A COUNTERMEASURE TO RESIST BLOCK REPLACEMENT ATTACKS

Gwenaël Doërr and Jean-Luc Dugelay

Multimedia Communications Department
Eurécom Institute, Sophia-Antipolis, France

ABSTRACT

Security issues have almost been ignored during the first decade of digital watermarking. As a result, many released watermarking algorithms are weak against hostile intelligence. For instance, block replacement attacks defeat watermarking systems which do not consider the self-similarities of the host signal during embedding. Such attacks replace each signal block with another one, or a combination of other ones, taken at a different location. In this paper, a novel strategy will be presented to generate a signal coherent watermark to achieve immunity against block replacement attacks. The basic idea consists in imposing a linear relationship between watermark samples embedded at different locations, with respect to their local neighborhoods which are characterized with Gabor features.

1. INTRODUCTION

Digital watermarking was initially introduced in the 90's to ensure Intellectual Property (IP) protection [1]. Encryption alone is indeed not enough and a complementary technology is required. Sooner or later, encrypted multimedia content is decrypted to be eventually presented to a human being. At this very moment, multimedia content is left unprotected and can be easily copied and/or manipulated. Digital watermarking basically consists in hiding some information into digital content in an imperceptible manner. Research has mainly investigated how to improve the trade-off between three key parameters: robustness, imperceptibility and capacity. However, despite the fact that watermarking technologies are likely to be released in a hostile environment, security issues have almost been neglected. This explains in part why recent initiatives to introduce digital watermarking into Digital Right Management (DRM) frameworks have failed e.g. copy/playback control for the Digital Versatile Disk (DVD) [2] and for music [3]. Those setbacks seem to have significantly reduced the original enthusiasm from industries. As a result, security evaluation is now a key issue in digital watermarking.

To assess the security of different watermarking schemes, researchers try to anticipate hostile behaviors from malicious customers. In particular, collusion attacks have been shown to defeat many video watermarking algorithms [4]. The basic idea consists in combining several watermarked video frames to obtain unwatermarked content. This approach has been further extended to a finer resolution by using signal blocks instead of full frames [5]. Multimedia content is highly repetitive and it is consequently possible to exploit the self similarities of the signal to replace each signal block with another perceptually similar one. Those Block Replacement Attacks (BRA) defeat common watermarking schemes such as Spread Spectrum (SS) and Quantization Index Modulation (QIM) [6]. BRA basically exploit the fact that watermarking algorithms do not consider the self-similarities of the host signal during embedding. As a result, similar signal blocks are likely to carry uncorrelated watermark samples. This is a weak link that an attacker can exploit to defeat the protection system.

Intuitively, if *similar signal blocks carry similar watermarks*, BRA are likely to be ineffective. In other terms, the embedded watermark has to be coherent with the self-similarities of the host signal. In this paper, a possible way to obtain such *coherent watermarks* for still images is presented. In Section 2, Gabor filters are introduced to characterize the neighborhood of each pixel in the image. Next, in Section 3, a linear form is defined in the Gabor space to ensure that watermark samples inherit the same linear relationships as the Gabor-defined neighborhoods of the host signal. The resilience of this novel watermark against BRA is then evaluated in Section 4 in comparison with standard SS watermarks. Finally, conclusions are drawn in Section 5 and tracks for future work are given.

2. NEIGHBORHOOD CHARACTERIZATION WITH GABOR FILTERS

In order to impose a linear relationship between watermark samples with respect to the neighborhood of the considered pixel, it is first necessary to isolate some features to characterize this neighborhood. Gabor features are among the most popular ones and have been now used for a long time for a broad range of applications including image analysis and compression [7], texture segmentation [8], face authentication [9] and facial analysis [10]. Images are classically viewed either as a collection of pixels (spatial domain) or as the sum of sinusoids of infinite extent (frequency domain). But those representations are just two opposite extremes in a continuum of possible joint space/frequency representations. In such a perspective, frequency is viewed as a local phenomenon that can vary with position throughout the image. Furthermore Gabor wavelets have also received an increasing interest since they are particularly close to 2-D receptive fields profiles of the mammalian cortical simple cells [11].

