Comparison of secure spread-spectrum modulations applied to still image watermarking

Benjamin Mathon, Patrick Bas, François Cayre, Benoît Macq

Ann. Telecommun.
DOI 10.1007/s12243-009-0119-9

Comparison of secure spread-spectrum modulations applied to still image watermarking

Benjamin Mathon · Patrick Bas · François Cayre · Benoît Macq

Received: 11 June 2008 / Accepted: 4 June 2009

Abstract This article shows the results obtained when using secure spread-spectrum watermarking on gray-scale images in the watermark only attack (WOA) framework. Two secure modulations, natural watermarking (NW) and circular watermarking (CW), are compared with classical insecure modulations, spread spectrum (SS) and improved spread spectrum (ISS), from distortion, robustness, and security points of view. Implementations of CW and NW for still images are proposed: they use a wavelet transform and variable strength embedding with bounded distortion. Robustness of these schemes is assured by using JPEG compression and security is quantified by using a source separation technique: independent component analysis (ICA). Finally, tests are conducted on 2,000 natural images. They allow to distinguish between WOA security classes.

B. Mathon (✉) · P. Bas · F. Cayre
GIPSA Lab - UMR CNRS 5216, 961 rue de la Houille Blanche, Domaine universitaire - BP 46,
38402 Saint Martin d'Hères cedex, France
e-mail: benjamin.mathon@gipsa-lab.grenoble-inp.fr

P. Bas
e-mail: patrick.bas@gipsa-lab.grenoble-inp.fr

F. Cayre
e-mail: francois.cayre@gipsa-lab.grenoble-inp.fr

B. Mathon · B. Macq
TELE - Laboratoire de télécommunications et télédétection,
Bâtiment Stévin, Place du Levant, 2,
1348 Louvain-la-Neuve, Belgium

B. Macq
e-mail: benoit.macq@uclouvain.be

Keywords Watermarking · Security · Still images

1 Introduction

The problem of security for watermarking applications has recently become a major concern for both researchers and industrialists. For specific setups, a security flaw can considerably reduce the usability of a watermarking scheme. In [1], the authors propose a classification of attacking scenarios for watermarking depending on the kind of knowledge of the adversary. Following the Diffie–Hellman classification, they came up with various scenarios, among which is the watermarked only attack (WOA) framework, where the adversary only owns several watermarked images (he knows neither embedded messages nor original images). Also in watermarking, security is based on Kerckhoffs' principle [2] and relates to the estimation of a part or all of the secret key [3]. In multi-bit watermarking, the key is the location of a set of codewords in a private subspace. Under the assumption that digital contents are represented by their feature vectors, embedding of a message in a host content consists of moving the feature vector into the decoding region of the right codeword. From [4], one can devise four embedding security classes in the WOA framework:

1. *Insecurity.* In this class, the conditional distribution (given the key) of the marked contents is different for all keys. By exhaustive search (or a more advanced technique), the adversary can estimate both the private subspace and the codewords.
2. *Key security.* A watermarking scheme belongs to the second class, *key security*, if, for a subset of keys

Published online: 01 July 2009

(including the actual secret key), the conditional distribution of the marked contents is the same. The adversary can find this subset of keys and he can find the secret subspace but he cannot obtain more information about the codewords.

3. *Subspace security.* The third class is called *subspace security*. In this case, the conditional distribution of marked contents is the same for all possible keys. Therefore, the adversary cannot gain any information about the key (he has no access to the secret subspace where the keys actually live).
4. *Stego security.* This last class relates to steganography: the distribution of marked contents is the same as the distribution of the host's contents. The adversary cannot decide whether the contents are marked or not.

The secret part of a watermarking scheme can be seen as twofold: on the one hand, there is the very private subspace in which the decoding regions are located, and on the other hand, there is the location of these decoding regions in the private subspace. In practice, the key-secure class determines the limitation between robustness and security: the optimal attack consists of focusing on the private subspace in order to minimize the attack distortion. However, the estimation of the location of the coding regions is more precise information and can be used to tamper the embedded message. We have the following relationships among the classes: *stego security* $\subset$ *subspace security* $\subset$ *key security* and *key security* $\cap$ *insecurity* $= \emptyset$. In [1], the authors propose the notion of security level for insecure watermarking (number of observations that are necessary to improve the adversary's knowledge about the secret key by an order of magnitude). Classical spread-spectrum modulations, such as spread spectrum (SS) [5] and improved spread spectrum (ISS) [6], are proved to be in the insecurity class. Recently, two secure modulations have been proposed in the WOA framework: the natural watermarking (NW), which can be made stego-secure, subspace-secure, or key-secure, and the circular watermarking (CW), which is key-secure. The main goal of this article is to assess and to compare the security of the four previous modulations in a practical case: image watermarking. Section 2 lists notational conventions used in the article. Section 3 recalls basics on SS watermarking schemes (insecure and secure modulations). Section 4 considers two points of view, namely distortion and robustness, to compare insecure and secure modulations over natural images and presents improvements regarding psychovisual constraints. Section 5 deals with the security of the SS schemes. In this section, we present techniques used to estimate the secret key and the different levels of security of the schemes. Key estimation is cast into a blind source separation problem where independent component analysis is used to find codeword locations (or, conversely, decoding regions in the private subspace, when possible). Also in this section, we compare the security of the modulations in the practical case of image watermarking.

2 Notations

We first list some notational conventions used in this article. Data are written in small letters. Vectors and matrices are set in bold fonts. Vectors are written in small letters and matrices in capital ones. $\mathbf{x}(i)$ is the i-th component of vector $\mathbf{x}$. As for the C programming language, all indexes start from zero. We write $(\mathbf{x}(0), \mathbf{x}(1), \mathbf{x}(2), \ldots)$ the content of a vector $\mathbf{x}$. Functions are noted in roman fonts, sets in calligraphy fonts, and variables in italic fonts. span $(\mathcal{A})$ represents the vector space spanned by $\mathcal{A}$. $\mathrm{p}\left(\mathbf{x}_j\right)$ denotes the distribution of vectors $\mathbf{x}_j$. $\sigma^2_{\mathbf{x}}$ represents the unbiased variance of a signal $\mathbf{x}$ and $\langle .|. \rangle$ denotes the usual scalar product. $\|\mathbf{x}\|$ denotes the Euclidean norm of a vector $\mathbf{x}$.

3 Spread-spectrum-based watermarking schemes

This section presents SS techniques and different modulations that are either insecure (SS and ISS) or that belong to a given security class (NW and CW).

3.1 Construction of marked signals

We want to hide a message $\mathbf{m}$ of N_c bits ($\mathbf{m} \in \mathbb{F}_2^{N_c}$) in a host Gaussian vector $\mathbf{x} \in \mathbb{R}^{N_v}$. This feature vector is obtained by selecting information from a linear transform (for example, DCT or DWT) of the content we want to watermark. The message is coded using N_c carriers $\mathbf{u}_i \in \mathbb{R}^{N_v}$. We generate the carriers with a pseudo-random number generator (PRNG) initialized with a seed K, the secret key. Carriers come as zero-mean Gaussian vectors obtained with the PRNG and are further orthogonalized (using Gram–Schmidt procedure) with unit variance in order to provide a basis of the private subspace, i.e., $\forall i \neq j$, $\langle \mathbf{u}_i | \mathbf{u}_j \rangle = 0$. Moreover, we have $\|\mathbf{u}_i\|^2 = N_v - 1$ (because we use the unbiased estimation of the variance). In WOA framework, security attacks are linked with the estimation of the carriers $\mathbf{u}_i$. It is not necessary to go back to the PRNG key K to successfully perform a security attack: contrarily to the arithmetics used in cryptography, in

watermarking, it is enough to have a good estimate of the secret carriers [7]. To create the watermark signal, we use a modulation. A modulation is an application $\mathrm{s} : \mathbb{F}_2 \times \mathbb{R}^{N_v} \to \mathbb{R}$. The watermark signal $\mathbf{w}$ is constructed as follows:

$$\mathbf{w} = \sum_{i=0}^{N_c-1} \mathrm{s}\left(\mathbf{m}(i), \mathbf{x}\right) \mathbf{u}_i. \tag{1}$$

As a convention, we set $\mathrm{s}(1, \mathbf{x}) > 0$ and $\mathrm{s}(0, \mathbf{x}) < 0$.

