How to Combat Block Replacement Attacks?

Gwenaël Doërr and Jean-Luc Dugelay

Eurécom Institute
Multimedia Communications Department
2229 route des Crêtes – B.P. 193
06904 Sophia-Antipolis Cédex, France
{doerr, dugelay}@eurecom.fr
http://www.eurecom.fr/∼image

Abstract. Block replacement attacks consist in exploiting the redundancy of the host signal to replace each signal block with another one or a combination of other ones. Such an attacking strategy has been recognized to be a major threat against watermarking systems e.g. additive spread-spectrum and quantization index modulation algorithms. In this paper, a novel embedding strategy will be introduced to circumvent this attack. The basic idea is to make the watermark inherit the self-similarities from the host signal. This can be achieved by imposing a linear structure on the watermark in a feature space e.g. the Gabor space. The relationship with existing multiplicative watermarking schemes will also be exhibited. Finally, experimental results will be presented and directions for future work will be discussed.

1 Introduction

Digital watermarking was initially introduced in the early 90's as a complementary protection technology [1] since encryption alone is not enough. Indeed, sooner or later, encrypted multimedia content is decrypted to be eventually presented to human beings. At this very moment, multimedia content is left unprotected and can be perfectly duplicated, manipulated and redistributed at a large scale. Thus, a second line of defense has to be added to address this issue. This is the main purpose of digital watermarking which basically consists in hiding some information into digital content in an imperceptible manner. Up to now, research has mainly investigated how to improve the trade-off between three conflicting parameters: imperceptibility, robustness and capacity. Perceptual models have been exploited to make watermarks less perceptible, benchmarks have been released to evaluate robustness, channel models have been studied to obtain a theoretical bound for the embedding capacity.

A lot of attention has focused on security applications such as Intellectual Property (IP) protection and Digital Rights Managements (DRM) systems. Digital watermarking was even thought of as a possible solution to combat illegal copying which was a forthcoming issue in the mid-90's. However the few attempts to launch watermarking-based copy-control mechanisms [2, 3] have resulted in partial failures, which have significantly lowered the initial enthusiasm

for this technology. These setbacks were mainly due to the claim that embedded watermarks would survive in a highly hostile environment even if very few works addressed this issue. Indeed, if the survival of the watermark against common signal processing primitives - filtering, lossy compression, global desynchronization - has been carefully surveyed, almost no work has considered that an attacker may exploit some knowledge on the watermarking systems to defeat it. Nevertheless, in applications such as copy control or fingerprinting, digital watermarking is usually seen as a disturbing technology. If content owners are glad to have means to protect their high valued multimedia items, customers on the other hand do not really appreciate that some hidden signal prevent them from freely copying digital material or that an invisible watermark identifies them as a source of leakage. Therefore, this protecting technology is likely to be submitted to strong hostile attacks when it is released to the public.

Security evaluation is now a growing concern in the watermarking community since recent studies have highlighted that most watermarking systems can be defeated by malicious attackers [4, 5]. In particular, collusion attacks have often been mentioned as a possible mean to evaluate security [6, 7]. Collusion consists in collecting several watermarked documents and combining them to obtain unwatermarked content. There are two basic cases. When different contents are watermarked with some kind of structure, colluders try to estimate this structure and exploit this knowledge in a second step to remove the watermark [8]. Alternatively, when similar contents carry uncorrelated watermarks, colluders can average them so that watermark samples sum to zero. Block Replacement Attacks (BRA) consist in replacing each signal block with another one or a combination of other ones and can thus be seen as an extension of this later strategy. BRA have been shown to defeat both additive Spread-Spectrum (SS) and Quantization Index Modulation (QIM) [9] and will thus be rapidly reviewed in Section 2. A novel embedding strategy is then designed in Section 3 to circumvent this attack by making the watermark inherit the self-similarities of the host signal. This is done by forcing a linear structure on the watermark in a feature space e.g. the Gabor space. At this point, an analogy between this new approach and previous multiplicative embedding schemes [10, 11] can even be exhibited. Next, the resilience of these signal coherent watermarks against BRA is evaluated in Section 4 in comparison with standard additive SS watermarks. Finally, conclusions are drawn in Section 5 and tracks for future work are given.

2 Block Replacement Attacks

Multimedia digital data is highly redundant: successive video frames are highly similar in a movie clip, most songs contain some repetitive patterns, etc. An attacker can consequently exploit these similarities to successively replace each part of the signal with a *similar* one taken from another location in the same signal. In particular, such approaches have already been investigated to obtain efficient compression tools [12]. The signal to be processed is first partitioned into a set of blocks $\mathbf{b}_T$ of size S_T. Those blocks can either overlap or not. The

