parameter N_v has absolutely no influence on the attack. Then, in an iterative process, the ICA 'rotates' the basis until it nullifies an objective function (often called a constrast function) of the estimated sources or symbols $\hat{\mathbf{a}}^{N_o}$. This function is usually an approximation of the mutual information of the estimated sources. We have supposed that the symbols are independent random variable, thus their mutual information is null. We are looking here the matrix $\mathcal{P}$ which renders the estimated sources statistically independent, thus which nullify their mutual information. Contrast functions depend on the distribution of the symbol sources. However, this measure reflects statistical independence only for large N_o. For a finite number of observations, ICA algorithms usually search for a minimum of the contrast function with the help of a gradient descent technique.

When successful, ICA reduces the set of ambiguity matrices $\mathcal{P}$ to the one of signed permutations. This is illustrated by Fig. 2. Subsection 1.4 lists the possibilities to hack watermarked pieces of content when the carriers are disclosed up to a signed permutation.

1.3. Watermarked Only Attack

The WOA case is quite similar to KOA, as it is related to BSS but in a noisy environment. The covariance matrix $\mathcal{R}_{\mathbf{y}}$ has the following expression:

$$\mathcal{R}_{\mathbf{y}} = \mathcal{R}_{\mathbf{x}} + \frac{\gamma^2}{N_c}\mathcal{U}E\{\mathcal{A}\mathcal{A}^T\}\mathcal{U}^T = \mathcal{R}_{\mathbf{x}} + \frac{\gamma^2}{N_c}\mathcal{U}\mathcal{R}_{\mathbf{a}}\mathcal{U}^T = \sigma_x^2\mathcal{I} + \frac{\gamma^2\sigma_a^2}{N_c}\mathcal{U}\mathcal{U}^T. \tag{5}$$

Its diagonalization leads to N_c eigenvalues equaling $\sigma_x^2 + \gamma^2\sigma_a^2/N_c$, and $N_v - N_c$ eigenvalues equaling σ_x^2. Hence, the eigenvectors related to the N_c biggest values constitute a basis of Span $(\mathcal{U})$, which is also known as the signal space in blind equalization for digital communications.

PCA estimates covariance matrix $\mathcal{R}_{\mathbf{y}}$ by $\mathcal{Y}\mathcal{Y}^T/N_o$, and outputs N_c eigenvectors whose eigenvalues are the biggest ones. Due to this rough estimation, these vectors do not live exactly in Span $(\mathcal{U})$. Compared to Fig. 2, these noisy estimation vectors would not lie in the plan of the page, regarded as subspace Span $(\mathcal{U})$ in this simple example. However, ICA will still try to rotate them in order to render the decoded symbols independent. Fig. 3

shows the locus of points such that $\eta = const$ for different values of N_c and N_o, with the DSSS method (*i.e.*, a BPSK modulation). The ICA algorithm meets the theoretical limit only for large N_o, and high energy of watermark signal per carrier: $\gamma^2 N_v/N_c$. Note that, for $N_c = 4$, the gap between experimental performances and theoretical limit gets larger.

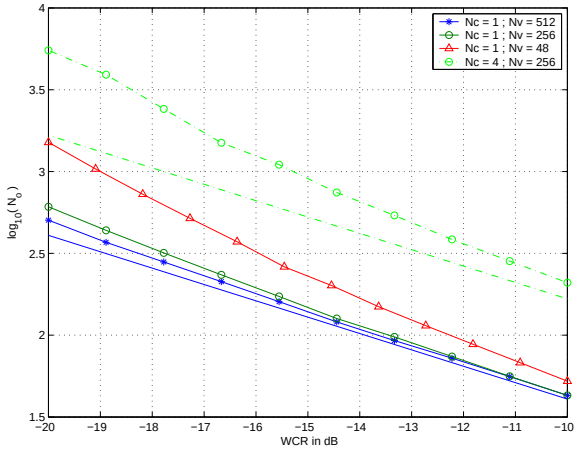

Figure 3. WOA for DSSS. Operating points achieving $\eta = 0.8$ for different parameters N_c and N_v. The solid line is the theoretical limit for $N_c = 1$, and curves with stars, circles and triangles are the experimental results. They capture the efficiency of the PCA, as only one carrier is used. The dashed line is the theoretical limit for $N_c = 4$ (*i.e.* the solid line translated of $\log_{10}(N_c)$), the dashed curve with circles is the experimental results with the FastICA algorithm [38].