We obtain the watermarked signal by a summation of $\mathbf{x}$ and $\mathbf{w}$:

$$\mathbf{y} = \mathbf{x} + \mathbf{w}. \tag{2}$$

Distortion is assessed by means of the watermark-to-content ratio (WCR):

$$\mathrm{WCR}_{[\mathrm{dB}]} = 10 \log_{10} \left(\frac{\sigma_{\mathbf{w}}^2}{\sigma_{\mathbf{x}}^2} \right). \tag{3}$$

One objective of watermarking schemes is to guarantee that the embedded watermark survives genuine signal processing operations (in the case of still images, one may want to resist compression or noise addition). These attacks are generally not intentional. We model robustness attacks by adding a Gaussian noise $\mathbf{n}$ to $\mathbf{y}$ and we consider the attacked vector $\mathbf{r} = \mathbf{y} + \mathbf{n}$. Attack strength is assessed by means of the watermarked content-to-noise ratio (WCNR):

$$\mathrm{WCNR}_{[\mathrm{dB}]} = 10 \log_{10} \left(\frac{\sigma_{\mathbf{y}}^2}{\sigma_{\mathbf{n}}^2} \right). \tag{4}$$

Decoding is performed by computing the normalized correlations z between carriers and (possibly) attacked signal:

$$\mathrm{z}_{\mathbf{r},\mathbf{u}_i} = \frac{1}{\|\mathbf{u}_i\|^2} \langle \mathbf{r} | \mathbf{u}_i \rangle = \frac{1}{N_v - 1} \langle \mathbf{r} | \mathbf{u}_i \rangle. \tag{5}$$

We have:

$$\mathrm{z}_{\mathbf{r},\mathbf{u}_i} = \mathrm{z}_{\mathbf{x},\mathbf{u}_i} + \mathrm{s}\left(\mathbf{m}(i), \mathbf{x}\right) + \mathrm{z}_{\mathbf{n},\mathbf{u}_i}. \tag{6}$$

The first term represents host interference, which can be used for side-informed embedding but which is negligible statistically compared to the second term for the classical SS modulation because of the normalization by $N_v - 1$ (see Section 3.2). The third term is unpredictable at the time of embedding because it depends on the power of the attacking signal. If $\hat{\mathbf{m}}$ denotes the estimated message, for each bit, we have:

$$\hat{\mathbf{m}}(i) = \begin{cases} 0 \text{ if } \mathrm{z}_{\mathbf{r},\mathbf{u}_i} < 0, \\ 1 \text{ if } \mathrm{z}_{\mathbf{r},\mathbf{u}_i} > 0. \end{cases} \tag{7}$$

It comes from the classical decoding rule of SS communications. We measure decoding performance with bit error rate (BER) figures between the estimated and the original message:

$$\mathrm{BER}(\mathbf{m}, \hat{\mathbf{m}}) = \frac{1}{N_c} \sum_{i=0}^{N_c-1} \mathbf{m}(i) \oplus \hat{\mathbf{m}}(i). \tag{8}$$

3.2 Various spread-spectrum modulations

The general formula for the modulation s of SS is given by:

$$\mathrm{s}(\mathbf{m}(i), \mathbf{x}) = \alpha(i, \mathbf{x})(-1)^{\mathbf{m}(i) \oplus 1} - \lambda(\mathbf{x}) \mathrm{z}_{\mathbf{x},\mathbf{u}_i}. \tag{9}$$

where:

- $\alpha(i, \mathbf{x})$ allows to adjust the distortion of each carrier.
- $\lambda(\mathbf{x})$ allows to adjust informed embedding.

3.2.1 Spread spectrum (SS)

The SS classical modulation is the analog of the binary phase shift keying (BPSK) modulation for numerical communications. We have:

$$\alpha_{\mathrm{SS}}(i, \mathbf{x}) = \sqrt{\frac{\sigma_{\mathbf{x}}^2 10^{\mathrm{WCR}/10}}{N_c}}, \tag{10}$$

and

$$\lambda_{\mathrm{SS}}(\mathbf{x}) = 0. \tag{11}$$

The value of $\alpha(i, \mathbf{x})$ is proportional to the strength of the embedding: the higher the value, the more robust but the less imperceptible the embedding. Note that informed embedding is not enabled ($\lambda(\mathbf{x}) = 0$).

3.2.2 Improved spread spectrum (ISS)

In the previous modulation, the embedding process does not depend on the host interference $\mathrm{z}_{\mathbf{x},\mathbf{u}_i}$. However, it can be interesting to cancel this interference to improve both robustness and error probability. In [6], the authors propose a new modulation, the ISS with:

$$\alpha_{\mathrm{ISS}}(i, \mathbf{x}) = \sqrt{1 - \frac{\lambda^2 N_c}{N_v 10^{\mathrm{WCR}/10}}}, \tag{12}$$

and

$$\lambda_{\mathrm{ISS}}(\mathbf{x}) = \frac{1}{2} \left(1 + 10^{\mathrm{NCR}/10} + \frac{N_v 10^{\mathrm{WCR}/10}}{N_c} - \sqrt{\left(1 + 10^{\mathrm{NCR}/10} + \frac{N_v 10^{\mathrm{WCR}/10}}{N_c} \right)^2 - 4 \frac{N_v 10^{\mathrm{WCR}/10}}{N_c}} \right). \tag{13}$$

$\alpha(i, \mathbf{x})$ and $\lambda(\mathbf{x})$ are computed to achieve host-interference rejection and error probability minimization given a target noise-to-content power ratio:

$$\mathrm{NCR}_{[\mathrm{dB}]} = 10 \log_{10} \left(\frac{\sigma_{\mathbf{n}}^2}{\sigma_{\mathbf{x}}^2} \right), \tag{14}$$

where $\mathbf{n}$ denotes Gaussian noise (possibly added to the watermarked signal).

If $\lambda(\mathbf{x}) = 0$, this scheme is reduced to the classical SS scheme, and if $\lambda(\mathbf{x}) = 1$, the host interference is totally achieved.

3.2.3 Natural watermarking (NW)

In [8], a new modulation is proposed, namely, NW. Its goal is to preserve the distribution of $z_{\mathbf{x},\mathbf{u}_i}$, which is supposed to have circular (cf. Eq. 21) pdf, possibly scaled by a factor $\eta \geq 1$ (in order to set distortion):

$$\alpha_{\mathrm{NW}}(i, \mathbf{x}) = \eta |z_{\mathbf{x},\mathbf{u}_i}|, \tag{15}$$

and

$$\lambda_{\mathrm{NW}}(\mathbf{x}) = 1, \tag{16}$$

with $\eta = \sqrt{\frac{N_v - 1}{N_c} 10^{\mathrm{WCR}/10} - 1}$. For each carrier $\mathbf{u}_i$, one has:

$$|z_{\mathbf{y},\mathbf{u}_i}| = |\eta| |z_{\mathbf{x},\mathbf{u}_i}|. \tag{17}$$

If $\eta = 1$, considering the Gaussian hypothesis of the host signals, the distribution of host and marked contents in the subspace spanned by the carriers is the same. The higher this value, the more robust the embedding. However, we change the security of the scheme. Section 5.2 shows a discussion on the choice of the parameters and on the security of NW. In the remainder, this modulation is called NW when $\eta = 1$ and robust natural watermarking (robust-NW) when $\eta > 1$.

