Practical Performance Analysis of Secure Modulations for WOA Spread-Spectrum based Image Watermarking

Benjamin Mathon, Patrick Bas, François Cayre

Practical Performance Analysis of Secure Modulations for WOA Spread-Spectrum based Image Watermarking

Benjamin Mathon
GIPSA-Lab, dept. IS – UMR CNRS 5216
961 rue de la Houille Blanche
Domaine universitaire – BP 46
F-38402 Saint-Martin d'Heres cedex
benjamin.mathon@gipsa-lab.inpg.fr

Patrick Bas
GIPSA-Lab, dept. IS – UMR CNRS 5216
961 rue de la Houille Blanche
Domaine universitaire – BP 46
F-38402 Saint-Martin d'Heres cedex
patrick.bas@gipsa-lab.inpg.fr

Francois Cayre
GIPSA-Lab, dept. IS – UMR CNRS 5216
961 rue de la Houille Blanche
Domaine universitaire – BP 46
F-38402 Saint-Martin d'Heres cedex
francois.cayre@gipsa-lab.inpg.fr

ABSTRACT

This paper presents the first practical analysis of secure modulations for watermarking of still images in the case of a WOA (Watermarked Only Attack) attack framework (the attacker observes only marked contents). Two recent spread spectrum modulations, namely Natural Watermarking (NW) and Circular Watermarking (CW) are compared against classical modulations, namely Spread Spectrum (SS) and Improved Spread Spectrum (ISS). Results are discussed from the distorsion point of view, as well as from the the robustness and security point of view. We emphasize that the experiments were carried out on a rather significant number of images (2000) and demonstrate the relevance of these modulations in a real-world application.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*performance measures*

General Terms

Experimentation, Performance, Security

Keywords

Security, Spread-Spectrum, Still Images

1. INTRODUCTION

This paper deals with the Watermark Only Attack (WOA) where an attacker has only access to watermarked contents and tries to estimate the secret key (see [6] for details). In particular, he does not know about the embedded messages and/or the original works. A security *level* [6] is the number of contents required to produce an estimation of the secret that is better by an order of magnitude. A security level essentially deals with unsecure data-hiding methods. A security *class* [4] deals with the very nature of the secret that can be learned (or not) from watermarked contents. From [4], one can devise four embedding security classes in the WOA framework:

- *stego-security*: the embedding method fulfills Cachin's requirement for steganography (which is obviously to be straightforwardly fitted into a WOA framework). In particular, one cannot make any difference between host and watermarked contents;
- *subspace-security*: assuming the secret key lives in some private space, one cannot exhibit this space by observing watermarked contents – but steganalysis is possible;
- *key-security*: an attacker is able to disclose the private secret subspace but cannot make decision about the very secret that lives in it – optimal watermark removal is possible but security attacks are not;
- *insecurity*: an attacker is able to estimate the very secret that was used to embed the message into the host contents.

Recent works have proposed the use of two new modulations for spread-spectrum-based watermarking. These modulations were shown to belong to a better security class than unsecurity [1, 2]. Most specifically, Natural Watermarking (NW) can be either stego-secure or subspace-secure. Circular Watermarking (CW) is key-secure. SS and ISS, on the other hand, were shown to be unsecure.

The very goal of this paper is to experiment a practical implemenation of such secure modulations in the WOA framework. Results were collected with the help of a 2000-image database. This paper is organized as follows: Sec. 2 recalls the basics of spread-spectrum based watermarking and the modulations that are to be used throughout this paper. Sec. 3 shows attacks that the new modulations prevent. Sec. 4 presents an implementation of the practical algorithm we used for benchmarking purposes against security attack and robustness against JPEG compression. Finally, Sec. 5 draws some conclusions about the present results.

MM&Sec'07, September 20–21, 2007, Dallas, Texas, USA.

2. SPREAD-SPECTRUM WATERMARKING

Let $\mathbf{m} \in [0,1]^{N_c}$ be a binary message to be embedded into a host content $\mathbf{x} \in \mathbb{R}^{N_v}$ to produce a watermarked version $\mathbf{y}$ of $\mathbf{x}$. The watermark embedding operation can always be seen as follows:

$$\mathbf{y} = \mathbf{x} + \mathbf{w}, \tag{1}$$

where $\mathbf{w}$ is the watermark signal. One achieves the construction of $\mathbf{w}$ with the help of a modulation $s : [0,1] \to \mathbb{R}$:

$$\mathbf{w} = \sum_{i=1}^{N_c} \mathbf{u}_i s(\mathbf{m}(i)), \tag{2}$$

where the components of the $\mathbf{u}_i \in \mathbb{R}^{N_v}$ are $\mathcal{N}(0,1)$ real-valued carriers s.t.:

$$\forall i \neq j \quad <\mathbf{u}_i|\mathbf{u}_j> = 0, \tag{3}$$

where $<.|.>$ denotes the usual scalar product. The carriers are the output of a PRNG seeded with K the private secret key. From an attacker point of view, there's no difference between estimating the carriers and getting K because the dimension of the carriers is set and public (see for example watermarking schemes based on replication of a fixed-length pattern [8]). Security attacks traditionally target estimation of the $\mathbf{u}_i$. In a spread-spectrum framework, the private subspace is obviously $\mathrm{Span}(\mathbf{u}_i)$ and the secret is the set of the $\mathbf{u}_i$. Distorsion is assessed by means of the WCR (Watermark-to-Content Ratio) in dB:

$$WCR = 10 \log_{10} \left(\frac{\sigma^2_{\mathbf{w}}}{\sigma^2_{\mathbf{x}}} \right), \tag{4}$$

where $\sigma^2_{\mathbf{w}}$ (resp. $\sigma^2_{\mathbf{x}}$) is the variance of $\mathbf{w}$ (resp. $\mathbf{x}$).

Message decoding aims at producing an estimate $\hat{\mathbf{m}}$ of the original message by using the normalized correlation z:

$$z_{\mathbf{u}_i,\mathbf{y}'} = \frac{1}{N_v} \sum_j \mathbf{u}_i(j)\mathbf{y}'(j), \tag{5}$$

where $\mathbf{y}'$ is a (possibly) attacked version of $\mathbf{y}$. Estimated message is produced as follows:

$$\hat{\mathbf{m}}(i) = \mathrm{sign}(z_{\mathbf{u}_i,\mathbf{y}'}) \tag{6}$$

2.1 Unsecure modulations

Classical modulations, although unsecure, include Spread-Spectrum (SS) and Improved Spread Spectrum (ISS). SS modulation [7] is the analogon of the BPSK modulation for communications:

$$s_{SS}(\mathbf{m}(i)) = \gamma(-1)^{\mathbf{m}(i)}. \tag{7}$$

ISS [10] uses side-information to improve both robustness and error probability:

$$s_{ISS}(\mathbf{m}(i)) = \alpha(-1)^{\mathbf{m}(i)} - \lambda \frac{<\mathbf{x}|\mathbf{u}_i>}{\|\mathbf{u}_i\|^2}, \tag{8}$$

where α and λ are computed to achieve host-interference rejection and error probability minimization. Please refer to [10] to see how they are computed.

SS and ISS are unsecure, as they were already shown to allow for carriers estimation given enough watermarked images [5].

2.2 Secure modulations

Recently, Natural Watermarking (NW) and Circular Watermarking (CW) modulations were introduced to provide better security at the cost of some robustness. Contrary to ISS, NW uses side-information to enhance security:

$$s_{NW}(\mathbf{m}(i)) = \left(\eta(-1)^{\mathbf{m}(i)} \frac{<\mathbf{x}|\mathbf{u}_i>}{|<\mathbf{x}|\mathbf{u}_i>|} - 1 \right) \frac{<\mathbf{x}|\mathbf{u}_i>}{\|\mathbf{u}_i\|^2}. \tag{9}$$

When $\eta = 1$, NW belongs to the so-called stego-secure class and is suitable for steganography applications. When $\eta > 1$, NW is only subspace-secure (at best).

Decreasing security requirements, one can devise another modulation called CW, based on ISS:

$$s_{CW}(\mathbf{m}(i)) = \alpha(-1)^{\mathbf{m}(i)}\mathbf{d}(i) - \lambda \frac{N_v z_{\mathbf{x},\mathbf{u}_i}}{\|\mathbf{u}_i\|}, \tag{10}$$

where α and λ are computed the same way than with ISS and $\mathbf{d}$ is generated at each embedding as follows from $\mathbf{g} \sim \mathcal{N}(0,1)$:

$$\mathbf{d}(i) = \frac{|\mathbf{g}(i)|}{\|\mathbf{g}\|}. \tag{11}$$

This parameter is used to randomly spread the correlations of the mixed signals on the whole decoding regions. CW belongs to the so-called key-secure security class. Next section recalls what are embedding security classes.

3. SECURITY ATTACKS

The very problem of assessing data-hiding security involves the knowledge of several (possibly) watermarked contents. Let N_o denotes this number of observations a pirate has access to. Stating the N_o watermarking operations column-wise, one has the following matrix relation:

$$\mathbf{Y} = \mathbf{X} + \mathbf{W} = \mathbf{X} + \mathbf{US}, \tag{12}$$

where $\mathbf{S}$ are the modulations of the embedded messages, $\mathbf{U}$ is the matrix of the carriers and $\mathbf{X}$ is the matrix of the host contents. Given $\mathbf{Y}$ the matrix of the watermarked contents, the problem of disclosing $\mathbf{U}$ and $\mathbf{S}$ is known as blind source separation (BSS). One particular family of BSS method is called independent component analysis (ICA). It performs well when the sources (i.e. the modulations) are independently drawn. Obviously, ICA is relevant in the case of WOA attack where the messages can be supposed to be independent.

The goal of NW and CW is to break this independence so that ICA returns random sources and, most of all, random mixing matrix (i.e. random carriers). The reader may refer to [6] and [9] for further details. Methods to solve BSS when the sources are dependent is the subject of ongoing works.