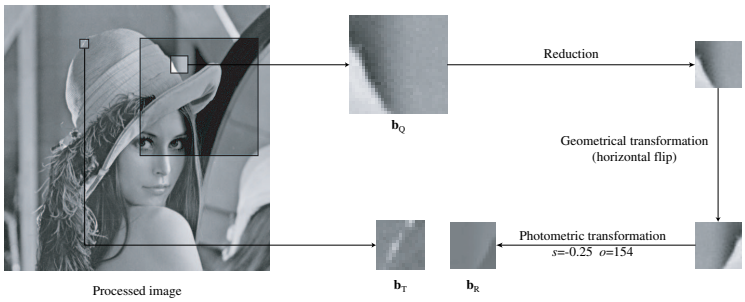

Fig. 1. BRA implementation using a fractal coding strategy: each block is replaced by the one in the search window which is the most similar modulo a geometrical and photometric transformation.

asset of using overlapping blocks is that it prevents strong blocking artifacts on the border of the blocks by averaging the overlapping areas. The attack processes then each one of these blocks sequentially. For each block, a search window is defined. It can be chosen in the vicinity of the target block $\mathbf{b}_T$ or randomly for security reasons. This search window is partitioned to obtain a codebook $\mathcal{Q}$ of blocks $\mathbf{b}_{Q_i}$ of size S_Q. Once again, these blocks can overlap or not. Next a candidate block for replacement $\mathbf{b}_R$ is computed using the blocks present in the codebook. Of course, the larger the codebook $\mathcal{Q}$ is, the more choices there are to compute a replacement block which is *similar* enough to the input block $\mathbf{b}_T$ so that they can be substituted without introducing strong visual artifacts. On the other hand, the larger the codebook $\mathcal{Q}$ is, the higher the computational complexity is and a trade-off has to be found. The Mean Square Error (MSE) can be used to evaluate how similar are two blocks with the following formula:

$$\mathrm{MSE}(\mathbf{b}_R, \mathbf{b}_T) = \frac{1}{S_T} \sum_{i=1}^{S_T} \big(\mathbf{b}_R(i) - \mathbf{b}_T(i)\big)^2, \tag{1}$$

where the summation index i can be one-dimensional (sound) or multidimensional (image, video). The lower the MSE is, the more similar are the two blocks. Thus, the original block $\mathbf{b}_T$ is substituted by the replacement block $\mathbf{b}_R$ associated with the lowest MSE.

There are many ways of computing the replacement block $\mathbf{b}_R$. One of the first proposed implementation was based on fractal coding [13] and is illustrated in Figure 1. The codebook is first artificially enlarged by also considering geometrically transformed versions of the blocks within the search window. For complexity reasons, a small number of transformations are considered e.g. downsampling by a factor 2 and 8 isometries (identity, 4 flips, 3 rotations). Next, the candidate replacement blocks are computed with a simple affine photometric compensation. In other terms, each block $\mathbf{b}_{Q_i}$ of the codebook is transformed in $s\mathbf{b}_{Q_i} + o\mathbf{1}$, where $\mathbf{1}$ is a block containing only ones, so that the MSE with the

target block $\mathbf{b}_T$ is minimized. This is a simple least squares problem and the scale s and offset o can be determined as follows:

$$s = \frac{(\mathbf{b}_T - \mathrm{m}_T\mathbf{1}) \cdot (\mathbf{b}_{Q_i} - \mathrm{m}_{Q_i}\mathbf{1})}{|\mathbf{b}_{Q_i} - \mathrm{m}_{Q_i}\mathbf{1}|^2} \tag{2}$$

$$o = \mathrm{m}_T - s.\mathrm{m}_{Q_i} \tag{3}$$

where m_T (resp. m_{Q_i}) is the mean value of block $\mathbf{b}_T$ (resp. $\mathbf{b}_{Q_i}$), $\cdot$ is the linear correlation defined as:

$$\mathbf{b} \cdot \mathbf{b}' = \frac{1}{S_T} \sum_{i=1}^{S_T} \mathbf{b}(i)\mathbf{b}'(i) \tag{4}$$

and $|\mathbf{b}|$ is the norm defined as $\sqrt{\mathbf{b} \cdot \mathbf{b}}$. At this point, the transformed blocks $s\mathbf{b}_{Q_i} + o\mathbf{1}$ are sorted in ascending order according to their similarity with the target block $\mathbf{b}_T$ and the most similar one is retained for replacement. In the same fashion, an alternative approach consists in building iteratively sets of similar blocks and randomly shuffling their positions [14, 9] until all the blocks have been replaced.

The baseline of the algorithm has then been improved to further enhance the performances of the attack. The main drawback of the previous implementation is that it is not possible to modify the strength of the attack. Furthermore, the computation of the replacement block is not properly managed: either it is too close from the target block $\mathbf{b}_T$ and the watermark is reintroduced, or it is too distant and strong visual artifacts appear. Optimally, one would like to ensure that the distortion $\Delta = \mathrm{MSE}(\mathbf{b}_R, \mathbf{b}_T)$ remains within two bounds τ_{low} and τ_{high}. To this end, several blocks $\mathbf{b}_{Q_i}$ can be combined to compute the replacement block instead of a single one as follows:

$$\mathbf{b}_R = \sum_{i=1}^{N} \lambda_i \mathbf{b}_{Q_i} \tag{5}$$

where the λ_i are mixing parameter chosen in such a way that Δ is minimized. This combination can take into account a fixed number of blocks [15] or also adapt the number of considered blocks for combination according to the nature of the block to be reconstructed [16]. Intuitively, approximating flat blocks require to combine fewer blocks than for highly textured ones. However, the computational load induced by computing optimal mixing parameters in Equation (5) for each candidate replacement block has motivated the design of an alternative implementation which is described in Table 1 [16]. First, for each block $\mathbf{b}_T$, the codebook $\mathcal{Q}$ is built and photometric compensation is performed. Next, a Principal Component Analysis (PCA) is performed considering the different blocks $\mathbf{b}_{Q_i}$ in the codebook. This gives a centroid $\mathbf{c}$ defined as follows:

$$\mathbf{c} = \frac{1}{|\mathcal{Q}|} \sum_{\mathbf{b}_{Q_i} \in \mathcal{Q}} \mathbf{b}_{Q_i} \tag{6}$$

Table 1. BRA procedure using block projection on a PCA-defined subspace.

For each block $\mathbf{b}_T$ of the signal
1 Build the block codebook $\mathcal{Q}$
2 Perform photometric compensation
3 Performs the PCA of the blocks in $\mathcal{Q}$ to obtain a set of orthogonal eigenblocks $\mathbf{e}_i$ associated with their eigenvalues ϵ_i
Set $N = 1$ and flag $= 0$
4 While (flag $= 0$) AND ($N \leq S_T$)
(a) Build the optimal replacement block $\mathbf{b}_R$ using the eigenblocks $\mathbf{r}_i$ associated with the first N eigenvalues
(b) Compute $\Delta = \text{MSE}(\mathbf{b}_R, \mathbf{b}_T)$
(c) If $\tau_{\text{low}} \leq \Delta \leq \tau_{\text{high}}$, set flag $= 1$
(d) Else increment N
5 Replace $\mathbf{b}_T$ by $\mathbf{b}_R$

and a set of eigenblocks $\mathbf{e}_i$ associated with their eigenvalues ϵ_i. These eigenblocks are then sorted by descending eigenvalues i.e. the direction $\mathbf{e}_1$ contains more information than any other one in the basis. Then, a candidate block for replacement $\mathbf{b}_R$ is computed using the N first eigenblocks so that the distortion Δ is minimized. In other terms, the block $\mathbf{b}_T - \mathbf{c}$ is projected onto the subspace spanned by the N first eigenblocks and $\mathbf{b}_R$ can be written:

$$\mathbf{b}_R = \mathbf{c} + \sum_{i=1}^{N} \frac{(\mathbf{b}_T - \mathbf{c}) \cdot \mathbf{e}_i}{|\mathbf{e}_i|^2} \mathbf{e}_i \tag{7}$$

Of course, the distortion Δ gracefully decreases as the number N of combined eigenblocks increases. Thus, an adaptive framework is introduced to identify which value N should have so that the distortion Δ falls within the range $[\tau_{\text{low}}, \tau_{\text{high}}]$. It should be noted that the underlying assumption is that most of the watermark energy will be concentrated in the last eigenblocks since the watermark can be seen as details. As a result, if a valid candidate block can be built without using the last eigenblocks, the watermark signal will not be reintroduced.

3 Signal Coherent Watermarks

As reminded in the previous section, for each signal block, BRA look for a linear combination of neighboring blocks resulting in a block which is similar enough to the current block so that a substitution does not introduce strong visual artifacts. Since watermarking systems do not perform today anything specific to ensure that the embedded watermark is coherent with the self-similarities of the host signal, most of them are defeated by such attacks. Intuitively, to ensure that the watermark will survive BRA, the embedding process should guarantee that *similar signal blocks carry similar watermarks* or alternatively that *pixels with*

similar neighborhood carry watermark samples with close values. In this perspective, assuming that it is possible to characterize the neighborhood in each point with a feature vector, signal coherent watermarking can be achieved if watermark samples are considered as the output of a linear form in this feature space as it is theoretically demonstrated in Subsection 3.1. A practical implementation of this approach using Gabor features is then described in Subsection 3.2. Finally, a relationship with existing multiplicative watermarking scheme in the frequency space is exhibited in Subsection 3.3.