The response of an input image $\mathbf{i}$ to a Gabor Elementary Function (GEF) with radius ρ and orientation θ is obtained by computing:

$$\mathbf{g}_{\rho,\theta} = \mathbf{i} * \mathbf{h}_{\rho,\theta} \tag{1}$$

where $*$ denotes convolution and $\mathbf{g}_{\rho,\theta}$ is the resulting filtered image. The GEF $\mathbf{h}_{\rho,\theta}$ is basically a complex 2D sinusoid whose orientation and frequency are given by (θ, ρ) modulated by a Gaus-

Jean-Luc Dugelay can be reached by email at jld@eurecom.fr. The authors acknowledge the European Commission for financial support through the IST Program under Contract IST-2002-507932 ECRYPT.

0-7803-9134-9/05/$20.00 ©2005 IEEE

sian envelope. For computational complexity reasons, Gabor filtering is usually performed in the FFT domain since it then comes down to a simple multiplication with the following filter:

$$\mathbf{H}_{\rho,\theta}(u,v) = \exp\left[-\frac{1}{2}\left(\left(\frac{u'-\rho}{\sigma_\rho}\right)^2 + \left(\frac{v'}{\sigma_\theta}\right)^2\right)\right]$$
$$\text{with} \quad \begin{pmatrix} u' \\ v' \end{pmatrix} = \mathbf{R}_\theta \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix} \quad (2)$$

where σ_ρ and σ_θ characterize the bandwidth of the GEF. In other terms, $\mathbf{H}_{\rho,\theta}$ is a 2D Gaussian that is shifted ρ frequency units along the frequency u-axis and rotated by an angle θ. Thus, it acts as a bandpass filter with a center frequency controlled by ρ and θ and a bandwidth regulated by σ_ρ and σ_θ. To obtain real valued features $\mathbf{g}_{\rho,\theta}$ in the spatial domain, GEFs are paired as follows $\mathbf{H}_{\rho,\theta} \leftarrow \mathbf{H}_{\rho,\theta} + \mathbf{H}_{\rho,\theta+\pi}$.

The idea is then to build a filter bank of such GEF pairs for M frequencies and N orientations to obtain a Gabor decomposition of the image. So, for each pixel position $\mathbf{p} = (x, y)$ in the image $\mathbf{i}$, the response of the different filters can be collected in a single column vector $\mathbf{g}(\mathbf{i},\mathbf{p}) = \{\mathbf{g}_{\rho_{i,j},\theta_{i,j}}(\mathbf{i},\mathbf{p})\}$ with $1 \leq i \leq M$ and $1 \leq j \leq N$. This vector can be regarded as the local power spectrum of the image and thus be used to characterize the neighborhood. Based on previous work [9], the different parameters of the GEF pairs are computed as follows:

$$\rho_{i,j} = \rho_{\min} + b\frac{(s+1)s^{i-1}-2}{s-1} \quad (3)$$
$$\sigma_{\rho_{i,j}} = tbs^{i-1} \quad (4)$$
$$\theta_{i,j} = \frac{(j-1)\pi}{N} \quad (5)$$
$$\sigma_{\theta_{i,j}} = t\frac{\pi\rho_{i,j}}{2N} \quad (6)$$
$$b = \frac{\rho_{\max}-\rho_{\min}}{2}\left(\frac{s-1}{s^M-1}\right) \quad (7)$$

The whole filter bank is specified by the 6 parameters M, N, $\rho_{\min}$, $\rho_{\max}$, s and t. The first two parameters determine the number of orientations and frequencies in the filter bank. The next two ones specify the bandwidth in which the filters are bound. The parameter s controls how much the radial bandwidth increases when the radius increases. Typically when it is set to 2, frequency bands are distributed in octave steps with a frequency bandwidth which doubles at each step. Finally, the parameter t sets the value at which neighboring filters intersect. For instance, with $t = 1$, they cross at equal value $1/e$ along their principal axis.

3. SIGNAL COHERENT WATERMARKS

For each signal block, BRA look for a linear combination of neighboring blocks which results in a block which is similar enough to the current block so that a substitution does not introduce strong visual artifacts [5]. Today, most watermarking systems are defeated by such attacks since nothing specific is done to ensure that the embedded watermark is coherent with the self-similarities of the host signal. Alternative watermark generation algorithms have consequently to be designed to obtain such signal coherent watermarks. An intuitive specification is that *similar signal blocks should carry similar watermarks* or alternatively that *pixels with similar neighborhood should carry watermark samples with close values*. In Subsection 3.1, it will be shown that such watermarks can be theoretically obtained if watermark samples are considered as the output of a linear form in the Gabor space. A discussion is then conducted in Subsection 3.2 to get a practical implementation of this approach.

