

ROBUSTNESS OF ASYMMETRIC WATERMARKING TECHNIQUE

Teddy FURON

THOMSON multimedia,
Security Team,
1, av. Belle Fontaine,
35510 Cesson Sévigné, France

Pierre DUHAMEL

Ecole Nationale Supérieure
des Télécommunications de Paris,
Laboratoire Traitement Signaux et Images,
46 rue Barrault, 75013 Paris, France

ABSTRACT

Asymmetric schemes belong to second generation of watermarking. Whereas their need and advantage are well understood, many doubts have been raised about their robustness. According to a method presented in [1], a very robust symmetric technique is derived into an asymmetric scheme. Tests show that it is as robust as the symmetric version. Yet, asymmetric schemes undergo malicious attacks that confuse the detection process. Tests reveal that the quality loss due to these malicious attacks is too important for the signal to be used after the attack.

1. INTRODUCTION

To build a copy protection system for consumer electronic devices, we are looking for a technique, which could embed in an original content a signal commonly called watermark. Compliant devices such as players or recorders are able to detect the presence of this watermark. In this particular case, its presence means that the content is protected and thus it is illegal to copy it. This embedded watermark must not be perceptible.

To assess the security of watermarking, we made a threat analysis of these techniques. Some achieve good results in non-perceptibility and robustness, and all of them are symmetric schemes. Symmetric means that the detection process make use of the parameters used by the embedding process. The knowledge of these parameters allow pirates to forge illegal contents by modifying or removing watermark. This set of parameters is called the secret key and must be stored safely. This is not possible in consumer electronics. Tamper proof device is too expensive.

This is the reason why asymmetric watermarking schemes inspired from the cryptography domain have been recently studied ([2], [3] and [1]). They should be as robust as symmetric techniques with a detector needing a set of parameters called the public key different from the embedding's secret key. knowing the public key, it should be neither possible to deduce the private key nor possible to remove the watermark .

In this paper, we choose a symmetric technique achieving very good results in robustness. According to our method, we render it asymmetric. The first issue is to test the derived technique against common image transformations (for instance, JPEG, filtering, cropping...). The main result is that this derived technique is as robust as the symmetric

one. The second issue is the vulnerability against malicious attacks. These attacks are specific of the asymmetric method, but their visual impact depends on the watermarking technique.

2. ALGORITHMS

We describe in this section the algorithms of the symmetric technique, its derivation into a public key scheme, and its human perceptual model.

2.1. The symmetric technique

In this subsection, we give the technical details about the implementation of the symmetric technique invented by Alessia De Rosa and al.[4]. Its robustness is impressive, especially with the optimal version of the detector.

From a cover content C_o belonging to the "media space" the extraction function $X(\cdot)$ maps cover data into a vector in the "watermark space": $r_o = X(C_o)$. The "media space" is the spatial domain. $X(\cdot)$ orders in a vector r_o a subset of the magnitude of N discrete Fourier transform coefficients of C_o . These coefficients are extracted between the k -th and the $(k+n)$ -th diagonal in the first quadrant and their symmetrical images in the second quadrant [4]. They are ordered in a pseudo random manner, so that we can assume the sequence $\{r_o[m]\}$ is a white stationary process.

The role of the "mixing function" $f(\cdot)$ is to modify the extracted vector r_o into a vector r_w which is sufficiently similar to the watermark signal w : $r_w = f(r_o, w)$. In this paper, it modifies the amplitude of the DFT coefficients store in r_o proportionally to their value:

$$r_w[m] = r_o[m] \cdot (1 + \gamma \cdot w[m]) \quad \forall m \in [0..N-1]$$

where $\gamma > 0$ fixes the embedding depth. If $w[m] < \frac{-1}{\gamma}$, then $r_w[m]$ is clipped to 0.