4. TESTS ON STILL IMAGES

4.1 Implementation

We do not pay any attention to the problem of synchronization. Rather, we focus on robustness and security. Consider we want to watermark images of size $M \times N$ pixels. Using 9/7 Daubechies wavelete lifting scheme, we are able to construct a signal $\mathbf{x}_t \in \mathbb{R}^{N_t}$ from the finest detail subbands.

The signal $\mathbf{x}_t$ is known not to have Gaussian distribution. Since we need the Gaussian assumption for the host content when using NW, we use the central limit theorem to construct the host (asymptotically Gaussian) signal $\mathbf{x} \in \mathbb{R}^{N_v}$:

$$\mathbf{x}(i) = \frac{2}{\sqrt{3N_t}} \sum_{j}^{N_t} \mathbf{x}_t(j)\mathbf{a}_i(j), \tag{13}$$

where the $\mathbf{a}_i$ are pseudo-random uniformly-distributed vectors and the ratio $\frac{2}{\sqrt{3}}$ is used to normalize w.r.t. the variance of a uniformly-distributed variable. We produce $\mathbf{w}$, the watermark signal. It depends on $\mathbf{x}$ for ISS, NW and CW modulations. Proper retro-projection is performed to produce $\mathbf{w}_t$ the version of $\mathbf{w}$ in the wavelet domain:

$$\mathbf{w}_t(i) = \frac{2}{\sqrt{3N_t}} \sum_{j}^{N_v} \mathbf{w}(j)\mathbf{a}_j(i). \tag{14}$$

Also note that we set the $\mathbf{a}_i$ only quasi-orthogonal to one another for computing complexity reasons. Distorsion results will show the influence of this quasi-orthogonality compared to strict orthogonality. Finally, we obtain $\mathbf{y}_t$, the marked signal, by summation of $\mathbf{x}_t$ and $\mathbf{w}_t$.

In practice, we have $M = N = 512$, $N_t = 258048$, $N_v = 256$ and our tests were conducted with $N_c = 10$-bit payloads. Fig. 1 depicts our implementation.

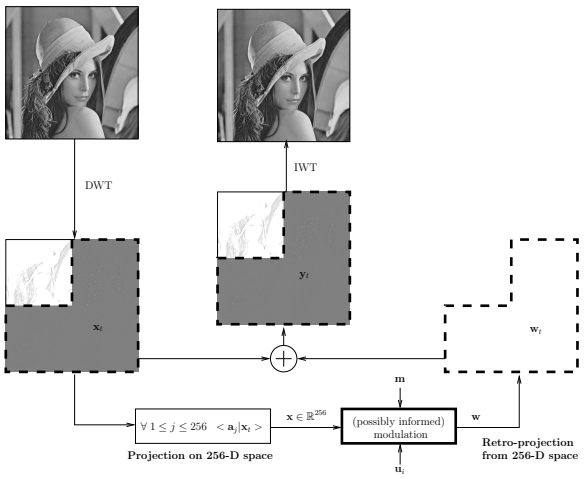

Figure 1: Flowchart of our implementation for stills. Essential work of this paper deals with the modulation box.

4.2 Measures and specifications

4.2.1 Distortion

Distortion was specified not to be less than a certain PSNR. We linked the PSNR and the WCR with the formulae obtained from Appendix. A. We were able to specify distorsion both for constant embedding strength and variable embedding strength [11]. Variable strength embedding is perfomed as follows:

$$\forall\ k,\ \mathbf{y}_t(k) = \mathbf{x}_t(k) + d|\mathbf{x}_t(k)|\mathbf{w}_t(k) \tag{15}$$

We report on Tab. 1 the distorsion results we had for our database of 2000 images with constant embedding strength, we also report the mean and the standard deviation of the original/marked PSNR for all images. As shown in Sec. 4.3, adding psycho-visual masking does not impair security.

Modulation	$\mathbb{E}[PSNR](dB)$	$\sigma_{PSNR}(dB)$
SS	44.75	1.18e-1
ISS	44.76	2.17e-1
NW	45.19	1.89e0
CW	44.76	2.16e-1

Table 1: Distorsion caused by payload insertion. Target PSNR: 45dB.

We depict on Fig. 2 to Fig. 5 the histograms of the obtained PSNR for the four modulations and constant embedding strength.

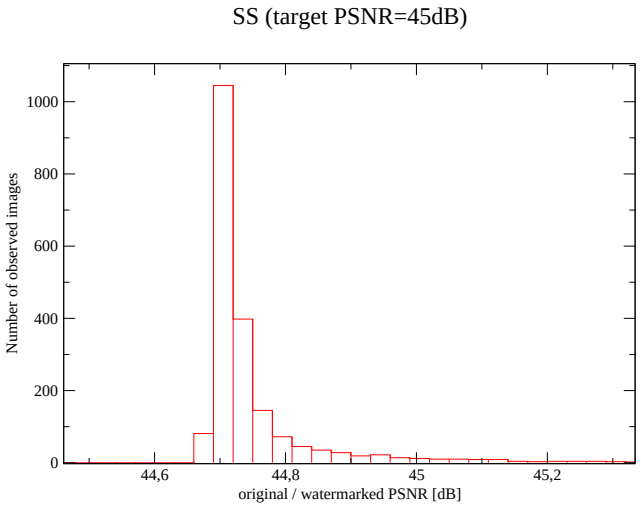

Figure 2: PSNR histogram for SS modulation.