3.2.4 Circular watermarking (CW)

One drawback of NW is the fact that, since a great number of watermarked contents are close to the decoding regions borders, the robustness of the scheme is not very high. In [9], another modulation, namely CW, based on ISS modulation is proposed:

$$\alpha_{\mathrm{CW}}(i, \mathbf{x}) = \mathbf{d}(i) \alpha_{\mathrm{ISS}}(i, \mathbf{x}), \tag{18}$$

and

$$\lambda_{\mathrm{CW}}(\mathbf{x}) = \lambda_{\mathrm{ISS}}(\mathbf{x}). \tag{19}$$

Parameter $\mathbf{d}$ is generated at each embedding from a zero mean Gaussian signal $\mathbf{g}$. This perturbation is used to randomly spread the correlations of the mixed signals on the whole decoding regions:

$$\mathbf{d}(i) = \frac{|\mathbf{g}(i)|}{\|\mathbf{g}\|}. \tag{20}$$

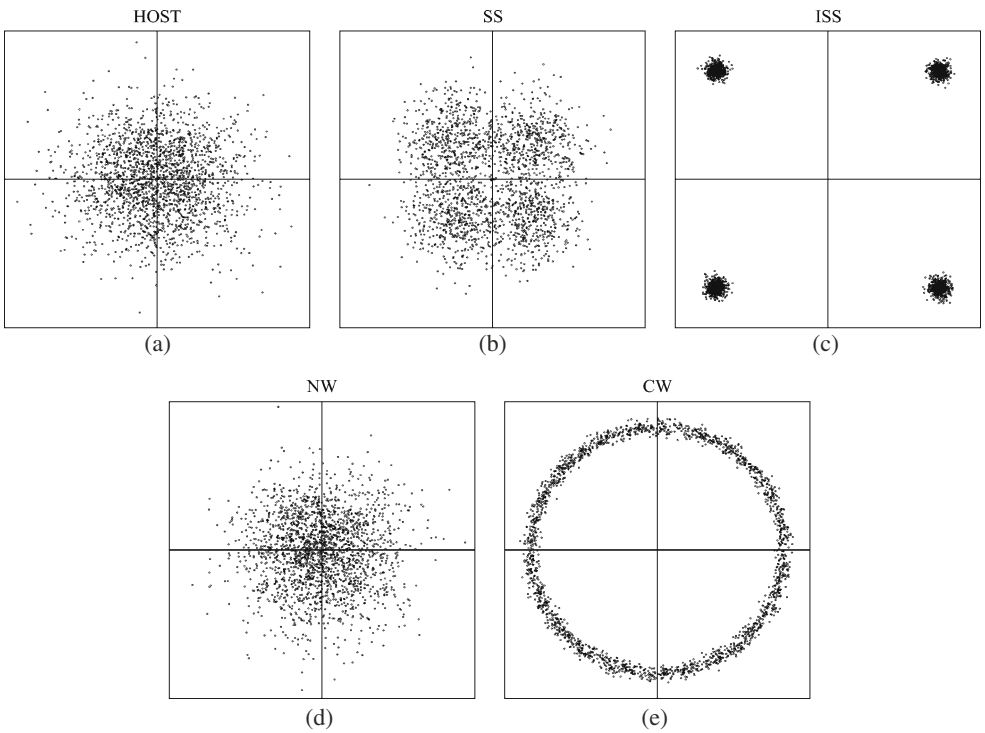

Fig. 1 Distributions of the correlations of the host signals (**a**) and of the marked signals (**b**, **c**, **d**, **e**) over the carriers. These distributions have been generated with 2,000 host Gaussian signals, $N_v = 256$, $N_c = 2$, WCR $= -20$ dB

Perturbation $\mathbf{d}$ enables the following property of circularity:

$$\mathrm{p}\left(\mathrm{z}_{\mathbf{y},\mathbf{u}_0},\ldots,\mathrm{z}_{\mathbf{y},\mathbf{u}_{N_c-1}}\right)=\mathrm{p}\left(\sqrt{\sum_{i=0}^{N_c-1}\mathrm{z}^2_{\mathbf{y},\mathbf{u}_i}}\right). \tag{21}$$

Circularity means that the distribution of the correlations can be reduced to a distribution that depends only on the Euclidean norm of the correlations. However, in this implementation, correlations can be spread on a wrong codeword region due to the host interference. This problem can occasionally lead to decoding errors (even in an attack-free context). We propose to replace the usual CW implementation with a new stochastic version of CW, called "zero-error-bit CW." The process consists of randomly generating a new perturbation $\mathbf{d}$ until the watermarked content $\mathbf{y}$ is located inside the right decoding region after embedding (this is checked using Eq. 7). Note that the complexity of this process depends on the host interference rejection, if $\lambda(\mathbf{x})=1$, the watermarked content is always in the right decoding region. This way, we have a CW implementation that has zero BER in an attack-free context without changing the probability density function of $\mathbf{y}$. In the remainder of the article, CW means "zero-error-bit CW" since it has intrinsically better performance than the original implementation.

Figure 1 shows correlations of host and marked signals for the four modulations over two secret carriers. We can see that, for each modulation, the four decoding regions depend on the sign of each secret carrier.

4 Implementation on still images

Theoretical watermarking schemes assume a Gaussian host signal $\mathbf{x}$. However, this model does not match usual distributions of feature vectors: DCT coefficients are traditionally modeled by a Laplace distribution and wavelet coefficients are usually modeled by a generalized Gaussian distribution. In order to get closer to this assumption (mainly for NW testing), we adopt the following trick: we perform a projection of the host feature vector on a set of pseudo-random signals, so that the vector that will be used for watermarking will be composed of the values of the projected host vector over the pseudo-random signals. This projected host feature vector can be assumed to have Gaussian pdf.

Note that embedding produces dependencies along the directions of the carriers, and independent component analysis (ICA) might be used to recover these directions. However, this is not possible in practice: ICA cannot be used on very high-dimensional observations (here, the vectors are more than 200,000 samples long) and only second-order analysis like principal component analysis can be reliably performed.

4.1 Image watermarking scheme

The scheme of the implementation of secure modulations on images is presented in Fig. 2. We want to watermark 8-bit grayscale images of $M\times N$ pixels. After a three-level 9/7 Daubechies wavelet forward transform, we arrange the nine first subbands (high-pass coefficients) of the host image into a vector $\mathbf{x_t}\in\mathbb{R}^{N_t}$. In order to have a Gaussian distribution, we construct the host signal $\mathbf{x}\in\mathbb{R}^{N_v}$ as follows:

$$\mathbf{x}(i)=\frac{2\sqrt{3}}{\sqrt{N_t}}\sum_{j=0}^{N_t-1}\mathbf{x_t}(j)\mathbf{a}_i(j). \tag{22}$$

$\mathbf{a}_i$ are pseudo-random, uniformly distributed vectors and the ratio $2\sqrt{3}$ is used to account for the variance of a uniformly distributed variable. This projection is orthogonal. Then, the watermark vector $\mathbf{w}$ is produced by SS watermarking from a random message $\mathbf{m}$ and carriers $\mathbf{u}_i$.

Distortion is assessed by means of the peak signal-to-noise ratio (PSNR). The link between PSNR and WCR is given by (proof in Appendix):

$$\mathrm{WCR}=10\log_{10}\left(\frac{255^2}{\sigma^2_{\mathbf{x}}}\times\frac{M\times N}{N_v}\right)-\mathrm{PSNR}. \tag{23}$$

Retroprojection of $\mathbf{w}$ in the wavelet domain is defined by:

$$\mathbf{w_t}(i)=\frac{2\sqrt{3}}{\sqrt{N_t}}\sum_{j=0}^{N_v-1}\mathbf{w}(j)\mathbf{a}_j(i). \tag{24}$$

Finally, we construct the marked signal $\mathbf{y_t}$ in the wavelet domain by constant strength embedding:

$$\mathbf{y_t}=\mathbf{x_t}+\mathbf{w_t}, \tag{25}$$

and we produce a marked image by applying the inverse wavelet transform.