The application of the "inverse extraction" function $Y(\cdot)$ concludes the embedding process. It maps back from the "watermark space" to the "media space": $C_w = Y(r_w, C_o)$. Here, $Y(\cdot)$ copies the DFT coefficients of C_o and changes the amplitude of those used at the extraction according to the watermarked vector r_w .

2.2. The asymmetric version

In [1], we described a method allowing to derive an asymmetric technique from classical spread spectrum ones. The

detection process does not compare the extracted signal r_u to a specific signal w , but checks if r_u has a specific statistical property due to the presence of w . Under several conditions, one can even demonstrate that it is not possible to estimate it from r_w . In this asymmetric scheme, the signal w is a filtered Gaussian central white noise v with unity variance:

$$r_w = f(r_o, (h \otimes v))$$

The normalized filter h and the signal v are private parameters.

The detection process does not need these private parameters, but it needs the amplitude of the frequency response of the filter h . This public parameter $|H(f)|$ characterizes the expected statistical property: the spectrum of r_u is shaped by $|H(f)|^2$. A simple hypothesis test decides to which hypothesis the unknown content C_u is more likely to belong:

- \mathcal{H}_0 : The extracted signal r_u is not watermarked, so it does not share this specific statistical property: Its estimated spectrum $g_0(f)$ is flat (we assumed that r_o is a white stationary processes). $g_0(f) = \sigma_{r_u}^2 + \mu_{r_u}^2 \cdot \delta(f)$ where μ_{r_u} and $\sigma_{r_u}^2$ are the mean and the variance of the tested extracted signal r_u .

- \mathcal{H}_1 : The extracted signal r_u has been watermarked. The following relations hold:

$$\begin{aligned} \varphi_{r_u}[l] &= E(r_u[m] \cdot r_u[m+l]) = \varphi_{r_o}[l] + \gamma^2 \cdot \varphi_{r_o}[l] \cdot \varphi_w[l] \\ g_1(f) &= \Phi_{r_o}(f) + \gamma^2 \cdot \Phi_{r_o}(f) \otimes \Phi_w(f) \end{aligned}$$

$E(\cdot)$ is the mathematical expectation, $\varphi_{r_o}[\cdot]$ the correlation function of the sequence $\{r_o[m]\}$ and $\Phi_{r_o}(f)$ its Fourier transform, which is the power spectral density of this sequence. We make the assumption that the sequences $\{r_o[m]\}$ and $\{w[m]\}$ are statistically independent. Here again, $\Phi_{r_o}(f) = \sigma_{r_o}^2 + \mu_{r_o}^2 \cdot \delta(f)$. Because $w = h \otimes v$, $\Phi_w(f) = |H(f)|^2$. The filter h is normalized so that $\int |H(f)|^2 \cdot df = 1$. Finally, the power spectral density expected if \mathcal{H}_1 is true, is:

$$\begin{aligned} g_1(f) &= \sigma_{r_o}^2 + \mu_{r_o}^2 \cdot \delta(f) + \gamma^2 \cdot (\mu_{r_o}^2 \cdot |H(f)|^2 + \sigma_{r_o}^2) \\ &= \mu_{r_u}^2 \cdot \delta(f) + \sigma_{r_u}^2 + \gamma^2 \cdot \mu_{r_u}^2 \cdot (|H(f)|^2 - 1) \end{aligned}$$

The critical region $R'(|H(f)|^2)$ is the set of extracted vector of the “watermark space” sharing this specific statistical property. It depends only on the public parameter, and can be defined as follows:

$$R'(|H(f)|^2) = \{r_u | U(r_u, g_0) - U(r_u, g_1) \geq Thr'\}$$

where $U(r_u, g_i)$ is the Whittle’s principal part of the likelihood that the spectrum of the random process r_u matches the power spectral density $g_i(f)$. Its simplified expression is $U(r_u, g_i) = 2N \int_{-\frac{1}{2}}^{\frac{1}{2}} (\log(g_i(f)) + I_N(f)/g_i(f)) \cdot df$ where $I_N(f)$ is the periodogram of the sequence r_u : $I_N(f) = \left| \sum_{k=0}^{N-1} r_u(k) \cdot e^{2\pi i k f} \right|^2 \quad \forall f \in]-\frac{1}{2}, \frac{1}{2}[$.