From a complexity point of view, the bottleneck is the decomposition in eigenvectors 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 into p chunks 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. 4. 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* [39]. Let us summarize the algorithm:

- **Initialization: ICA algorithm.** Extract from the extracted vectors 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: $\hat{\mathcal{A}}(k+1) = \text{Decoder}(\mathcal{Y}, \hat{\mathcal{U}}(k))$.

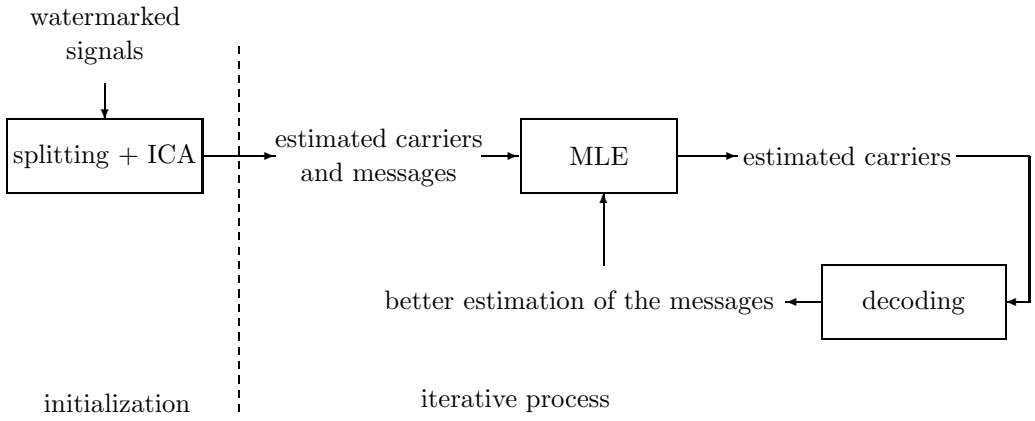

Figure 4. Final attack for the WOA case.

1.4. Possible hacks

The conclusion of this security analysis stands in the different possibilities to forge pirated content.

- The pirate discloses secret subspace Span (U). He can now focus attack's noise in this subspace to jam the communication far more efficiently. He can also nullify the watermarked signals projection in this subspace to remove the watermark.
- The pirate discloses the secret carriers up to a signed permutation. The above-mentioned hacks are still possible. Besides, he can detect whether two watermarked pieces of content share the same hidden message. He can also flip some randomly chosen bits. Moreover, the accidental knowledge of hidden messages in few watermarked pieces of content might remove this ambiguity. This extra security analysis indeed pertains the substitutive case studied in subsection I-3.4.
- The pirate discloses the secret carriers. He has a full access to the watermarking channel to read, write or erase hidden message.

Of course, the quality of the pirated pieces of content depends on the accuracy of his estimation. The authors focus on this aspect in section 2.

1.5. Extension to spread transform side information watermarking

This subsection presents experiments with side information watermarking using the process of spread spectrum. In these methods, the symbols $a_j(\ell)$ depend on the host signal in the following way:

$$a_j(\ell) = f(m_j(\ell), \mathbf{u}_\ell^T \mathbf{x}_j) \tag{6}$$

Three techniques were investigated: Improved Spread Spectrum (ISS) [22], Scalar Costa Scheme (SCS) [20], and Maximized Robustness Embedding (MRE) [21]. Two implementations of SCS have been done. The carriers have disjoint supports in the first one, which is a possible interpretation of [20]: $\mathbf{u}_1 = (\mathbf{u}^T \mathbf{0}_\tau^T \dots \mathbf{0}_\tau^T)^T$, $\mathbf{u}_2 = (\mathbf{0}_\tau^T \mathbf{u}^T \dots \mathbf{0}_\tau^T)^T$, and so on with $\tau N_c = N_v$. The second implementation is called SCS with Subspace Projection (SSP) [40]: the carriers have a full support and are orthonormal. The embedding distortion, the vector length and the number of hidden bits are the same for a fair comparison.