3.1 Linear Watermarking with Neighborhood Characteristics

Let us assume for the moment that it is possible to associate to each pixel position $\mathbf{p} = (x, y)$ with $1 \leq x \leq X$ and $1 \leq y \leq Y$ in the image $\mathbf{i}$ a feature vector $\mathbf{f}(\mathbf{i}, \mathbf{p})$ which characterizes *in some sense* the neighborhood of the image around this specific position. Thus, this function can be defined as follows:

$$\begin{aligned} \mathbf{f} : \mathcal{I} \times \mathcal{P} &\rightarrow \mathcal{F} \\ (\mathbf{i}, \mathbf{p}) &\mapsto \mathbf{f}(\mathbf{i}, \mathbf{p}) \end{aligned} \tag{8}$$

where $\mathcal{I}$ is the image space, $\mathcal{P} = [1 \dots X] \times [1 \dots Y]$ the position space and $\mathcal{F}$ the feature space. From a very low-level perspective, generating a digital watermark can be regarded as associating a watermark value $\mathrm{w}(\mathbf{i}, \mathbf{p})$ to each pixel position in the image. However, if the embedded watermark is required to be immune against BRA, the following property should also be verified:

$$\mathbf{f}(\mathbf{i}, \mathbf{p}_0) \approx \sum_k \lambda_k \mathbf{f}(\mathbf{i}, \mathbf{p}_k) \Rightarrow \mathrm{w}(\mathbf{i}, \mathbf{p}_0) \approx \sum_k \lambda_k \mathrm{w}(\mathbf{i}, \mathbf{p}_k) \tag{9}$$

In other terms, if at a given position $\mathbf{p}_0$, the local neighborhood is similar to a linear combination of neighborhoods at other locations $\mathbf{p}_k$, then the watermark sample $\mathrm{w}(\mathbf{p}_0)$ embedded at position $\mathbf{p}_0$ should be close to the linear combination (with the same mixing coefficients λ_k) of the watermark samples $\mathrm{w}(\mathbf{p}_k)$ at these locations. A simple way to obtain this property is to make the watermarking process be the composition of a feature extraction operation and a linear form φ.

Hence, one can write $\mathrm{w} = \varphi \circ \mathbf{f}$ where $\varphi : \mathcal{F} \rightarrow \mathbb{R}$ is a linear form which takes F-dimensional feature vectors in input. Next, to completely define this linear form, it is sufficient to set the values $\xi_f = \varphi(\mathbf{b}_f)$ for a given orthonormalized basis $\mathcal{B} = \{\mathbf{b}_f\}$ of the feature space $\mathcal{F}$. Without loss of generality, one can consider the canonical basis $\mathcal{O} = \{\mathbf{o}_f\}$ where $\mathbf{o}_f$ is a F-dimensional vector filled with 0's except the fth coordinate which is equal to 1. The whole secret of the algorithm is contained in the values ξ_f and they can consequently be pseudo-randomly generated using a secret key K. 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:

$$\mathrm{f}_{\varphi|\mathcal{U}}(w) = \frac{1}{\Xi\sqrt{\pi}} \frac{\Gamma\left(\frac{F}{2}\right)}{\Gamma\left(\frac{F-1}{2}\right)} \left[1 - \left(\frac{w}{\Xi}\right)^2\right]^{\frac{F-3}{2}} \tag{10}$$

where $\Xi^2 = \sum_{f=1}^{F} \xi_f^2$ and $\Gamma(.)$ is the Gamma function. When the dimension F of the feature space $\mathcal{F}$ grows large, this probability density function tends towards a Gaussian distribution with zero mean and standard deviation $\Xi/\sqrt{F}$. Thus if the ξ_f's are chosen to have zero mean and unit variance, this ensures that the values of the linear form restricted to the unit sphere $\mathcal{U}$ are normally distributed with also zero mean and unit variance. Then, keeping in mind that φ is linear and that the following equation is valid,

$$\mathrm{w}(\mathbf{i},\mathbf{p}) = \varphi\left(\|\mathbf{f}(\mathbf{i},\mathbf{p})\|\frac{\mathbf{f}(\mathbf{i},\mathbf{p})}{\|\mathbf{f}(\mathbf{i},\mathbf{p})\|}\right) = \|\mathbf{f}(\mathbf{i},\mathbf{p})\|\varphi\big(\mathbf{u}(\mathbf{i},\mathbf{p})\big) \quad \text{with } \mathbf{u}(\mathbf{i},\mathbf{p}) \in \mathcal{U} \quad (11)$$

it is straightforward to realize that the obtained watermark is equivalent to a Gaussian watermark with zero mean and unit variance multiplied by some local scaling factors. The more textured is the considered neighborhood, the more complicated it is to characterize it and the greater the norm $\|\mathbf{f}(\mathbf{i},\mathbf{p})\|$ is likely to be. Looking back at Equation 11, it results that the watermark is amplified in textured area whereas it is attenuated in smooth ones. This can be regarded as some kind of perceptual shaping [17].

3.2 A Practical Implementation Using Gabor Features

In order to impose a linear relationship between watermark samples with respect to some characteristics of the neighborhood, it is first necessary to define the features which will be used to differentiate between neighborhoods i.e. it is needed to define the feature extraction function $\mathbf{f}$ mentioned in Equation (8). In this perspective, 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 [18], texture segmentation [19], face authentication [20] and facial analysis [21]. Images are classically viewed either as a collection of pixels (spatial domain) or as a sum of sinusoids of infinite extent (frequency domain). But these representations are just two opposite extremes in a continuum of possible joint space/frequency representations. Indeed, frequency can be viewed as a local phenomenon that may vary with position throughout the image. Moreover, Gabor wavelets have also received an increasing interest in image processing since they are particularly close to 2-D receptive fields profiles of the mammalian cortical simple cells [22].