4.2 Psychovisual constraints

It is possible to exploit the weaknesses of the human visual system (HVS) in order to better embed the watermark signal. The following assumptions hold in practice:

- HVS is less sensitive to high-activity areas (e.g., textures).
- HVS is more sensitive to low-activity areas (e.g., flat areas).

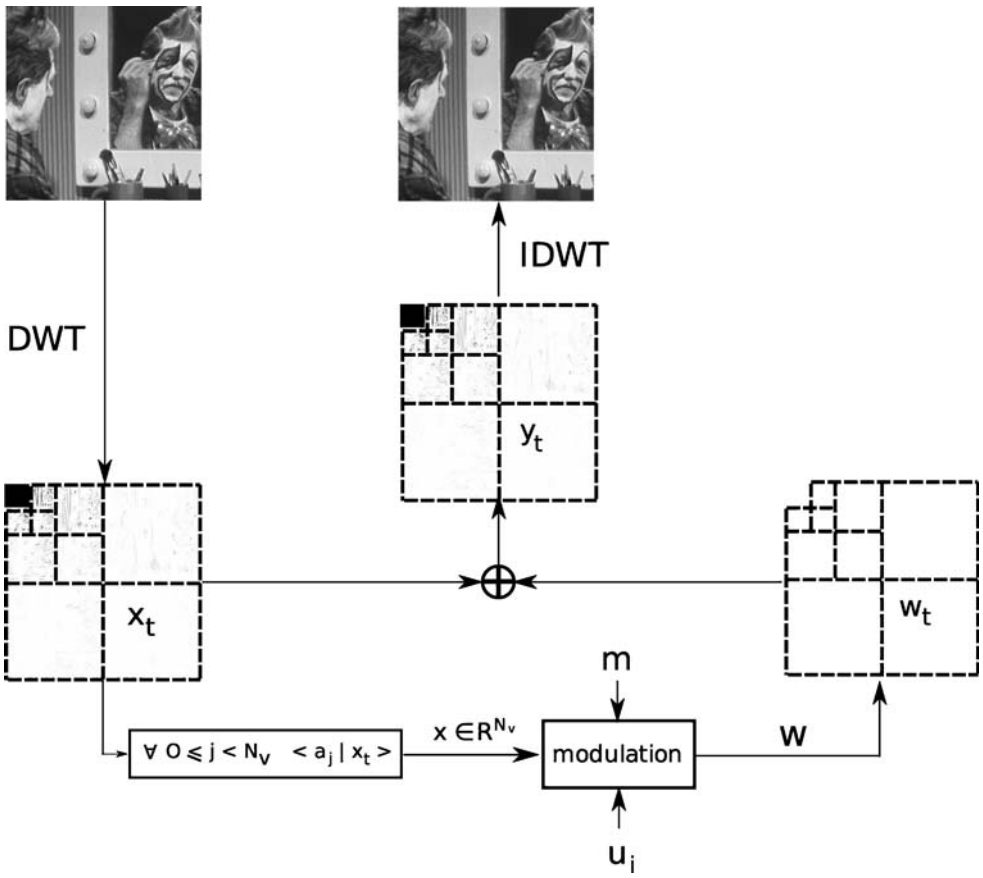

Fig. 2 Experimental watermarking scheme

Exploitation of these weaknesses in the current watermarking algorithm is made by adding psychovisual masking in the wavelet domain. We use variable strength embedding [10]:

$$\mathbf{y_t} = \mathbf{x_t} + \mathbf{w}'_\mathbf{t} \text{ with } \mathbf{w}'_\mathbf{t}(i) = \frac{|\mathbf{x_t}(i)|}{\mathbb{E}[|\mathbf{x_t}|]^2}\mathbf{w_t}(i), i \in \{0, ..., N_t - 1\}. \tag{26}$$

The factor $\frac{1}{\mathbb{E}[|\mathbf{x_t}|]^2}$ is used in order to preserve correlations in the N_c-dimensional private space and to avoid detection errors [11]. Therefore, the relationship between PSNR and WCR becomes (proof in Appendix):

$$\text{WCR} = 10\log_{10}\left(\frac{255^2}{\sigma_\mathbf{x}^2} \times \frac{M \times N}{N_v} \times \frac{\mathbb{E}[|\mathbf{x_t}|]^2}{\mathbb{E}[\mathbf{x_t}^2]}\right) - \text{PSNR}. \tag{27}$$

4.3 Numerical values and assessment

In practice, we use $M = N = 512$, $N_t = 258{,}048$, $N_v = 256$. We want to hide $N_c = 10$ bits on each image. We set a constant target PSNR of 45 dB for the four modulations: SS, ISS, robust-NW, and CW. Unless it is explicitly mentioned otherwise, we use constant strength embedding. Implementation of stego-secure NW is not evident because of the weakness of this modulation. In fact, variance of the watermark signal is too low and produces a fragile mark (because of the quantization of the marked image on 8 bpp). A way to circumvent this problem is to produce the mark in the medium frequencies subbands of the host image. Section 5.4.1 presents a practical implementation of NW.

Moreover, we use $N_c = 2$ on the figures in order to have a 2D representation of the distribution of the contents before and after embedding. This allows for intuitive reasoning on the BPSK constellation. However, one could hide as many bits as needed. Tests are made on 2,000 of the BOWS2-original images [12].

Table 1 Distortion caused by message embedding

Modulation	$\mathbb{E}$[PSNR](dB)	σ_{PSNR}(dB)
SS	44.75	1.2e-1
ISS	44.75	2.15e-1
Robust-NW	45.18	1.9e0
CW	44.76	2.16e-1

Target PSNR is 45 dB

4.4 Distortion

We have implemented the four modulations on the image database, and the distortion is computed on average on the marked contents. However, the quantization that consists of writing the marked image on 8 bpp causes a negligible variation on the target distortion.

Table 1 shows results obtained by implementation on the image database by setting a distortion equal to 45 dB.

Fig. 3 Comparison between constant (*top*) and variable (*bottom*) strength embedding. PSNR is 30 dB. Modulation: CW

Fig. 4 Comparison between constant (*top*) and variable (*bottom*) strength embedding (zoom on the upper-right corner of Fig. 3). PSNR is 30 dB. Modulation: CW

Robust-NW modulation is the one which exceeds the target distortion with the higher standard deviation. Because of the correct quality of embedding (45 dB), the visual distortion is negligible.

We have applied the previous psychovisual masking on CW modulation. Figure 3 shows an example of a picture marked with constant strength embedding (Eq. 25) and variable strength embedding (Eq. 26) with $\mathbb{E}[\text{PSNR}] = 30$ dB (in order to highlight the differences between the two approaches). Figure 4 shows zooms on the upper-right corners of the pictures of Fig. 3.

4.5 Robustness assessment

We have tested the robustness of the four modulations against JPEG compression, a very frequent attack on images, see Fig. 5. For each image, we embed a random message in it, we compress it, and we measure the BER. We measure BER on average with relation to JPEG quality factor on 2,000 marked images.