The KMA case has not been investigated. The knowledge of the messages does not usually imply the disclosure of the symbols. In SCS, function $f(.)$ of (6) is private and depends on a secret key (*i.e.*, a dithering vector). However, information about the symbols may leak from the message. Symbols are Gaussian variables centered on $\gamma(-1)^{m_j(\ell)}$ for the ISS technique:

$$a_j(\ell) = \gamma(-1)^{m_j(\ell)} - \lambda \mathbf{u}_\ell^T \mathbf{x}_j. \tag{7}$$

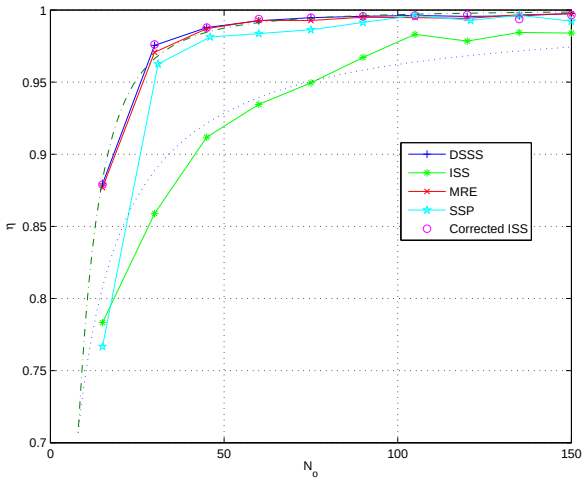

Figure 5. KOA for four different watermarking techniques($N_c = 4$, $N_v = 512$). Dotted line: $\eta = (1 + k/N_o)^{-1}$; Dash-dotted line: $\eta = (1 + (k/N_o)^2)^{-1}$.

We foresee that the MLE algorithm could easily be tuned to exploit this information leakage.

The KOA is simpler, as the basic assumption is still valid: $\mathbf{u}_\ell^T \mathbf{x}_j$ and $\mathbf{u}_k^T \mathbf{x}_j$ ($k \neq \ell$) are Gaussian distributed and non correlated; thus, the symbols are statistically independent. Yet, the efficiency of BSS depends on the symbols distribution, so that we expect different performances. Once again, in our simulation, the opponent always uses the same generic ICA algorithm. No fine tuning according to the expected symbols distribution is done. Fig. (5) shows the results, except for SCS[‡]. Surprisingly, the rate of the noise estimation variance is in ${1/N_o}^2$ for DSSS, SSP and MRE. This seems to be due to the bounded support feature of the symbols in these methods, despite of the use of a generic algorithm. For ISS, the rate is in $1/N_o$. Please, note that, according to (7), the KOA for ISS is similar to a WOA for the SS method, with a watermark to host power ratio of $\gamma^2/\lambda^2\sigma_x^2$. A smarter attack on ISS stems from this remark. First, difference vectors are used to disclose the secret subspace with a PCA. Then, they are corrected in adding the projection of the original vectors scaled by a factor λ. We are now in a situation similar to a KOA with DSSS. Finally, ICA finishes the job working on the corrected vectors. The last curve named 'Corrected ISS' in Fig. (5) shows the dramatic improvement. The security level of ISS is in practice as low as the DSSS one.

The WOA is also straightforward as we applied the same ICA algorithm for DSSS, ISS, MRE, and SSP. For SCS, the observed watermarked vectors are split by chunks of τ samples. Thus, the opponent has ${N_o}' = N_o\tau$ vectors whose length is ${N_v}' = N_v/\tau$, watermarked with ${N_c}' = 1$ secret carrier. The algorithm is thus a simple PCA in this case. Fig. (6) shows the results. SCS (or more precisely the way we have implemented it) is obviously the less secure. But the simple change brought in the implementation of SSP is sufficient to correct this security flaw [§]. The other techniques share the same security level. ISS seems to be slightly more secure; however, remember that we did not tune the contrast function of the ICA algorithm. In the same way, the embedding parameters (γ, λ) play a big role in the symbols distribution, and the attack might thus perform differently. This is the reason why we prefer to look at the global shape of the curves, rather than to draw erroneous conclusions from these meager differences.