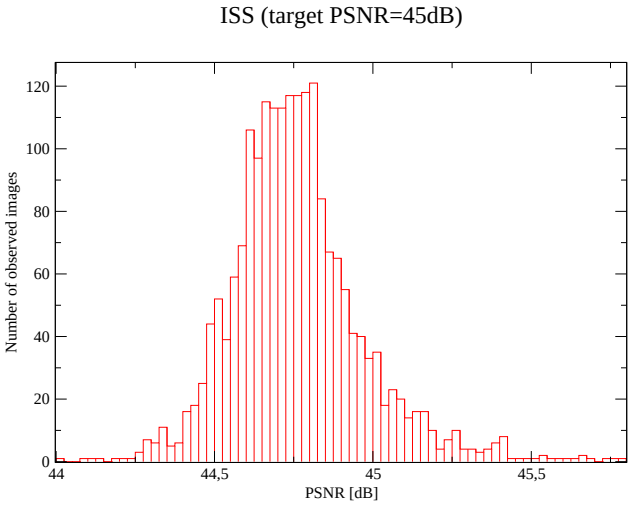

Figure 3: PSNR histogram for ISS modulation.

It is worth noting that CW, NW and ISS are all supposed to achieve the target PSNR on average. Further, since our projection over the $\mathbf{a}_i$ is only quasi-orthogonal, we pay this imprecision by 0.15dB on average, plus some additional variation as shown by the results for SS which should have yield

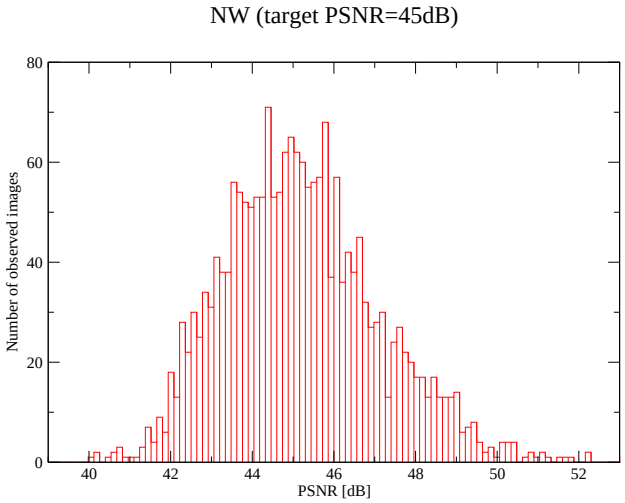

Figure 4: PSNR histogram for NW modulation.

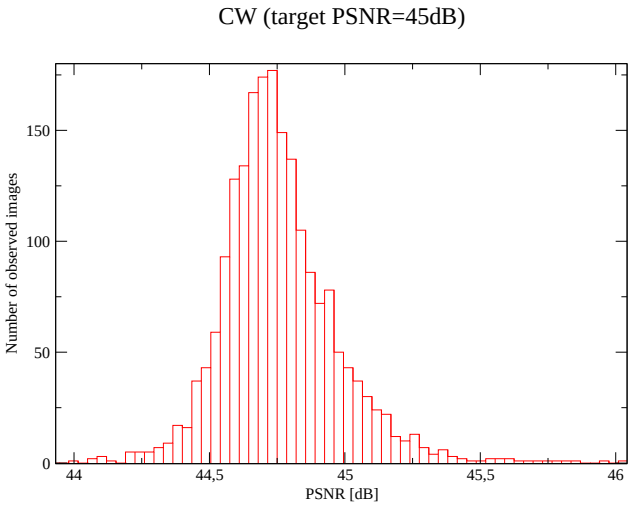

Figure 5: PSNR histogram for CW modulation.

perfect target PSNR. We believe this is not a big issue in practice.

4.2.2 *Robustness*

Robustness was assessed using JPEG compression. It was not the goal of this paper to treat optimal removal attack. Robustness results against JPEG compression are reported on Fig. 6

Cost in robustness due to security requirement vary substantially according to the desired target JPEG quality factor. However, we believe secure modulations (although not yet optimized) are already usable in practice. Optimal robustness attack deals with cancelling estimated carriers projections in the estimated subspace and is subject to ongoing work for the CW modulation.