3.1. Linear Form in Gabor Space

From a very low-level perspective, generating a digital watermark can be defined as associating a watermark value $\mathrm{w}(\mathbf{i},\mathbf{p})$ to each pixel location in the image. Then, if the watermark is required to be immune against BRA, the following property should be verified:

$$\mathbf{g}(\mathbf{i},\mathbf{p}_0) \approx \sum_k \lambda_k \mathbf{g}(\mathbf{i},\mathbf{p}_k) \Rightarrow \mathrm{w}(\mathbf{i},\mathbf{p}_0) \approx \sum_k \lambda_k \mathrm{w}(\mathbf{i},\mathbf{p}_k) \quad (8)$$

In other terms, if at a given position, the local neighborhood is similar to a linear combination of neighborhoods at other locations, then the watermark sample should be close to the linear combination (with the same mixing coefficients λ_k) of the watermark samples at these locations. In order to obtain this property, one can write $\mathrm{w} = \varphi \circ \mathbf{g}$ where $\varphi(.)$ is a linear form in the MN dimensional Gabor space $\mathcal{G}$. Subsequently, it is sufficient to choose an orthonormalized basis $\mathcal{B} = \{\mathbf{b}_l\}$ of the Gabor space and the values $\xi_l = \varphi(\mathbf{b}_l)$ to completely define the linear form $\varphi(.)$. This is where some secret can be injected into the framework i.e. the basis and its associated values can be pseudo-randomly generated using a secret key K. Furthermore it should be noted that one can decide to have less than MN basis vectors if other constraints have to be imposed. For instance, it happens that neighborhoods which are the same modulo a small set of geometrical operations, e.g. 8 isometries and scaling by a factor 2, are required to carry the same watermark samples to achieve robustness [12]. The dimension of the vector space spanned by the $\mathbf{b}_l$'s will be denoted $D \leq MN$. Now, if the values taken by the linear form on the unit sphere $\mathcal{U}$ of this subspace are considered, the following probability density function is obtained [13]:

$$\mathrm{f}_{\varphi|\mathcal{U}}(w) = \frac{1}{\Xi\sqrt{\pi}}\frac{\Gamma\left(\frac{D}{2}\right)}{\Gamma\left(\frac{D-1}{2}\right)}\left[1-\left(\frac{w}{\Xi}\right)^2\right]^{\frac{D-3}{2}} \quad (9)$$

where $\Xi^2 = \sum_{l=1}^{D}\xi_l^2$ and $\Gamma(.)$ denotes the Gamma function. When D grows large, this tends towards a Gaussian distribution with zero mean and standard deviation $\Xi/\sqrt{D}$. Thus if the ξ_l's are chosen to have zero mean and unit variance, this ensures that the obtained watermark is equivalent to a Gaussian watermark with zero mean and unit variance multiplied by some local scaling factors, which can be regarded as perceptual shaping. As a matter of fact, by linearity, $\mathrm{w}(\mathbf{i},\mathbf{p}) = \|\mathbf{g}(\mathbf{i},\mathbf{p})\|.\varphi\big(\mathbf{g}(\mathbf{i},\mathbf{p})/\|\mathbf{g}(\mathbf{i},\mathbf{p})\|\big)$. The greater the norm $\|\mathbf{g}(\mathbf{i},\mathbf{p})\|$ is, the more textured the neighborhood is and the more amplified is the normally distributed watermark sample $\varphi\big(\mathbf{g}(\mathbf{i},\mathbf{p})/\|\mathbf{g}(\mathbf{i},\mathbf{p})\|\big)$. On the other hand, in smooth areas, $\|\mathbf{g}(\mathbf{i},\mathbf{p})\|$ is small and the watermark is attenuated.