Figure 5 depicts the superior robustness of SS and ISS modulations against robust-NW and CW modulations: this is not surprising since the former do not have any security constraint to meet. In fact, there is a

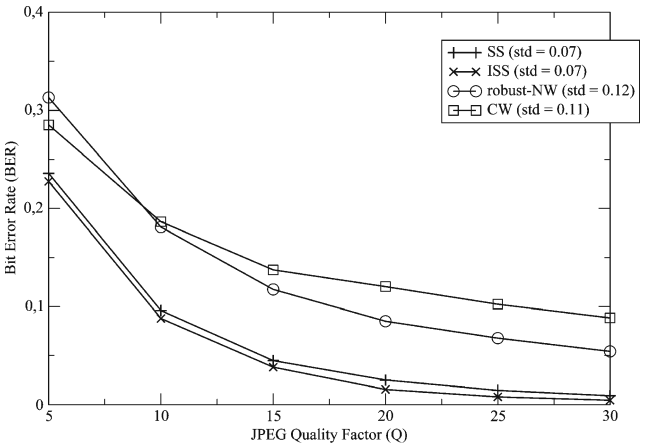

Fig. 5 Robustness of the four proposed modulations against JPEG compression. For each modulation, the legend includes the standard deviation of the experiments

compromise between security and robustness. If we do not want an adversary to estimate the carriers, projections of marked signals in corresponding codewords must be separated (presence of clusters).

5 Security of spread-spectrum schemes

5.1 Carriers estimation and blind source separation

We have seen that the problem of assessing data-hiding security in the WOA framework involves the knowledge of several (possibly) watermarked contents. If N_o denotes the number of observations an adversary has access to, considering the N_o watermarked contents column-wise gathered in the $\mathbf{Y}$ matrix, one has the following matrix relation:

$$\mathbf{Y} = \mathbf{X} + \mathbf{W} = \mathbf{X} + \mathbf{US}, \tag{28}$$

with:

- $\mathbf{Y} \in \mathcal{M}_{N_v,N_o}(\mathbb{R})$ watermarked signals
- $\mathbf{X} \in \mathcal{M}_{N_v,N_o}(\mathbb{R})$ host signals
- $\mathbf{W} \in \mathcal{M}_{N_v,N_o}(\mathbb{R})$ watermark signals
- $\mathbf{U} \in \mathcal{M}_{N_v,N_c}(\mathbb{R})$ carriers
- $\mathbf{S} \in \mathcal{M}_{N_c,N_o}(\mathbb{R})$ modulations of embedded messages

The problem of disclosing $\mathbf{U}$ and $\mathbf{S}$ is a blind source separation (BSS) problem, where $\mathbf{Y}$ represents the *observations*, $\mathbf{X}$ stands for a matrix representing *noise*, $\mathbf{U}$ represents the *mixing matrix*, and $\mathbf{S}$ represents the *sources*. ICA [13] performs BSS when modulations are independently drawn. We assume the following hypothesis: the information bits that are embedded in the images are independently drawn.

5.1.1 Principal component analysis

Principal component analysis deals with subspace security. It allows for an adversary to estimate the N_c-dimensional private subspace spanned by the carriers $\mathbf{U}$ (the private key), if the embedding alters the covariance matrix of the contents. Principal component analysis (PCA) is a technique used to reduce multidimensional data sets to lower dimensions for analysis. It involves the calculation of the eigenvalues of the covariance matrix of $\mathbf{Y}$ (taken column-wise). PCA has the distinction of being the optimal linear transformation that keeps the subspace with the largest variance (message embedding increases the variance of the signal in the directions of the carriers). The knowledge of this subspace for an adversary allows him/her to tamper with the hidden channel at minimum distortion.

5.1.2 Independent component analysis

In the context of SS embedding, ICA enables to establish whether a scheme is key-secure or insecure. It allows, after PCA, for an estimation of the carriers $\mathbf{U}$ when modulations are independently drawn and are not Gaussian. Note that the first assumption is relevant for SS and ISS in the case of WOA attack because the messages are supposed to be independent. ICA finds independent components by maximizing the statistical independence of the estimated components. However, ICA has three limitations:

- It can only estimate the carriers up to their sign.

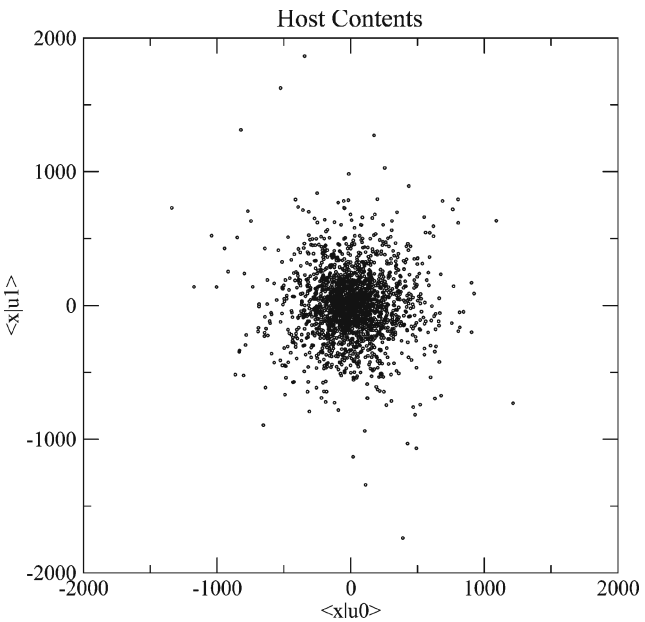

Fig. 6 Distribution of the projections of the host signals over two carriers

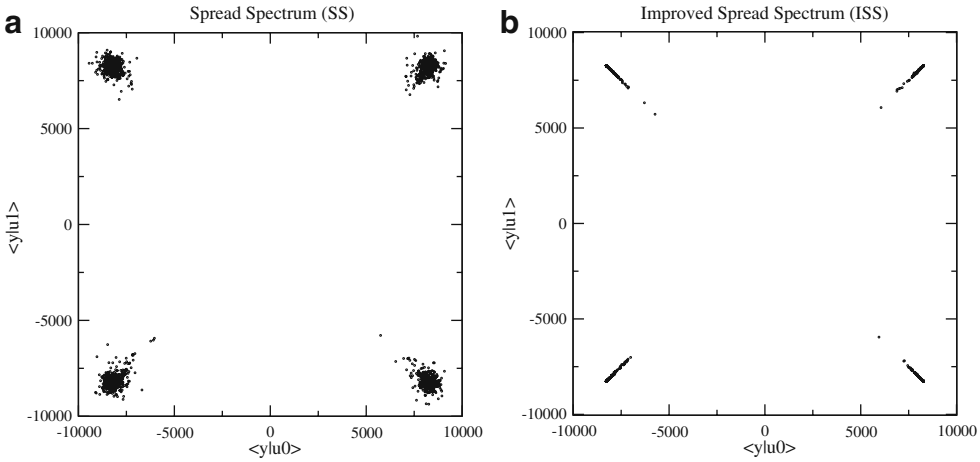

Fig. 7 Distribution of the projection of the marked signals for SS (**a**) and ISS (**b**) over two carriers

- It cannot estimate the order of the carriers (in the WOA framework—one would need the knowledge of a number of messages to do so, in the order of $\log_2(N_c)$).
- It cannot estimate the carriers if the sources are Gaussian distributed or dependent.