2.3. Human perception model

Up to now, the watermark’s invisibility issue is only tackled by the embedding depth γ . But, this action is very limited because the watermark signal, once mapped in the media space, is spread all over the image. Uniform areas of the image are very sensitive to watermark addition so they only support extremely small embedding depth γ , whereas edge areas, for instance, support deeper watermark addition. This issue leads to a spatial domain based human perceptual model giving the amount of noise each pixel can support. We selected the human perception model proposed by Bartolini and al. [5]. This empirical human perception model, gives good experimental results. It is based on the computation of the variance of the 9×9 windowed signal and by normalizing the obtained arrays with respect to its maximum value. Thus, $\forall(x, y) 0 \leq M(x, y) \leq 1$. Finally, the watermarked content is given by a new inverse extraction function $Y'(\cdot)$:

$$C_w = Y'(r_w, C_o) = (1 - M) \cdot C_o + M \cdot Y(r_w, C_o)$$

This choice allows us to compare the symmetric technique [4] and its asymmetric version because we share exactly the same condition of experiments. We simply model the influence of this masking function putting $\bar{\gamma} = \gamma \sum_{l,c} M(l, c)/N^2$ in place of γ in $g_1(f)$.

3. TESTS OF ROBUSTNESS

C_o is the $512 * 512$ pixels image “Lena”. For each test, It has been watermarked several times with different random permutations ordering DFT coefficients in r_o . The average performances are given here. Parameters are set to $n = 72, k = 78, \bar{\gamma} = 0.22$

3.1. Common transformations

3.1.1. JPEG Compression

The quality factor Q of the JPEG compression scheme varies from 100% to 0%. Detector’s response is normalized so that for $Q = 100\%$, its response is set to 1. Fig. 1 plots averaged results to a quality factor from 0% to 30%. Detection performs well until $Q = 5\%$. The symmetric technique robustness quality factor were $Q = 5\%$ with a correlation detector and $Q = 3\%$ with the optimum decoding¹.

3.2. Malicious attacks

The pirate knows exactly the detection process, especially what the detector is looking for. So, he can forge a pirated content C_p that the detector does not classify as watermarked. The asymmetric technique is robust to these malicious attacks if the quality of pirated contents is so poor that they have no commercial value. J. Eggers and B. Girod introduced in [3] the measure $\mathcal{D} = \frac{D(C_o, C_p)}{D(C_o, C_w)}$, where $D(B, C)$ represents a perceptual distance between two contents B and C . The greater \mathcal{D} is, the more robust the technique is. We assume $D(B, C)$ to be the Euclidean distance between

¹Test on filtering, cropping and noise addition are detailed in the full paper.

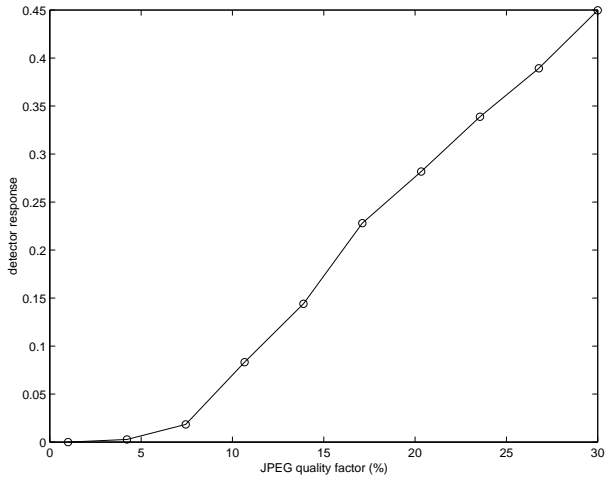


Figure 1: Robustness against JPEG

the two vectors B and C in the “media space”. Thanks to Parseval’s theorem, in our case, $D(B, C) = d(r_B, r_C)/M$ with $d(r_B, r_C)$ the Euclidean distance between the two vectors r_B and r_C in the “watermark space”.