3.2. Practical Implementation

When M and N grow, more and more Gabor responses $\mathbf{g}_{\rho_{i,j},\theta_{i,j}}$ need to be computed which can rapidly get prohibitive. A practical implementation of the presented watermark generation algorithm has consequently been investigated. To obtain the watermark sample $\mathrm{w}(\mathbf{i},\mathbf{p})$ at a given position, the local Gabor power spectrum

$\mathbf{g}(\mathbf{i},\mathbf{p})$ is first projected onto the basis $\mathcal{B}$. Then, the inner product between the obtained vector and the column vector $\boldsymbol{\xi}$ containing the ξ_l's is computed. This can be written:

$$\mathrm{w}(\mathbf{i},\mathbf{p}) = \left(\mathbf{b}^{\mathrm{T}}\mathbf{g}(\mathbf{i},\mathbf{p})\right)^{\mathrm{T}}\boldsymbol{\xi} = \mathbf{g}(\mathbf{i},\mathbf{p})^{\mathrm{T}}\boldsymbol{\psi} \qquad (10)$$

where $.^{\mathrm{T}}$ denotes the transposition operation, $\mathbf{b}$ is a matrix whose columns are the $\mathbf{b}_l$'s and $\boldsymbol{\psi} = \mathbf{b}\boldsymbol{\xi}$. It should be noted that the whole secret of the algorithm is contained in the MN values $\psi_{i,j}$ of the column vector $\boldsymbol{\psi}$. Now, looking at Equation (10), the watermark can be written as:

$$\mathbf{w} = \sum_{i=1}^{M}\sum_{j=1}^{N} \psi_{i,j}\mathbf{g}_{\rho_{i,j},\theta_{i,j}} \qquad (11)$$

In other terms, the watermark is a linear combination of the Gabor responses $\mathbf{g}_{\rho_{i,j},\theta_{i,j}}$. Since the Fourier transform is linear, the same property is also valid in the frequency domain:

$$\begin{aligned}\mathbf{W} &= \sum_{i=1}^{M}\sum_{j=1}^{N} \psi_{i,j}\mathbf{G}_{\rho_{i,j},\theta_{i,j}} \\ &= \left(\sum_{i=1}^{M}\sum_{j=1}^{N} \psi_{i,j}\mathbf{H}_{\rho_{i,j},\theta_{i,j}}\right)\mathbf{I} = \mathbf{HI} \qquad (12)\end{aligned}$$

where $\mathbf{I}$ is the Fourier transform of the input image $\mathbf{i}$. Thus, the watermark can be generated in one row in the frequency domain by computing $\mathbf{H}$ which significantly reduces the computational cost.

4. EXPERIMENTS

The major claim of this paper is that a watermark whose samples have inherited the same linear relationships as the Gabor-defined neighborhoods of the host signal should not be affected by BRA. It is thus necessary to check whether or not the proposed watermark is degraded by such attacks in comparison with more current watermarks e.g. additive SS watermarks. To this end, large-scale experiments have been conducted. The experimental protocol is detailed in Subsection 4.1 and the results are presented in Subsection 4.2.

4.1. Protocol

A watermark with zero mean and unit variance $\mathbf{w}_K(\mathbf{i})$ is embedded in the input image $\mathbf{i}$ to obtain a watermarked image $\mathbf{i}_{\mathrm{w}}$ according to the following equation:

$$\mathbf{i}_{\mathrm{w}} = \mathbf{i} + \alpha\mathbf{w}_K(\mathbf{i}) \qquad (13)$$

where K is a secret key used to generate the watermark and α an embedding strength equal to 3 so that the embedding process results in a distortion about 38.5 dB in terms of Peak Signal to Noise Ratio (PSNR). For SS watermarking, the embedded watermark is completely independent of the host content ($\mathbf{w}_K(\mathbf{i}) = \mathbf{w}_K$). It is pseudo-randomly generated using the secret key K and is normally distributed. On the other hand, the signal coherent watermark presented in Section 3 is related with the host signal $\mathbf{i}$ through the linear form imposed on the Gabor-defined neighborhoods. In the reported experiments, the Gabor filter bank has been configured as follows: $N = 16$, $\rho_{\min} = 0.01$, $\rho_{\max} = 0.45$, $s = 2$ and $t = 1.5$. The number of frequencies M has remained variable.