5.2 Security classes of spread-spectrum modulations

We use the assumption that host vectors are Gaussian-distributed. Since it is possible to estimate the carriers $\mathbf{u}_i$ [14] for SS and ISS in the WOA setup, these modulations are insecure. CW only allows for an estimation of the private subspace span($\mathbf{u}_i$). The circularity of the distribution allows to say that, for a subset of keys (all bases of span($\{\mathbf{u}_i\}$)), the distribution of marked signals will be identical (CW belongs to the so-called key-secure security class). NW modulation can theoretically be used for steganography. In fact, we have $D_{\mathrm{KL}}(\mathrm{p}(\mathbf{x})\|\mathrm{p}(\mathbf{y})) = 0$, where D_{KL} denotes the Kullback–Leibler divergence. With this hypothesis, one cannot decide whether a content is marked or not. In the case of NW on a circular (cf Eq. 21) pdf for the host feature vector, things go differently, depending on η, N_c, and N_v:

- If $N_c = N_v$:
  - If $\eta = 1$, then NW is stego-secure,

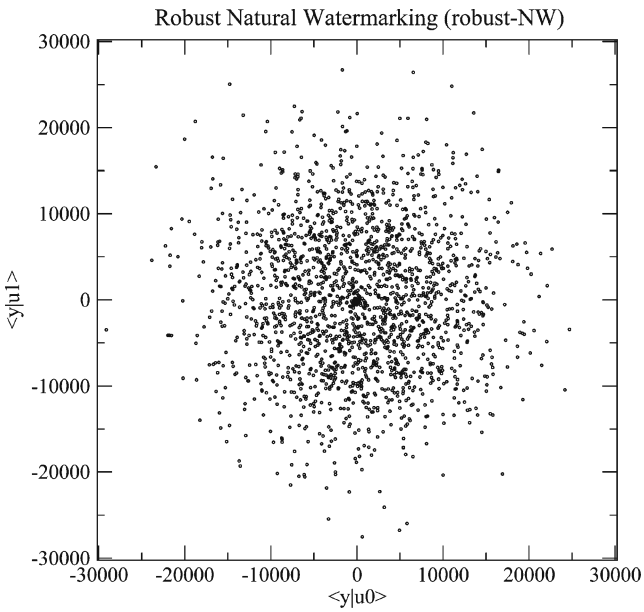

Fig. 8 Distribution of the projection of the marked signals for robust-NW over two carriers

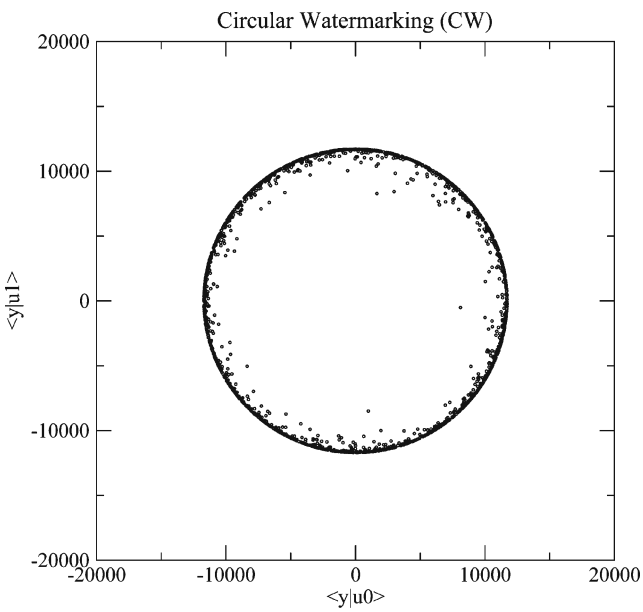

Fig. 9 Distribution of the projection of the marked signals over two carriers for CW, $N_c = 2$

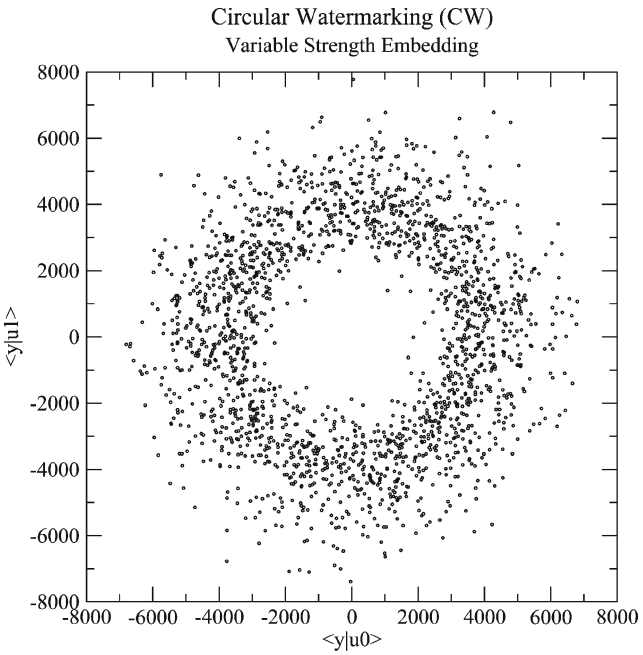

Fig. 10 Distributions of the projection of the marked signals over two carriers for CW with variable strength embedding: variance of correlations is stronger but does not impair security

- If $\eta > 1$, then NW is subspace-secure.

- If $N_c < N_v$:
  - If $\eta = 1$, then NW is subspace-secure,
  - If $\eta > 1$, then NW is key-secure.

5.3 Assessment of key security

5.3.1 Comparison of distributions of original and watermarked contents

Figure 6 represents the distribution of the host contents in the secret subspace spanned by the carriers. As expected, this distribution is asymptotically Gaussian. We further show the distributions of these contents after embedding using SS, ISS, CW, and robust-NW.

Figure 7 shows, respectively, SS and ISS distributions on two carriers. We can see the presence of four clusters (i.e., a constellation in classical communications), which correspond to the four possible messages with $N_c = 2$: (0, 0), (0, 1), (1, 0) and (1, 1). As expected, ISS modulation decreases the variance of the correlations in order to improve robustness. By PCA, the adversary can find the private subspace spanned by the carriers; by ICA, he/she can locate codewords. These schemes are insecure.

Figure 8 shows robust-NW distribution and Fig. 9 shows CW distribution on two carriers. These distributions are circular, and we can conclude that, for all bases $(\hat{\mathbf{u}}_0, \hat{\mathbf{u}}_1)$ of span$(\mathbf{u}_0, \mathbf{u}_1)$, the distribution $p(\mathbf{y}_1, ..., \mathbf{y}_{N_o} | \hat{\mathbf{u}}_0, \hat{\mathbf{u}}_1)$ will be identical (i.e., rotations of the secret subspace). It is consistent with the definition of circularity and, consequently, these schemes are key-secure, so that the adversary can access the subspace of the codewords but has no information about the decoding regions.

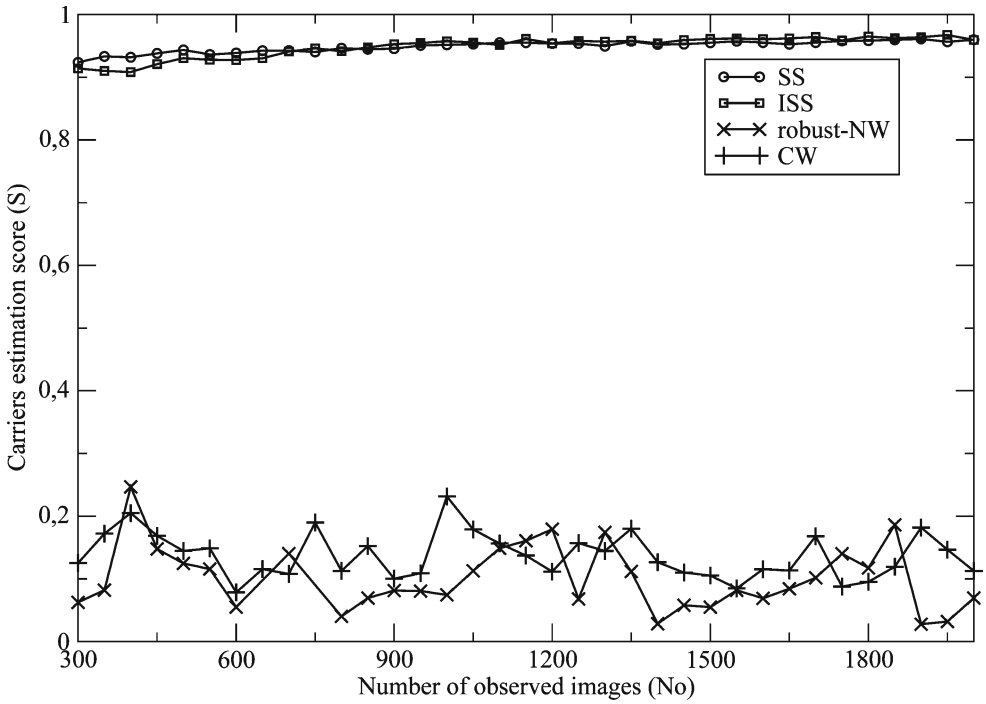

Fig. 11 Carrier estimation for the four modulations with relation to the number of observed marked contents. We apply ICA on marked signals (N_v by N_o matrix) to obtain the estimated carriers $\hat{\mathbf{u}}_j$. Next, we compute the score S according to Eq. 29. A score S close to one is equivalent to a correct estimation. Note that we just add columns to the matrix of marked signals when N_o grows up