A Gabor Elementary Function (GEF) $\mathbf{h}_{\rho,\theta}$ is defined by a radius ρ and an orientation θ and the response of an input image $\mathbf{i}$ to such a GEF can be computed as follows:

$$\mathbf{g}_{\rho,\theta} = \mathbf{i} * \mathbf{h}_{\rho,\theta} \quad (12)$$

where $*$ denotes convolution and $\mathbf{g}_{\rho,\theta}$ is the resulting filtered image. The GEF is a complex 2D sinusoid whose orientation and frequency are given by (θ, ρ) restricted by a Gaussian envelope. For computational complexity reasons, Gabor filtering is usually performed in the Fourier domain since it then comes down to

a simple multiplication with the following filter:

$$\mathbf{H}_{\rho,\theta}(u,v) = \exp\left[-\frac{1}{2}\left(\left(\frac{u'-\rho}{\sigma_\rho}\right)^2 + \left(\frac{v'}{\sigma_\theta}\right)^2\right)\right]$$
$$\text{with} \quad \begin{pmatrix} u' \\ v' \end{pmatrix} = \mathbf{R}_\theta \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} \tag{13}$$

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 $\mathbf{H}_{\rho,\theta} \leftarrow \mathbf{H}_{\rho,\theta} + \mathbf{H}_{\rho,\theta+\pi}$.

A single GEF pair associates to each pixel $\mathbf{p}$ of the image a single feature value $\mathbf{g}_{\rho,\theta}(\mathbf{i},\mathbf{p})$. As a result, the idea is now to design a filter bank of such GEF pairs to obtain for each pixel a multi-dimensional feature vector $\mathbf{g}(\mathbf{i},\mathbf{p}) = \{\mathbf{g}_{\rho,\theta}(\mathbf{i},\mathbf{p})\}$ with $1 \leq i \leq M$ and $1 \leq j \leq N$. Based on previous work [20], the different parameters of the GEF pairs are computed as follows:

$$\rho_{i,j} = \rho_{\min} + b\frac{(s+1)s^{i-1}-2}{s-1} \tag{14}$$
$$\sigma_{\rho_{i,j}} = tbs^{i-1} \tag{15}$$
$$\theta_{i,j} = \frac{(j-1)\pi}{N} \tag{16}$$
$$\sigma_{\theta_{i,j}} = t\frac{\pi\rho_{i,j}}{2N} \tag{17}$$
$$b = \frac{\rho_{\max}-\rho_{\min}}{2}\left(\frac{s-1}{s^M-1}\right) \tag{18}$$

The whole filter bank is specified by the 6 parameters M, N, $\rho_{\min}$, $\rho_{\max}$, s and t. The first two parameters determine respectively the number of orientations and frequencies in the filter bank. The next two ones specify the bandwidth within which the GEFs are bound. The parameter s controls how much the radial bandwidth increases when the radius increases. For instance, 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. As an example, with $t = 1$, they cross at equal value $1/e$ along their principal axis. Figure 2 depicts how GEFs are scattered throughout a specified frequency ring in the Fourier domain.

In each pixel position $\mathbf{p}$, the resulting MN-dimensional vector $\mathbf{g}(\mathbf{i},\mathbf{p})$ can be regarded as the local power spectrum of the image and thus be used to characterize the neighborhood. It should be noted that if the Gabor filter bank is properly designed, it is possible to impose higher constraints. For instance, if the fractal approach depicted in Figure 1 is enforced, neighborhoods which are the same modulo a small set of geometrical operations, e.g. 8 isometries and downsampling by a factor 2, are required to carry the same watermark samples

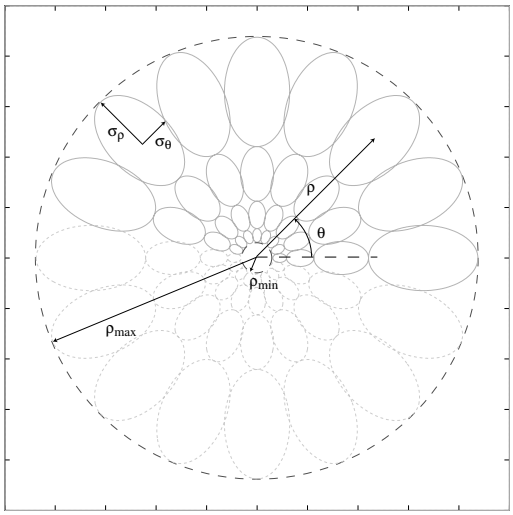

Fig. 2. Graphical representation in the Fourier domain of the GEFs levelset for value $1/e$ with $M = 8$, $N = 4$, $s = 2$ and $t = 1$.

to achieve robustness [13]. Such constraints need to be taken into account to define the kernel of the linear form φ i.e. the non null vectors $\mathbf{v}$ for which $\varphi(\mathbf{v}) = 0$. However, more constraints induce a lower dimensional subspace for watermarking which can rapidly become critical.