3.2.1. Adding filtered noise

The strategy of the pirate here² is to whiten the sequence r_w adding a filtered noise r_n to it: $r_p[m] = r_w[m] + r_n[m] = r_o[m] * (1 + \gamma \cdot w[m]) + \eta \cdot (h' \otimes v')[m]$. The expected spectrum of r_p is then

$$\Phi_{r_p}(f) = \mu_{r_u}^2 \delta(f) + \sigma_{r_u}^2 + \gamma^2 \mu_{r_u}^2 (|H(f)|^2 - 1) + \eta^2 |H'(f)|^2$$

The pirate achieves his goal if $\Phi_{r_p}(f)$ looks like $g_0(f)$. This means $\gamma^2 \mu_{r_u}^2 |H(f)|^2 + \eta^2 |H'(f)|^2 = \sigma_P^2 \forall f \in]-\frac{1}{2}, \frac{1}{2}]$. The minimum value for σ_P^2 is $\gamma^2 \mu_{r_u}^2 \max(|H(f)|^2)$. This defines the following parameters.

$$\eta^2 * |H'(f)|^2 = (\gamma \mu_{r_u})^2 (\max(|H(f)|^2) - |H(f)|^2)$$

The pirate creates a filter h' which matches the required spectrum $|H'(f)|^2$. The Egger’s measure is then:

$$\mathcal{D} = \frac{\|\gamma \cdot \text{diag}(r_o) \cdot w + \eta(h' \otimes v')\|^2}{\|\gamma \cdot \text{diag}(r_o) \cdot w\|^2} = \max(|H(f)|^2)$$

This gives us a clue how to design the filter h . We must maximize \mathcal{D} choosing h as selective as possible. View Fig.(3.2.1) to assess quality loss.

4. CONCLUSION

The high performances against classical image transformations are mainly due to the implementation layout, especially the design of the extraction function $X(\cdot)$. The derivation into an asymmetric technique does not spoil these performances. We also achieved a fair robustness against malicious attacks especially designed for this asymmetric method.

²Malicious attack by re-filtering is also presented in the final paper.



Figure 2: forged content by malicious attack

This performance highly depends on the design of the filter h . This paper brings credits to the feasibility of asymmetric techniques.

5. ACKNOWLEDGEMENTS

The authors greatly thank Alessia De Rosa from the Communication and Image Processing Laboratory of the Department of Electronic Engineering of the University of Florence.

6. REFERENCES

- [1] T. Furon and P. Duhamel “An Asymmetric Public Detection Watermarking Technique” in *Proc. of the 3rd Int. Work. on Information Hiding*, Dresden, Sept 1999.
- [2] R.G. van Schyndel, A.Z. Tirkel, and I.D. Svalbe “Key Independent Watermark Detection” in *ICMCS’99*, Florence, Italy, 1999.
- [3] J. Eggers and B. Girod “Robustness of Public Key Watermarking Schemes”, *V³D² Watermarking Workshop*, Erlangen, Germany, Oct 1999.
- [4] A. De Rosa, M. Barni, F. Bartolini, V. Cappellini, and A. Piva “Optimum decoding of non-additive full frame DFT watermarks” in *Proc. of the 3rd Int. Work. on Information Hiding*, Dresden, Sept 1999.
- [5] F. Bartolini, M. Barni, V. Cappellini, and A. Piva “Masking building for perceptually hiding frequency embedded watermarks” in *Proc. IEEE. ICIP*, Chicago, Illinois, USA, Oct. 1998.