4.2.3 *Security*

If projection (Eq. 13) and retro-projection (Eq. 14) are considered known by the attacker, the BSS problem stated

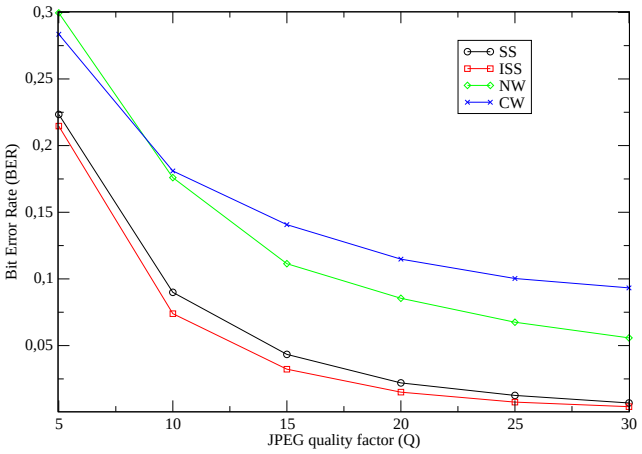

Figure 6: JPEG compression robustness results for the four modulations.

in Eq. 12 is to be solved in the N_v-D space. One has to keep the following general BSS limitations in mind:

- BSS can only recover the carriers up to the sign;
- BSS cannot recover the order of the carriers (WOA framework – one would need the knowledge of some messages to do so).

Therefore, since we try to estimate basis vectors of the private subspace, we can use the following measure of precision for estimation of the carriers [3]:

$$S = \frac{1}{N_c} \sum_i \left(\max_j^1 |z(\mathbf{u}_j, \hat{\mathbf{u}}_i)| - \max_j^2 |z(\mathbf{u}_j, \hat{\mathbf{u}}_i)| \right), \qquad (16)$$

where $\hat{\mathbf{u}}_\mathbf{i}$ are the estimated carriers and $\max_j^1 |z(\mathbf{u}_j, \hat{\mathbf{u}}_i)|$ (resp. $\max_j^2 |z(\mathbf{u}_j, \hat{\mathbf{u}}_i)|$) is the first (resp. second) maximal absolute value of the normalized correlation $z(\mathbf{u}_j, \hat{\mathbf{u}}_i)$. The score S will be close to one if we performed accurate estimation of the carriers. It will vegetate at low values without convergence in the case of inaccurate estimation of the carriers. Security results are collected on Fig. 7.

To further illustrate our views, we plot on Fig. 8 to Fig. 12 the 2D-distribution of the $\mathbf{y}$ projected over the two carriers when $N_c = 2$. For ISS and CW modulations, we precise the NCR (Noise-to-Content Ratio) in dB we used [10]:

$$NCR = 10 \log_{10} \left(\frac{\sigma_\mathbf{n}^2}{\sigma_\mathbf{x}^2} \right). \qquad (17)$$

This parameter sets the strength of AWGN attack by signal $\mathbf{n}$ that the marked signal should resist.

One can see on Fig. 8 and Fig. 9 that both SS and ISS modulations produce clusters in the private space. This is why it gets possible to estimate the carriers (up to the sign and the order).

On Fig. 10, no cluster appear. Actually, the watermarked distribution has the same shape than the host one. This allows for superior security in the WOA framework. Fig. 10

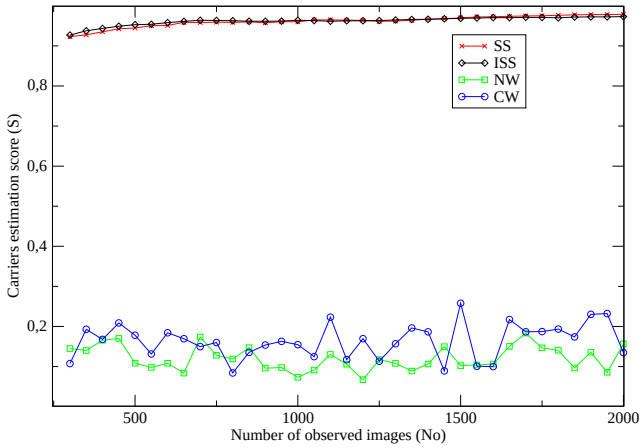

Figure 7: Security results for the four modulations. The used estimator is presented in Eq. 16. SS and ISS are confirmed not to be secure. CW and NW show good security also in practice.

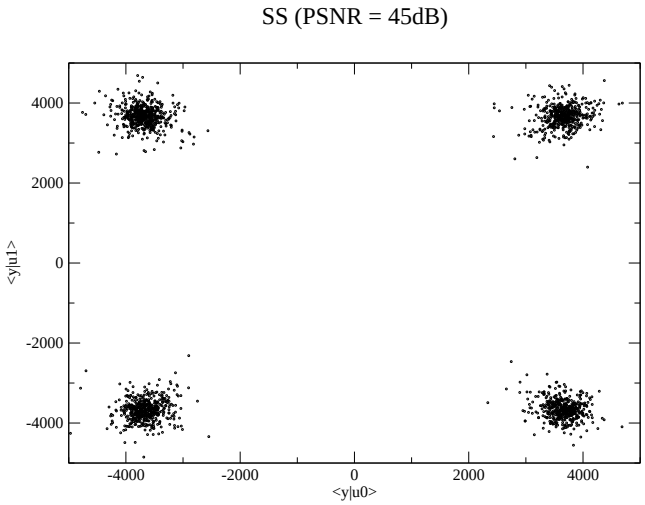

Figure 8: Distribution of the projection of the y over the two carriers for SS.