3.3 Analogy with Multiplicative Watermarking Schemes

Since the values ξ_f of the linear form φ are defined on the canonical basis $\mathcal{O}$ when Gabor features are considered, the watermark sample obtained at position $\mathbf{p}$ is simply given by:

$$\mathrm{w}(\mathbf{i}, \mathbf{p}) = \sum_{f=1}^{F} \xi_f \mathbf{g}_f(\mathbf{i}, \mathbf{p}) \tag{19}$$

where $\mathbf{g}_f(\mathbf{i}, \mathbf{p})$ is the fth coordinate of the F-dimensional Gabor feature vector $\mathbf{g}(\mathbf{i}, \mathbf{p})$. In other terms, the watermark is a linear combination of different Gabor responses $\mathbf{g}_f$. However, when M and N grow, more and more Gabor responses need to be computed which can be quickly computationally prohibitive. Hopefully, when the Fourier domain is considered, the watermark can be computed as follows:

$$\begin{aligned}\mathbf{W}(\mathbf{i}, \mathbf{q}) &= \sum_{\mathbf{p} \in \mathcal{P}} \left(\sum_{f=1}^{F} \xi_f \, \mathbf{g}_f(\mathbf{i}, \mathbf{p}) \right) \omega_{\mathbf{p},\mathbf{q}} \\ &= \sum_{f=1}^{F} \xi_f \left(\sum_{\mathbf{p} \in \mathcal{P}} \mathbf{g}_f(\mathbf{i}, \mathbf{p}) \, \omega_{\mathbf{p},\mathbf{q}} \right) = \sum_{f=1}^{F} \xi_f \, \mathbf{G}_f(\mathbf{i}, \mathbf{q})\end{aligned}$$

$$= \sum_{f=1}^{F} \xi_f \, \mathbf{H}_f(\mathbf{q}) \, \mathbf{I}(\mathbf{q}) = \mathbf{H}(K, \mathbf{q}) \, \mathbf{I}(\mathbf{q}) \tag{20}$$

$$\text{with} \quad \mathbf{H}(K, \mathbf{q}) = \sum_{f=1}^{F} \xi_f \, \mathbf{H}_f(\mathbf{q})$$

where $\omega_{\mathbf{p},\mathbf{q}} = \exp\left[-j2\pi\left((x-1)(u-1)/X + (y-1)(v-1)/Y\right)\right]$, capital letters indicate FFT-transformed variables and $\mathbf{q} = (u, v)$ denotes a frequency position with $1 \leq u \leq U$ and $1 \leq v \leq V$. In other terms, the watermark can be generated in one row in the Fourier domain by computing $\mathbf{H}$ and such an approach is likely to significantly reduce the computational cost.

Looking closely at Equation (20), it is straightforward to realize that the watermark generation process comes down to a simple multiplication between the image spectrum $\mathbf{I}$ and some pseudo-random signal $\mathbf{H}(K)$. In other terms, it really looks similar to basic well-known multiplicative embedding schemes in the frequency domain [10, 11]. When the bandwidth of a GEF is close to 0, the 2D Gaussian in the Fourier domain tends toward a Dirac impulse centered at coordinates (ρ, θ) i.e. it tends toward an infinite sinusoid in the spatial domain. Therefore, multiplicative embedding in the FFT domain[1] is equivalent to imposing a linear relationship on the watermark samples according to the neighborhood which is characterized by its response to infinite sinusoids. Under this new light, FFT multiplicative watermarks can be seen as a special case of the Gabor watermarks introduced in Subsection 3.2 and are thus coherent with the host signal. Next, keeping in mind that DCT coefficients are simply FFT coefficients of some periodic image [23], it is immediate to assert that DCT multiplicative watermarks [10] are also signal coherent watermarks. At this point, it is interesting to note that multiplicative watermarking in the frequency domain was initially motivated by contrast masking properties: larger coefficients can convey a larger watermark without compromising invisibility [24]. This can be related with the natural perceptual shaping of signal coherent watermarks exhibited in Equation (11).

4 Experiments

The major claim in this paper is that a watermark whose samples have inherited the same linear relationships as the neighborhoods of the host signal should not be affected by BRA. An embedding scheme using Gabor features has been designed in Subsection 3.2 so that the generated watermark exhibits this property. Moreover, it has been shown in Subsection 3.3 that previous embedding schemes based on multiplicative embedding in the frequency space is also likely to resist

[1] In this paper, multiplicative embedding in the FFT domain means that the *complex* FFT coefficients are multiplied by pseudo-random values. It is slightly different from the algorithm described in [11] where only the *magnitude* of the FFT coefficients were watermarked.

BRA. It is now necessary to check whether or not these identified watermarks are degraded by such attacks in comparison with more current watermarks e.g. additive SS watermarks in the spatial domain. 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 embedding rule:

$$\mathbf{i}_{\mathrm{w}} = \mathbf{i} + \alpha \mathbf{w}(K, \mathbf{i}) \tag{21}$$

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). Four different watermark generation processes will be surveyed during the experiments:

SS: The embedded watermark is completely independent of the host content i.e. $\mathbf{w}(K, \mathbf{i}) = \mathbf{r}(K)$ where $\mathbf{r}(K)$ is a pseuso-random pattern which is generated using the secret key K and which is normally distributed with zero mean and unit variance.