2. APPLICATION TO A ROBUST WATERMARKING TECHNIQUE FOR STILL IMAGES

The goal of this last subsection is to demonstrate the power of 'smart' attacks based on secret carriers estimation. To this end, the section deals with real still images. So far, this article has investigated the first phase of the attack, the secret disclosure, on synthetic data. The difficulty here is to adapt the algorithms of KMA and

[‡]For SCS, $N_o = 1$ is enough to disclose small length carrier $\mathbf{u}$ up to a sign.

[§]We only analyze here the security of the spreading transform. Yet, the dithering vector in SCS-like technique constitute a second barrier, which will be the subject of a future work.

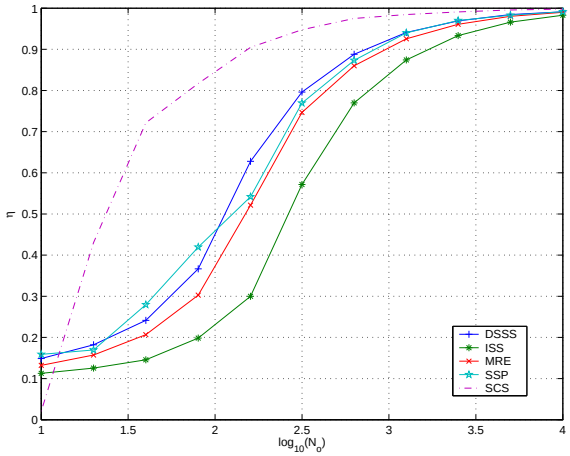

Figure 6. WOA for five different watermarking methods ($N_v = 512$, $N_c = 4$, WCR=-15dB). $\tau = 128$ for SCS. For SCS, SSP and ISS, the embedding parameters are optimal for an expected noise attack whose distortion equals the embedding distortion: WNR=0 dB.

WOAattack to real images. In a second phase, the opponent uses this *a posteriori* information to hack pieces of content, which were watermarked with the same secret key.

2.1. Robust Watermarking

We have chosen a robust watermarking technique [19] 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 $x_j(i)$ are modeled as independent random variables having their own distribution $\mathcal{N}(0, \sigma^2_{X_{i,j}})$.

The watermark amplitude factor $\gamma_j(i)$ is a complex function of the variance $\sigma^2_{X_{i,j}}$, the embedding distortion, and the targeted attack distortion. This is the result of a double Lagrangian optimization in the framework of game theory. Roughly, the amplitude factor follows the general law $\gamma_j(i) = G_j \sigma^{\beta_j}_{X_{i,j}}$, with G_j and β_j optimized for each image.

2.2. Adaptation to Real Images

We need to adapt the estimators that are based on the too simple model of Sect. 1. However, the selected technique is too complex and we prefer to simplify the attack and see how robust (in the statistical sense) it is with respect to more complex model. Therefore, we assume that the watermark amplitude factor is only proportional to the 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).

With this assumption, 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 $\mathbf{y}'_j$ 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.

2.3. 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.

Yet, the WOA attack leads to a sign and order ambiguity which must be 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). The Hungarian algorithm is run on the matrix $|\mathcal{U}^T \hat{\mathcal{U}}|$ (the matrix whose elements are the absolute value of the elements of $\mathcal{U}^T \hat{\mathcal{U}}$). It finds the permutation Π which maximizes the score $\eta = \mathrm{tr}(|\mathcal{U}^T \Pi \hat{\mathcal{U}}|)/N_c$. Fig. 7 shows the experimental results.