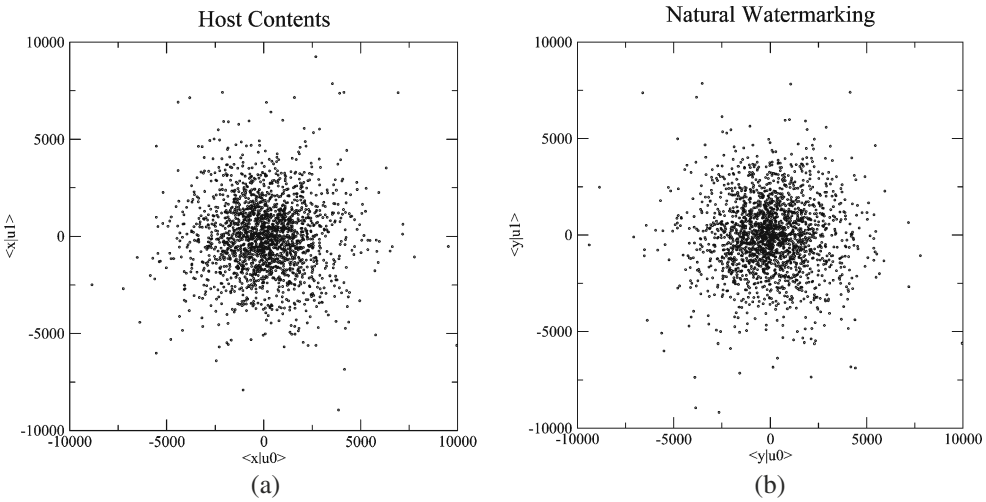

Fig. 12 Distributions of the projections of the host (**a**) and NW (**b**) signals over two carriers

Figure 10 shows CW correlations with variable strength embedding. As can be seen, the variance of the correlations is stronger than with constant strength embedding. However, it does not impair security: N_c-dimensional private subspace can still be found but codeword locations are kept unknown to an adversary.

5.3.2 Carriers estimation

In the WOA framework, the adversary has no access to the embedded messages. He/she can only assume that the messages are independently drawn. Moreover, we have seen in Section 5.1 that source separation techniques have limitations: they can only recover the carriers up to the sign and they cannot recover the order of the carriers. According to Kerckhoffs' principle, the adversary has access to the source code of the watermarking scheme. He/she can, therefore, estimate the projections of several marked contents in the N_v-dimensional space. Because of ICA limitations, we have to construct the following score S, obtained by the adversary, in order to quantify accuracy of the carrier estimation:

$$S = \frac{1}{N_c} \sum_i \left(\max_j^1 |z(\mathbf{u}_j, \hat{\mathbf{u}}_i)| - \max_j^2 |z(\mathbf{u}_j, \hat{\mathbf{u}}_i)| \right), \qquad (29)$$

where $\hat{\mathbf{u}}_i$ are the carriers estimated by ICA [15], and $\max^{1,\text{resp. }2}$ is the first (resp. second) maximum of the absolute value of the normalized correlation z between each correct carrier $\mathbf{u}_i$ and each estimated carrier $\hat{\mathbf{u}}_j$. This process was already successfully used [16].

Because $\mathbf{u}_{i \neq j}$ represents a basis of the secret N_c-dimensional space, we can conclude that a correct estimation of the carriers will produce a score S close to one when the number of contents that the adversary has access to is large enough. On the contrary, a wrong carrier estimation will produce a score S that will be close to zero or, equivalently, that will never converge. On Fig. 11, it is clear that the scores obtained by an adversary for the four modulations further illustrate the difference between embedding security classes: robust-NW and CW do not allow for a correct estimation of the carriers. Moreover, the score for these secure modulations does not depend on the number of observed contents.

5.4 Towards subspace-secure, spread-spectrum-based schemes

5.4.1 Robustness constraint for natural watermarking

Since NW provides subspace security, we also tried to provide an implementation of this method for digital images. However, we have seen in Section 4.3 that NW modulation cannot be applied directly because of the weakness of the modulation on high-pass wavelet coefficients when $\eta = 1$. To circumvent this problem,

Table 2 Tampering attack for NW and CW

	Before attack		After attack	
	E(PSNR) original/marked	BER	E(PSNR) marked/attacked	BER
NW	46.74	0.01	45.91	0.06
CW	46.65	0.04	45.55	0.29

we propose the new following parameters of the experimental watermarking scheme:

- We use four levels for wavelet decomposition.
- We arrange the subbands HL4 and LH4 of the host image in the vector $\mathbf{x_t}$.
- $N_t = 2{,}048$.

With these parameters, we obtain an average PSNR between host and marked images, which equals 46.7 dB. Here, we consider an average of the distortions in order to avoid to average infinite values when no modification of the host is needed during the embedding. Note that NW produces variable strength embedding and the PSNR is not fixed for each image. Although this scheme is not very robust, the main goal of this section is to illustrate the subspace-security property of NW (since we have $N_c \neq N_v$). We use PCA to estimate the subspace spanned by the carriers and we compare the results with CW (which does not allow for subspace security).

5.4.2 Comparison of NW and CW under a tampering attack

We have implemented NW on the image database. Figure 12 shows, respectively, host- and NW-marked distributions on two carriers. As can be seen, the distribution of correlations after NW is the same as the distribution of host correlations up to the variance, so that it is consistent with the definition of subspace security.

To assess the security of implementations of NW and CW, we propose a watermark tampering attack based on PCA, which aims at tampering the embedded message by putting the whole attack power into the estimated subspace. If an adversary can construct a basis of this subspace, he/she can make embedded messages unreadable. More precisely, the adversary recovers a basis with a PCA: $\{\hat{\mathbf{u}}_j\}_{j=0,\ldots,N_c-1}$ of the subspace defined by the carriers $\{\mathbf{u}_i\}_{i=0,\ldots,N_c-1}$. For each marked signal $\mathbf{y}$, the adversary constructs the attacked signal $\mathbf{r}$:

$$\mathbf{r} = \mathbf{y} - \sum_{j=0}^{N_c-1} \frac{\langle \mathbf{y}|\hat{\mathbf{u}}_j\rangle}{\langle \hat{\mathbf{u}}_j|\hat{\mathbf{u}}_j\rangle}\hat{\mathbf{u}}_j. \tag{30}$$

The tampering process tends to remove the components of $\mathbf{y}$ that are collinear with each estimated carrier $\hat{\mathbf{u}}_j$ in order to make the decoded bit random during the actual decoding step.

We have implemented this attack with NW and CW modulations on the 2,000 images. Results are shown in Table 2. In the case of CW, after attack, the distortion is correct and only 71% of the bits are correctly read. Therefore, we have verified that CW does not allow for subspace security. For NW, the attack is not efficient since 94% of the bits are decoded. In this case, the attack behaves as a classical AWGN attack. In fact, PCA on NW returns random mixing matrix (carriers) and sources (modulations). In accordance with theoretical approaches, subspace security of NW is verified.