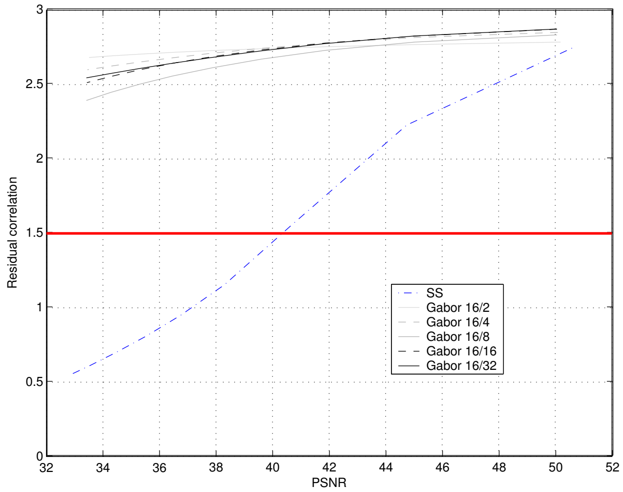

Fig. 1. Comparison of the impact of BRA with SS and Gabor-defined coherent watermarks: whereas the SS watermark is washed out when the attacking strength increases, the coherent watermark survives.

The watermarked image $\mathbf{i}_{\mathrm{w}}$ is then attacked using the latest version of BRA [5]. For each input signal block, a search window is defined and a codebook is built using the blocks within this window. Principal Component Analysis (PCA) is then performed and the obtained eigenvectors are sorted according to their eigenvalues. Finally, the replacement block is obtained by considering more or less eigenvectors (or eigenblocks) so that the distortion with the original signal block is as close as possible to a target value τ_{target}. In the experiments, 8×8 blocks have been considered with an overlap of 4 pixels and the search window size has been set to 64×64. The attacking strength τ_{target} has remained variable.

On the detector side, the only concern is to know whether or not the embedded watermark has survived. Non-blind detection can consequently be considered and the residual correlation is computed as follows:

$$\mathrm{d}(\mathbf{i},\tilde{\mathbf{i}}_{\mathrm{w}}) = (\tilde{\mathbf{i}}_{\mathrm{w}} - \mathbf{i}) \cdot \mathbf{w}_K(\tilde{\mathbf{i}}_{\mathrm{w}}) \qquad (14)$$

where $\tilde{\mathbf{i}}_{\mathrm{w}}$ is the attacked image and $\cdot$ denotes the linear correlation operation. To anticipate future blind detection, the watermark is generated considering the attacked image instead of the original image. This has no impact for SS since it is content independent, but this may have one with signal coherent watermarks. The residual correlation should be equal to α if the watermark has survived while it should drop down to 0 when the watermark signal has been completely washed out. As a result, the presence of the watermark can be asserted by comparing the residual correlation $\mathrm{d}(\mathbf{i},\tilde{\mathbf{i}}_{\mathrm{w}})$ with a detection score τ_{detect} which can be set to $\alpha/2$ for equal false positive and false negative probabilities.

4.2. Experimental Results

A database of 500 images of size 512×512 has been considered for experiments. It contains snapshots, synthetic images, drawings and cartoons. All the image are first watermarked using either SS or signal coherent watermarks with different values for M. Then, each watermarked image is submitted to BRA with varying attacking strength τ_{target} to obtain a distortion vs. residual correlation curve. Finally, all the curves associated with a given watermarking method are averaged to depict the statistical behavior of this scheme against BRA. Those results have been gathered in Figure 1. It should be reminded that the goal of the attacker is to decrease the residual correlation while maintaining the image quality. First of all, the proposed signal coherent watermark clearly outperforms additive SS watermarking when BRA are considered. Indeed, the residual correlation never goes below 2.5 with Gabor watermarks while it already drops below the detection threshold $\tau_{\text{detect}} = 1.5$ for a distortion of 40 dB when SS watermarks are considered. Experiments have also been done to further investigate the influence of the number of GEF pairs in the filter bank on the resilience of the embedded watermark against BRA. To this end, the same benchmark has been run for Gabor watermarks corresponding to different value of M (2, 4, 8, 16, and 32). Even if more images should be considered to allow a pertinent comparison, one can already assert that it has no drastic impact on the immunity of the watermark against BRA. Nevertheless, it is important to increase the number of GEF pairs so that watermarks generated with different secret keys K are as little correlated as possible and thus decrease the false positive probability. Of course, increasing the number of GEF pairs also raises the computational load. Future work will consequently investigated how to obtain a reasonable trade-off.