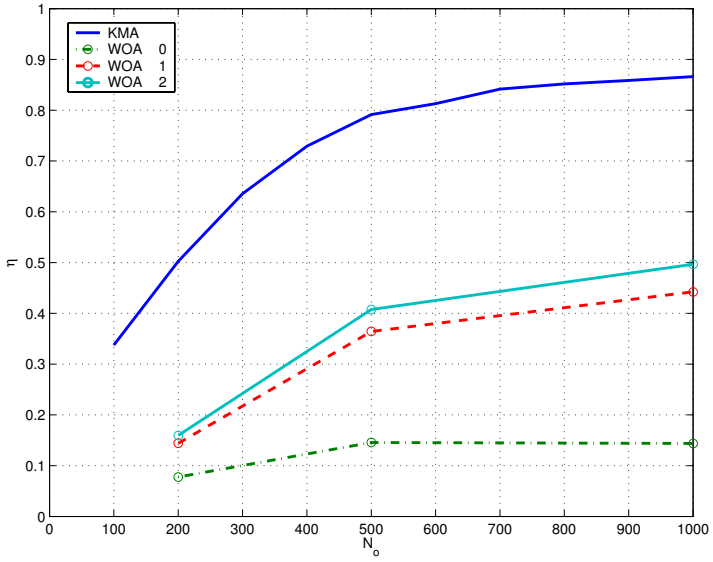

Figure 7. Mean normalized correlation η between the estimated carriers and the secret ones as the number of observations increases. With circles, correlations with $\hat{\mathcal{U}}(0)$, $\hat{\mathcal{U}}(1)$, and $\hat{\mathcal{U}}(2)$ (see EM algorithm in Sect. 1.3). The WOA EM-algorithm is initialized with the FastICA algorithm [38] on ${N_v}' = 2048$.

2.4. 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*). For instance, 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. We have also tested a JPEG2000 compression. Pirate B uses the following *smart* attack (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:

$$k^\star = \underset{1 \leq k \leq N_c}{\arg\min} \, |\hat{\mathbf{u}}_k^T \mathbf{y}| \, . \tag{8}$$

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 $\mathbf{z}$ is formed as follows:

$$z(i) = y(i) - G_{hack} . \sigma_{X_i} . \text{sign}(\hat{\mathbf{u}}_{k^\star}^T \mathbf{y}) . \hat{u}_{k^\star}(i) \quad \forall i \in \{1 \dots N_v\} \, , \tag{9}$$

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 8 clearly shows the power of smart attacks. They need a far smaller distortion 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. 9. 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.

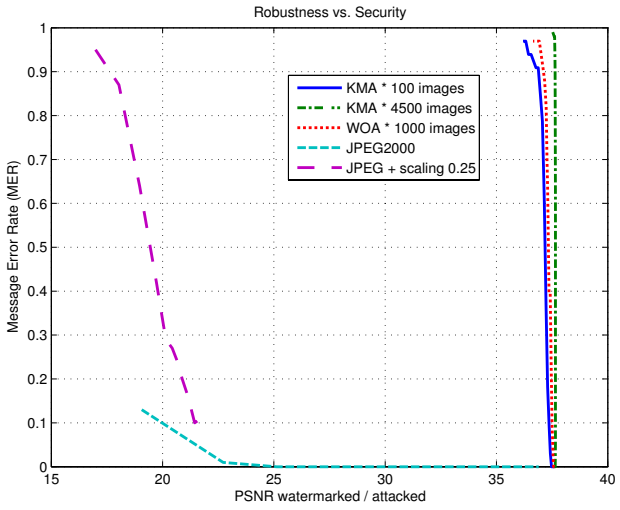

Figure 8. MER against the attack distortion - PSNR in dB.

(a) Pirate A (b) Pirate B

Figure 9. Comparison between the two pirated Lena images. This is their best quality for a successful attack. Pirate A: PSNR=21.8 dB, Pirate B: PSNR=35.8 dB.

3. CONCLUSION

As in cryptanalysis, measurement of information leakages is the fundamental principle underlying the theoretical framework for robust watermarking security assessment presented in this article. A watermarking technique, even robust, is not secure if the opponent can refine his knowledge on the presumably secret key while pieces of content are watermarked with the same key. The security level is then defined by the number of observations the opponent needs in order to accurately estimate the secret key.