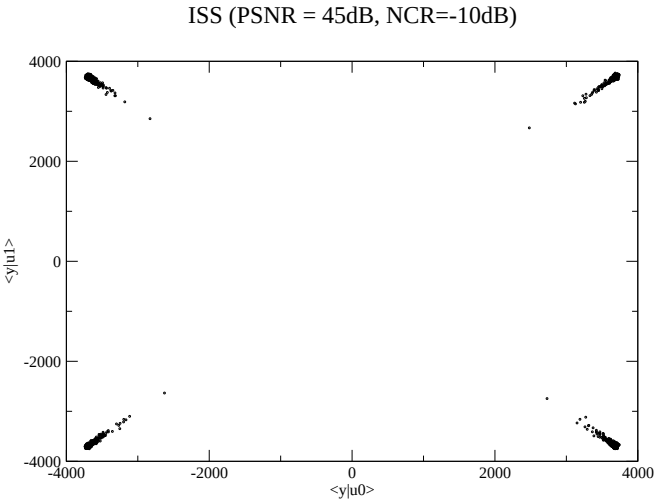

Figure 9: Distribution of the projection of the y over the two carriers for ISS.

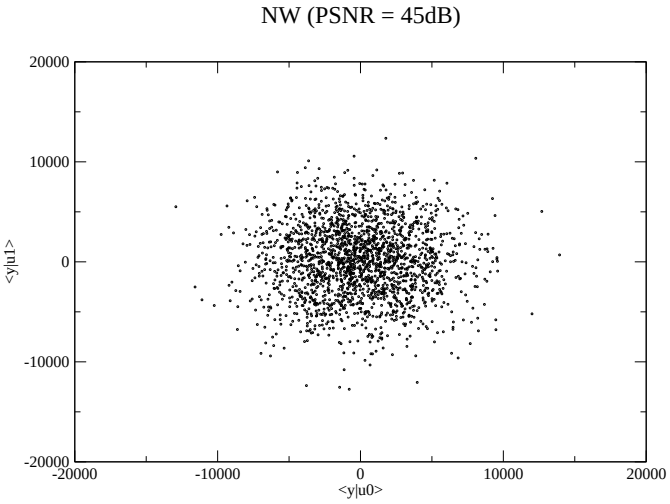

Figure 10: Distribution of the projection of the y over the two carriers for NW. Target PSNR obtained on average.

was obtained when specifying target PSNR *on average*. We think it is quite insightful to compare with Fig. 11 where target PSNR was reached exactly. In this case, NW should better be called constant-PSNR-NW. This enables to think of NW as a special case of CW, which it is actually. Note that this is consistent with the circularity definition given in [4].

On Fig. 12, we illustrate key-security with CW modulation. Even if the attacker can disclose the private subspace, he cannot make any decision in it about where are the secret carriers. CW delimits the fine line between robustness and security.

4.3 Security and psycho-visual masking

To illustrate the versatile capabilities of our scheme, we depict on Fig. 13 the distribution of the projection of the $\mathbf{y}$ over the $\mathbf{u}_i$ in the case variable strength embedding (Eq. 15) is used with CW modulation. As one can see, adding psycho-visual masking capabilities to our scheme does not impair security (the circularity property still remains).

Further, we depict on Fig. 14 the two versions of the same image watermarked with CW modulation, with and without variable strength embedding.

5. CONCLUSION

This work demonstrates that while high security requirement in the WOA framework may induce robustness loss, CW and NW secure modulations are good candidates for practical use in real-world applications. This somewhat mitigates the conclusions of [4] where CW and NW where shown to be significantly outperformed by SS and ISS from the robustness point of view. This paper used quite little payloads for proof-of-concept purposes. However, we are in the

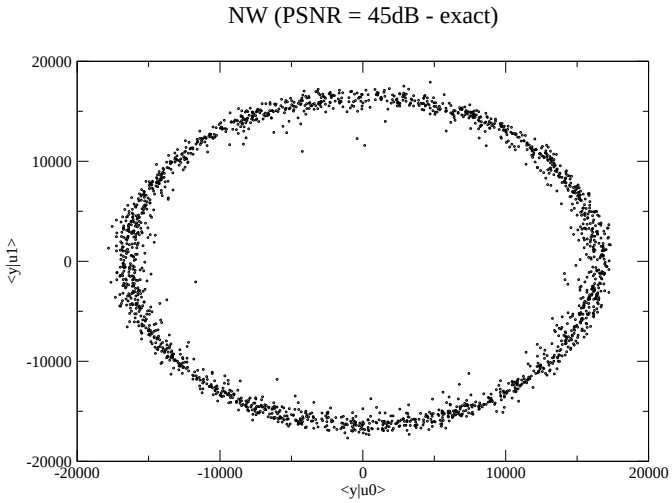

Figure 11: Distribution of the projection of the y over the two carriers for NW. The PSNR was obtained exactly.