5. CONCLUSION

After robustness, security evaluation is now growing a key issue in the watermarking community. Indeed, most of the original interest from the industry has come from the potential use of digital watermarking to ensure IP protection. In such applications, customers are likely to be willing to remove the embedded watermark which they can see as a disturbing signal since it reduces the potential usages of protected data. As a result, researchers have to anticipate possible hostile behaviors and, in this perspective, BRA are recognized to be among the most critical attacks against watermarking systems today. Typically, these attacks exploit the fact that *similar blocks do not carry similar watermark* to confuse the watermark detector. In this paper, a novel watermarking strategy has been introduced which has been shown to significantly enhance the survival of the embedded watermark against such attacks. The basic idea consists in removing the weak link exploited by BRA by ensuring that the embedded watermark inherits the self-similarities of the host signal. To this end, the neighborhood is characterized for each pixel using Gabor filters and a linear form is defined in the resulting Gabor space to obtain a signal coherent watermark.

From a more general points of view, this can be seen as some kind of informed watermarking [1, 14]. Digital watermarking can indeed be seen as moving a point in a high dimentional media space to a nearby location i.e. introducing a small displacement in a random direction. The introduced framework only stipulates that the host signal self-similarities have to be considered to resist BRA and that, in this case, some of the possible directions are now prohibited. Future work will investigate how to properly configure the Gabor filter bank and how to design a blind detector with such signal coherent watermarks i.e. a detector which does not require the original image $\mathbf{i}$ to assert whether a watermark is present or not. Furthermore, looking at Equation (12) more closely, it can be noted that the watermark generation process can be seen as a multiplication in the frequency domain. Thus, it may be interesting to revisit the results previously obtained with such a watermarking strategy under this new light [15, 16].

6. REFERENCES

[1] I. Cox, M. Miller, and J. Bloom, *Digital Watermarking*, Morgan Kaufmann Publishers, 2001.

[2] DVD Copy Control Association, “http://www.dvdcca.org,” .

[3] Secure Digital Music Initiative, “http://www.sdmi.org,” .

[4] G. Doërr and J.-L. Dugelay, “Security pitfalls of frame-by-frame approaches to video watermarking,” *IEEE Transactions on Signal Processing, Supplement on Secure Media*, vol. 52, no. 10, pp. 2955–2964, October 2004.

[5] G. Doërr, J.-L. Dugelay, and L. Grangé, “Exploiting self-similarities to defeat digital watermarking systems - a case study on still images,” in *Proceedings of the ACM Multimedia and Security Workshop*, September 2004, pp. 133–142.

[6] D. Kirovski and F. Petitcolas, “Blind pattern matching attack on watermarking systems,” *IEEE Transactions on Signal Processing*, vol. 51, no. 4, pp. 1045–1053, April 2003.

[7] J. Daugman, “Complete discrete 2-D Gabor transforms by neural network for image analysis and compression,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 36, no. 7, pp. 1169–1179, July 1988.

[8] D. Dunn, W. Higgins, and J. Wakeley, “Texture segmentation using 2-D Gabor elementary functions,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 2, pp. 130–149, February 1994.

[9] B. Duc, S. Fisher, and J. Bigün, “Face authentication with Gabor information on deformable graphs,” *IEEE Transactions on Image Processing*, vol. 8, no. 4, pp. 504–516, April 1999.

[10] G. Donato, M. Bartlett, J. Hager, P. Ekman, and T. Sejnowski, “Classifying facial actions,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 10, pp. 974–989, October 1999.

[11] D. Ringach, “Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex,” *Journal of Neurophysiology*, vol. 88, no. 1, pp. 455–463, July 2002.

[12] C. Rey, G. Doërr, J.-L. Dugelay, and G. Csurka, “Toward generic image dewatermarking?,” in *Proceedings of the IEEE International Conference on Image Processing*, September 2002, vol. III, pp. 633–636.

[13] G. Doërr, *Security Issue and Collusion Attacks in Video Watermarking*, Ph.D. thesis, Université de Nice Sophia-Antipolis, France, June 2005.

[14] J. Eggers and B. Girod, *Informed Watermarking*, Kluwer Academic Publishers, 2002.

[15] M. Barni, F. Bartolini, A. De Rosa, and A. Piva, “A new decoder for optimum recovery of nonadditive watermarks,” *IEEE Transactions on Image Processing*, vol. 10, no. 5, pp. 755–766, May 2001.

[16] Q. Cheng and T. Huang, “Robust optimum detection of transform domain multiplicative watermarks,” *IEEE Transactions on Signal Processing*, vol. 51, no. 4, pp. 906–924, April 2003.