6 Conclusions and perspectives

This paper presents a comparison between secure and insecure SS watermarking schemes for digital images that are either key-secure or subspace-secure; such schemes significantly differ from the classical modulations SS and ISS, which are insecure. We have shown that the implementations of the theoretical robust-NW and CW modulations on still images are possible considering robustness, security, and distortion constraints. Moreover, they provide a security level that is satisfactory for sensible applications.

One important conclusion is the fact that the security constraint has an impact on both the robustness and the induced distortion of the whole scheme. Parallel works show that it is possible to optimize the embedded distortion while guaranteeing security [17], and future works will focus on the joint optimization of security, robustness, and distortion. Other open research lines include the design of truly stego-secure schemes and the study of a scheme that does not need the trick we presented to obtain Gaussian-distributed host feature vectors.

Acknowledgements Benjamin Mathon, Francois Cayre, and Patrick Bas are partly supported by the European Commission through the National French projects Nebbiano ANR-06-SETIN-009, ANR-05-RIAM-01903 Estivale, and ARA TSAR. Moreover, Benjamin Mathon is partly supported by BCRYPT project, a Belgian Interuniversity Attraction Pole IAP-VI fund program. We also would like to thank the reviewers for their insightful and suggestive comments on this article.

Appendix A: Distortion specifications

We want to link the target PSNR for embedding with the theoretical WCR, used in the formulae of the four modulations. We give proofs of Eqs. 23 and 27.

A.1 Constant strength embedding

The first point is that, thanks to the nice normalization of retroprojection (Eq. 24), distortion stays constant in the wavelet domain and in the projected space:

$$\|\mathbf{w}_t\|^2 = \|\mathbf{w}\|^2 = d^2. \tag{31}$$

With renormalization against space dimensions, one gets:

$$\sigma^2_{\mathbf{w}_t} = \frac{d^2}{N_t}, \tag{32}$$

$$\sigma^2_{\mathbf{w}} = \frac{d^2}{N_v}. \tag{33}$$

So, we obtain:

$$\sigma^2_{\mathbf{w}} = \frac{N_t}{N_v}\sigma^2_{\mathbf{w}_t}. \tag{34}$$

Mean square error in the spatial domain is:

$$\text{MSE} = \frac{N_t}{M \times N}\sigma^2_{\mathbf{w}_t}; \tag{35}$$

therefore, PSNR equals:

$$\text{PSNR} = 10\log_{10}\left(\frac{255^2}{\frac{N_t}{M\times N}\sigma^2_{\mathbf{w}_t}}\right). \tag{36}$$

From the previous equation, one gets:

$$\sigma^2_{\mathbf{w}} = 255^2\frac{M \times N}{N_v}10^{-\frac{\text{PSNR}}{10}}, \tag{37}$$

which gives, once plugged into Eq. 3,

$$\text{WCR} = 10\log_{10}\left(\frac{255^2}{\sigma^2_{\mathbf{x}}} \times \frac{M \times N}{N_v}\right) - \text{PSNR}. \tag{38}$$

A.2 Variable strength embedding

From [10], watermark signal varies with the absolute value of the current wavelet coefficient we want to watermark, assuming that $\mathbf{x_t}(i)$ is independent from $\mathbf{w_t}(i)$. We have:

$$\|\mathbf{w}'_\mathbf{t}\|^2 = \frac{1}{\mathbb{E}[|\mathbf{x_t}|]^2}\sum_{i=0}^{N_t-1}|\mathbf{x_t}(i)|^2\mathbf{w_t}(i)^2 \tag{39}$$

$$\simeq \frac{1}{\mathbb{E}[|\mathbf{x_t}|]^2}\frac{1}{N_t}\sum_{i=0}^{N_t-1}\mathbf{x_t}(i)^2\sum_{i=0}^{N_t-1}\mathbf{w_t}(i)^2 \tag{40}$$

$$= \frac{\mathbb{E}[\mathbf{x_t}^2]}{\mathbb{E}[|\mathbf{x_t}|]^2}\|\mathbf{w_t}\|^2. \tag{41}$$

Equation 34 becomes:

$$\sigma^2_{\mathbf{w}} = \frac{\mathbb{E}[\mathbf{x_t}^2]}{\mathbb{E}[|\mathbf{x_t}|]^2}\frac{N_t}{N_v}\sigma^2_{\mathbf{w}'_\mathbf{t}}. \tag{42}$$

The same lines as above lead to the final equation for variable strength embedding:

$$\text{WCR} = 10\log_{10}\left(\frac{255^2}{\sigma^2_{\mathbf{x}}} \times \frac{M \times N}{N_v} \times \frac{\mathbb{E}[|\mathbf{x_t}|]^2}{\mathbb{E}[\mathbf{x_t}^2]}\right) - \text{PSNR}. \tag{43}$$

References

1. Cayre F, Furon T, Fontaine C (2005) Watermarking security: theory and practice. IEEE Trans Sig Proc 53(10):3976–3987
2. Kerckhoffs A (1883) La cryptographie militaire. J Sci Mil IX:5–38
3. Comesaña P, Pérez-Freire L, Pérez-González F (2005) Fundamentals of data hiding security and their application to spread-spectrum analysis. In: 7th information hiding workshop, IH05. Lecture notes in computer science. Springer, Barcelona
4. Cayre F, Bas P (2008) Kerckhoffs-based embedding security classes for woa data hiding. IEEE Trans Inf For Sec 3(1):1–15
5. Cox IJ, Killian J, Leighton FT, Shamoon T (1997) Secure spread spectrum watermarking for multimedia. IEEE Trans Image Process 6(12):1673–1687
6. Malvar HS, Flôrencio D (2003) Improved spread spectrum: a new modulation technique for robust watermarking. IEEE Trans Signal Process 53:898–905
7. Cox IJ, Doërr G, Furon T (2006) Watermarking is not cryptography. In: Proc. international workshop on digital watermarking (IWDW). http://www.adastral.ucl.ac.uk/~icox/papers/2006/IWDW2006.pdf
8. Bas P, Cayre F (2006) Natural watermarking: a secure spread spectrum technique for woa. In: Proc. information hiding, Alexandria
9. Bas P, Cayre F (2006) Achieving subspace or key security for woa using natural or circular watermarking. In: Proc. ACM multimedia security workshop, Geneva
10. Piva A, Barni M, Bartolini F, Cappellini V (1997) DCT-based watermark recovering without resorting to the uncorrupted original image. In: IEEE signal processing society 1997 international conference on image processing (ICIP'97), Santa Barbara
11. Furon T, Bas P (2008) Broken arrows. EURASIP J Inf Secur 2008(ID 597040)
12. Bas P, Furon T (2008) Break our watermarking system, 2nd edn. http://bows2.gipsa-lab.inpg.fr
13. Hyvärinen A, Karhunen J, Oja E (2001) Independent component analysis. Wiley, New York
14. Cayre F, Fontaine C, Furon T (2004) Watermarking attack: security of wss techniques. In: Proc. international workshop on digital watermarking (IWDW). Lecture notes on computer science (3304). Springer, New York, pp 171–183
15. Hyvarinen A (1999) Fast and robust fixed-point algorithm for independent component analysis. IEEE Trans Neur Net 10(3):626–634
16. Bas P, Doërr G (2007) Practical security analysis of dirty paper trellis watermarking. In: Proc. information hiding, St-Malo
17. Mathon B, Bas P, Cayre F, Pérez-González F (2008) Distortion optimization of model-based secure embedding schemes for data-hiding. In: Proc. information hiding, Santa-Barbara