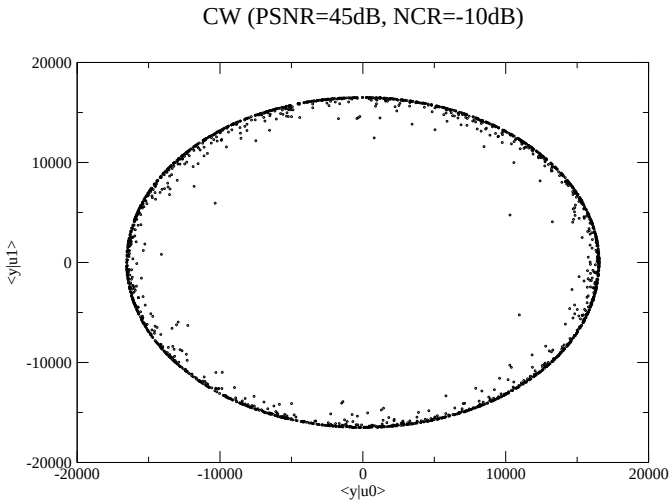

Figure 12: Distribution of the projection of the y over the two carriers for CW.

process of using these modulations in a more involved image watermarking scheme with pattern replication and more decent payload sizes. Future works therefore include assessment of security against the pattern size / the number of pattern replications and a more detailed study of the robustness / security tradeoff in a scheme designed to handle geometrical distorsions.

Acknowledgments

This was supported, in part, by IST-2002-507932 ECRYPT European grant, and ANR-06-SETI-009 Nebbiano, RIAM Estivale and ARA TSAR French grants.

6. REFERENCES

[1] P. Bas and F. Cayre. Achieving subspace or key security for woa using natural or circular watermarking. In *Proc. ACM Multimedia Security Workshop*, Geneva, Sept. 2006.

[2] P. Bas and F. Cayre. Natural watermarking: a secure spread spectrum technique for woa. In *Proc. Information Hiding*, Alexandria, VA, July 2006.

[3] P. Bas and G. Doerr. Practical security analysis of dirty paper trellis watermarking. In *Proc. Information Hiding*, St-Malo, June 2007.

[4] F. Cayre and P. Bas. Kerckhoffs based embedding security classes. *IEEE Trans. Inf. For. Sec. (under revision)*, 2007.

[5] F. Cayre, C. Fontaine, and T. Furon. Watermarking attack: Security of wss techniques. In *Proc. International Workshop on Digital Watermarking (IWDW), Springer-Verlag Lecture Notes on Computer Science, No. 3304, pp. 171–183*, Seoul, July 2004.

[6] F. Cayre, T. Furon, and C. Fontaine. Watermarking security: Theory and practice. *IEEE Trans. Sig. Proc.*, 53(10):3976–3987, Oct. 2005.

[7] I. Cox, J. Killian, F. Leighton, and T. Shamoon. Secure spread spectrum watermarking for multimedia. *IEEE Trans. Im. Proc.*, 6(12):1673–1687, Dec. 1997.

[8] F. Deguillaume, S. Voloshynovskiy, and T. Pun. Method for the estimation and recovering from general affine transforms. In *Proc. SPIE 2002*, San Jose, California, Jan. 2002.

[9] A. Hyvarinen. Fast and robust fixed-point algorithm for independent component analysis. *IEEE Trans. Neur. Net.*, 10(3):626–634, 1999.

[10] H. S. Malvar and D. F. rencio. Improved spread spectrum: a new modulation technique for robust watermarking. *IEEE Trans. Sig. Proc.*, 53:898–905, Apr. 2003.

[11] A. Piva, M. Barni, F. Bartolini, and V. Cappellini. DCT-based watermark recovering without resorting to the uncorrupted original image. In *IEEE Signal Processing Society 1997 International Conference on*

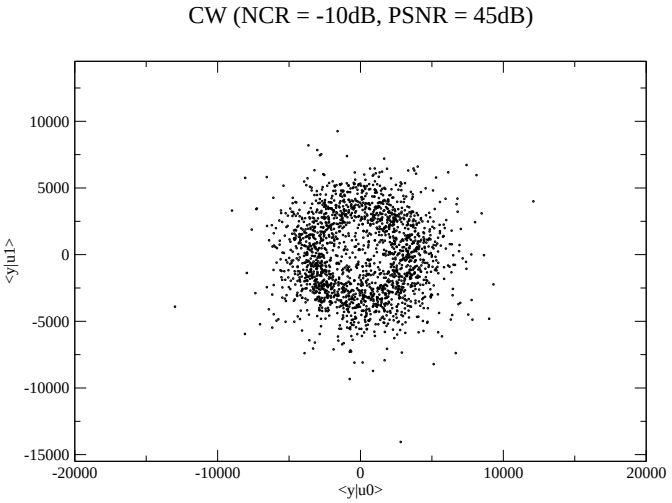

Figure 13: Distribution of the projection of the y over the two carriers for CW and variable strength embedding. Psycho-visual masking implies a higher variance in the target PSNR, which translates into a ring with larger width.