The conclusion of this article is not that spread spectrum based watermarking techniques or substitutive schemes are broken. The goal is to warn the watermarking community that security is a crucial issue. Designers should not only control the imperceptibility and the robustness of their schemes but also assess their security levels. Depending on the application designers are targeting (and especially on the observations available to the pirate), watermarking several pieces of content with the same key might bring threats. This potentially arises difficulties on the key management. For instance, it is not clear how a blind watermarking decoder will be

informed of the secret key, if this later one is to be changed on a regular basis according to the security levels assessed in this article.

REFERENCES

1. I. Cox, M. Miller, and J. Bloom, *Principles and Practice*, Morgan Kaufmann Publisher, 2001.
2. J. O'Ruanaidh and T. Pun, "Rotation, scale and translation invariant spread spectrum digital image watermarking," *Signal Processing*, vol. 66, no. 3, pp. 303–17, 1998.
3. I. Cox, M. Miller, and A. McKellips, "Watermarking as communication with side information," *Proc. IEEE*, vol. 87(7), pp. 1127–1141, July 1999.
4. B. Chen and G. Wornell, "Quantization index modulation: A class of provably good methods for digital watermarking and information embedding," *IEEE Trans. Inform. Theory*, vol. 47, pp. 1423–1443, May 2001.
5. P. Moulin, "The role of information theory in watermarking and its application to image watermarking," *Signal Processing*, vol. 81, pp. 1121–1139, 2001.
6. T. Kalker, "Considerations on watermarking security," in *Proc. MMSP*, Cannes, France, Oct. 2001, pp. 201–206.
7. S. Craver, N. Memon, B.-L. Yeo, and M.M. Yeung, "On the invertibility of invisible watermarking technique," in *Proc. ICIP*, Washington, DC, USA, Oct. 1997, IEEE, pp. 540–543.
8. M. Kutter, S. Voloshynovskiy, and A. Herrigel, "Watermark copy attack," in *Security and Watermarking of Multimedia Contents II*, P.W. Wong and E. Delp, Eds., San Jose, Cal., USA, Jan. 2000, vol. 3971.
9. I. Cox and J.-P. Linnartz, "Some general methods for tampering with watermarks," *IEEE J. Select. Areas Commun.*, vol. 16, no. 4, pp. 587–93, May 1998, Special issue on copyright and privacy protection.
10. J.P. Linnartz and M. van Dijk, "Analysis of the sensitivity attack against electronic watermarks in images," in *Proc. IHW*, D. Aucsmith, Ed., Portland, Oregon, USA, Apr. 1998, vol. 1525 of *Lecture Notes in Computer Science*, Springer Verlag.
11. T. Furon and P. Duhamel, "An asymmetric watermarking method," *in [41]*, pp. 981–995.
12. T. Mittelholzer, "An information-theoritic approach to steganography and watermarking," in *Proc. IHW*, A. Pfitzmann, Ed., Dresden, Germany, Sept. 1999, pp. 1–17, Springer Verlag.
13. M. Barni, F. Bartolini, and T. Furon, "A general framework for robust watermarking security," *Signal Processing*, vol. 83, no. 10, pp. 2069–2084, Oct. 2003, Special issue on Security of Data Hiding Technologies, invited paper.
14. A. Kerckhoffs, "La cryptographie militaire," *Journal des sciences militaires*, vol. 9, pp. 5–38, janvier 1883.
15. C.E. Shannon, "Communication theory of secrecy systems," *Bell system technical journal*, vol. 28, pp. 656–715, Oct. 1949.
16. W. Diffie and M. Hellman, "New directions in cryptography," *IEEE Trans. Inform. Theory*, vol. 22, no. 6, pp. 644–54, Nov. 1976.
17. S. Burgett, E. Koch, and J. Zhao, "Copyright labelling of digitized image data," *IEEE Commun. Mag.*, vol. 36, no. 3, pp. 94–100, Mar. 1998.
18. D. Kahn, "Cryptology and the origins of spread spectrum," *IEEE Spectr.*, pp. 70–80, Sept. 1984.
19. S. Pateux and G. Le Guelvouit, "Practical watermarking scheme based on wide spread spectrum and game theory," *Signal Processing: Image Communication*, vol. 18, pp. 283–296, Apr. 2003.
20. J.Eggers, R. Baüml, R. Tzschoppe, and B.Girod, "Scalar costa scheme for information embedding," *in [41]*, pp. 1003–1019.
21. M. Miller, I. Cox, and J. Bloom, "Informed embedding: exploiting image and detector information during watermark insertion," in *Proc. ICIP*, Vancouver, Canada, Sept. 2000.
22. H.S. Malvar and D.A.F. Florêncio, "Improved spread spectrum: A new modulation technique for robust watermarking," *in [41]*, pp. 868–905.
23. D.T. Pham and J.F. Cardoso, "Blind separation of instantaneous mixtures of non stationary sources," *IEEE Trans. Signal Processing*, vol. 49, no. 9, pp. 1837–1848, 2001.