Gabor: The generation process considers Gabor features to make the watermark inherit the self-similarities of the host signal. As discussed in Subsection 3.3, the watermark is generated in the Fourier domain using Equation (20) i.e. $\mathbf{W}(K, \mathbf{i}) = \mathbf{H}(K)\,\mathbf{I}$. Inverse FFT is then performed to come back to the spatial domain and the resulting watermark is scaled to have unit variance. In the reported experiments, the Gabor filter bank has been configured as follows: $M = 32$, $N = 16$, $\rho_{\min} = 0.01$, $\rho_{\max} = 0.45$, $s = 2$ and $t = 1.5$. Former investigations have demonstrated that the number MN of considered GEF pairs does not have a drastic impact on the performances of the algorithm with respect to the resilience against BRA [25].

FFT: The watermark is generated in the Fourier domain as follows $\mathbf{W}(K, \mathbf{i}) = \dot{\mathbf{r}}(K)\,\mathbf{I}$ where $\dot{\mathbf{r}}(K)$ is a pseudo-random pattern which is symmetric with respect to the center of the Fourier domain and which has value 0 at the DC coefficient position. This property has to be verified so that the resulting watermark is real-valued with zero mean after inverse transform. Once again, inverse FFT is performed to come back to the spatial domain and the resulting watermark is scaled to have unit variance. This algorithm can be regarded as an extension of the previous one when the GEFs are reduced to Dirac impulses in the frequency domain.

DCT: The watermark is generated in the frequency domain using the following formula $\hat{\mathbf{W}}(K, \mathbf{i}) = \mathbf{r}(K)\,\hat{\mathbf{I}}$ where "capital hat" denotes the DCT transform and $\mathbf{r}(K)$ is a normally distributed pseudo-random pattern which has value 0 at the DC coefficient position. Inverse DCT is then performed to come back to the spatial domain and the resulting watermark is scaled to have unit variance.

Next, the watermarked image $\mathbf{i}_\mathrm{w}$ is attacked using the version of BRA described in Table 1. 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. Furthermore, the two thresholds τ_low and τ_high have been set equal to the same value τ_target. As a result, the replacement block is obtained by considering more or less eigenblocks so that the distortion with the original signal block is as close as possible to the target value τ_target. This threshold can be used as an attacking strength which can be modified during experiments.

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

$$\mathrm{d}(\mathbf{i}, \tilde{\mathbf{i}}_\mathrm{w}) = (\tilde{\mathbf{i}}_\mathrm{w} - \mathbf{i}) \cdot \mathbf{w}(K, \tilde{\mathbf{i}}_\mathrm{w}) \tag{22}$$

where $\tilde{\mathbf{i}}_\mathrm{w}$ is the attacked image and $\cdot$ denotes the linear correlation operation. To anticipate future blind detection, the detector generates the watermark using 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 images are first watermarked using one of the watermarking system under study i.e. SS, Gabor, FFT or DCT. This results in 4 collections of 500 watermarked images each. 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 3. It should be reminded that the goal of the attacker is to decrease the residual correlation while maintaining the image quality. First of all, experimental results clearly show that signal coherent watermarking has a strong impact on the efficiency of BRA. As a matter of fact, the residual correlation never goes below 2.5 with signal coherent watermarks (Gabor, FFT or DCT) while it already drops below the detection threshold $\tau_\mathrm{detect} = 1.5$ for a distortion of 40 dB when SS watermarks are considered. Moreover, even if experiments at a larger scale should be carried out for a pertinent comparison, some kind of *ranking* appears amongst the signal coherent watermarking schemes. The observation that FFT behaves better than Gabor may be explained by the fact that the first algorithm is an extension of the second one. Therefore, the FFT curve would give some bound for the achievable performances with the Gabor scheme

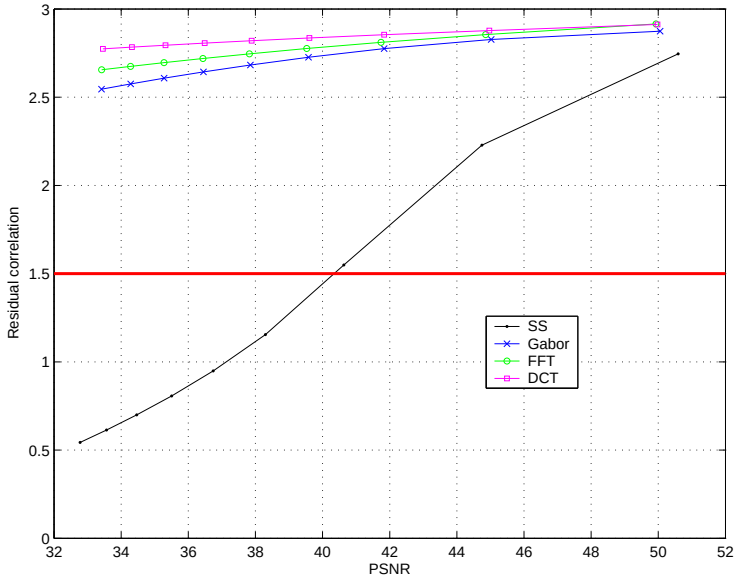

Fig. 3. Comparison of the impact of BRA with the 4 watermarking schemes under study: whereas non coherent watermarks (SS) are washed out when the attacking strength increases, coherent watermarks (Gabor/FFT/DCT) survive.