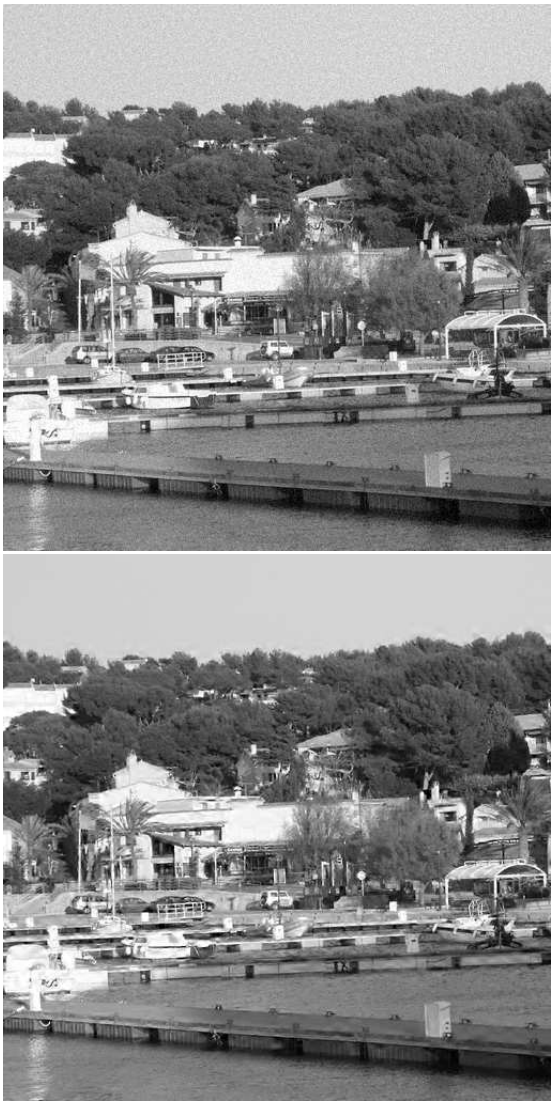

Figure 14: Comparison between constant (up) and variable (down) strength embedding. PSNR is 30dB. Modulation: CW.

Image Processing (ICIP'97), Santa Barbara, California, Oct. 1997.

APPENDIX

A. DISTORSION SPECIFICATIONS

The goal of this section is to rely the WCR to the PSNR in order to achieve the desired PSNR, possibly on average (depending on the modulation). We give results both for constant embedding strenght and for variable embedding strength [11].

A.1 Constant embedding strength

The first point is that, thanks to the nice normalization of projection (Eq. 13), distorsion keeps constant in the wavelet domain and in the projected space:

$$\|\mathbf{w}_t\|^2 = \|\mathbf{w}\|^2 = d^2 \tag{18}$$

Figure 15: Zoom of Fig. 14 on the sky above the trees.

With renormalization against space dimensions, one gets:

$$\sigma^2_{\mathbf{w}_t} = \frac{d^2}{N_t} \tag{19}$$

$$\sigma^2_{\mathbf{w}} = \frac{d^2}{N_v} \tag{20}$$

Thus leading to:

$$\sigma^2_{\mathbf{w}} = \frac{N_t}{N_v}\sigma^2_{\mathbf{w}_t} \tag{21}$$

MSE in the spatial domain is obviously:

$$MSE = \frac{N_t}{M \times N}\sigma^2_{\mathbf{w}_t}, \tag{22}$$

therefore, PSNR equals:

$$PSNR = 10\log_{10}\left(\frac{255^2}{\frac{N_t}{M\times N}\sigma^2_{\mathbf{w}_t}}\right). \tag{23}$$

From previous equation, one gets:

$$\sigma^2_{\mathbf{w}} = 255^2\frac{M \times N}{N_v}10^{-\frac{PSNR}{10}}, \tag{24}$$

which gives, once plugged into Eq. 4:

$$WCR = 10\log_{10}\left(\frac{255^2}{\sigma^2_{\mathbf{x}}} \times \frac{M \times N}{N_v}\right) - PSNR \tag{25}$$

A.2 Variable embedding strength

From [11], distorsion varies with the absolute value of the current wavelet coefficient to be watermarked. Eq. 21 therefore becomes:

$$\sigma_{\mathbf{w}}^2 = \frac{1}{\mathbb{E}[\mathbf{x}^2]} \frac{N_t}{N_v} \sigma_{\mathbf{w}_t}^2 . \tag{26}$$

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

$$WCR = 10 \log_{10} \left(\frac{255^2}{\sigma_{\mathbf{x}}^2} \times \frac{M \times N}{N_v} \times \frac{1}{\mathbb{E}[\mathbf{x}^2]} \right) - PSNR \tag{27}$$