24. J. Su, J. Eggers, and B. Girod, "Analysis of digital watermarks subjected to optimum linear filtering and additive noise," *Signal processing*, vol. 81, pp. 1141–1175, 2001.
25. P. Comon, "Independent component analysis, a new concept?," *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.
26. S.-I. Amari and J.F. Cardoso, "Blind source separation; semiparametric statistical approach," *IEEE Trans. Signal Processing*, vol. 45, no. 11, 1997, Special issue on neural networks.
27. J.-F. Cardoso, "Blind signal separation: statistical principles," *Proc. IEEE*, vol. 86, no. 10, pp. 2009–2025, Oct. 1998.
28. A.-J. van der Veen, "Blind separation of BPSK sources with residual carriers," *Signal Processing*, vol. 73, no. 10, pp. 67–79, Jan. 1999.
29. F. Gamboa and E. Gassiat, "Source separation when the input sources are discrete or have constant modulus," *IEEE Trans. Signal Processing*, vol. 45, no. 12, pp. 3062–3072, Dec. 1997.
30. P. Stoica and B.C. Ng, "On the Cramér-Rao bound under parametric constraints," *IEEE Signal Processing Lett.*, vol. 5, no. 7, pp. 177–179, 1998.
31. Y. Yao and G.B. Giannakis, "On regularity and identifiability of blind source separation under constant-modulus constraints," *IEEE Trans. Signal Processing*, 2004, To appear.
32. F.J. González-Serrano and J.J. Murillo-Fuentes, "Independent component analysis applied to image watermarking," in *Proc. ICASSP*, 2001.
33. S. Bounkong, B. Toch, D. Saad, and D. Lowe, "ICA for watermarking digital images," *Journal of Machine Learning Research*, vol. 1, pp. 1–25, 2002.
34. J. Du, C.-H. Lee, H.-K. Lee, and Y. Suh, "Watermark attack based on blind estimation without priors," in *Proc. IWDW*. 2002, Lecture Notes in Computer Science, Springer-Verlag.
35. G. Doërr and J.-L. Dugelay, "Danger of low-dimensional watermarking subspaces," in *Proc. ICASSP*, Montreal, Canada, may 2004, vol. 3.
36. P. Stoica and B. Ng, *Signal Processing Advances in Wireless and Mobile Communications*, vol. 1, chapter Performance Bounds for Blind Channel Estimation, pp. 41–62, Prentice Hall, 2001.
37. A. Hyvärinen and E. Oja, "Independent component analysis: a tutorial," *Neural Networks*, vol. 13, no. 4-5, pp. 411–430, 2000.
38. A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.
39. A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *J. Roy. Stat. Soc.*, , no. 39, pp. 1–38, 1977.
40. R. Fischer, R. Tzschoppe, and R. Bäuml, "Lattice costa schemes using subspace projection for digital watermarking," *European Trans. Telecommunications*, vol. 15, no. 4, pp. 351–362, Aug. 2004.
41. A. Akansu, E. Delp, T. Kalker, B. Liu, N. Memon, P. Moulin, and A. Tewfik, "Special issue on signal processing for data hiding in digital media and secure content delivery," *IEEE Trans. Signal Processing*, vol. 51, no. 4, Apr. 2003.