for different configurations. Finally, the superiority of DCT over FFT might be due to the properties of the DCT which ensure that the watermark will not be embedded in *fake* image frequencies revealed by the Fourier transform [24].

5 Conclusion

Security evaluation is now a growing concern in the watermarking community. Consumers are likely to attack the embedded watermark which they see as a disturbing signal and researchers have to anticipate these possible hostile behaviors. In this perspective, BRA are recognized to be among the most critical operations against watermarking systems today. Typically, these attacks exploit the fact that *similar blocks do not carry similar watermarks* to confuse the watermark detector. In this paper, a novel watermarking strategy has been investigated to remove this weak link. It basically aims at making the embedded watermark inherit the self-similarities of the host signal. Features are extracted in each pixel position to characterize the neighborhood and are exploited to export linear relationships between neighborhoods to watermark samples. A practical implementation using Gabor features has been presented and previous multiplicative embedding schemes in the frequency domain [10, 11] have been shown to also produce signal-coherent watermarks even if, to the best knowledge of the authors, such a property has never been foreseen in previous works.

From a more general points of view, signal coherent watermarking can be seen as some kind of informed watermarking [1, 26]. Digital watermarking can be seen as moving a point in a high dimensional 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 explore how former works [11, 27] can be used to design a blind detector for signal coherent watermarks. Furthermore, security investigations will be conducted to determine whether or not an attacker can gain some knowledge about the imposed watermarking structure. Indeed, using a redundant watermarking structure has been demonstrated to lead to security pitfalls in the past [7, 8].

6 Acknowledgment

This work has been supported in part by the European Commission through the IST Program under Contract IST-2002-507932 ECRYPT.

References

1. Cox, I., Miller, M., Bloom, J.: 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. Cayre, F., Fontaine, C., Furon, T.: Watermarking security, part I: Theory. In: Security, Steganography and Watermarking of Multimedia Contents VII. Volume 5681 of Proceedings of SPIE. (2005) 746–757
5. Cayre, F., Fontaine, C., Furon, T.: Watermarking security, part II: Practice. In: Security, Steganography and Watermarking of Multimedia Contents VII. Volume 5681 of Proceedings of SPIE. (2005) 758–768
6. Su, K., Kundur, D., Hatzinakos, D.: A novel approach to collusion resistant video watermarking. In: Security and Watermarking of Multimedia Contents IV. Volume 4675 of Proceedings of SPIE. (2002) 491–502
7. Doërr, G., Dugelay, J.-L.: Collusion issue in video watermarking. In: Security, Steganography and Watermarking of Multimedia Contents VII. Volume 5681 of Proceedings of SPIE. (2005) 685–696
8. Doërr, G., Dugelay, J.-L.: Security pitfalls of frame-by-frame approaches to video watermarking. IEEE Transactions on Signal Processing, Supplement on Secure Media **52** (2004) 2955–2964
9. Kirovski, D., Petitcolas, F.: Blind pattern matching attack on watermarking systems. IEEE Transactions on Signal Processing **51** (2003) 1045–1053
10. Cox, I., Kilian, J., Leighton, T., Shamoon, T.: Secure spread spectrum watermarking for multimedia. IEEE Transactions on Image Processing **6** (1997) 1673–1687
11. Barni, M., Bartolini, F., De Rosa, A., Piva, A.: A new decoder for optimum recovery of nonadditive watermarks. IEEE Transactions on Image Processing **10** (2001) 755–766

12. Fisher, Y.: Fractal Image Compression: Theory and Applications. Springer-Verlag (1994)
13. Rey, C., Doërr, G., Dugelay, J.-L., Csurka, G.: Toward generic image dewatermarking? In: Proceedings of the IEEE International Conference on Image Processing. Volume III. (2002) 633–636
14. Petitcolas, F., Kirovski, D.: The blind pattern matching attack on watermarking systems. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. Volume IV. (2002) 3740–3743
15. Kirovski, D., Petitcolas, F.: Replacement attack on arbitrary watermarking systems. In: Proceedings of the ACM Digital Rights Management Workshop. Volume 2696 of Lecture Notes in Computer Science. (2003) 177–189
16. Doërr, G., Dugelay, J.-L., Grangé, L.: Exploiting self-similarities to defeat digital watermarking systems - a case study on still images. In: Proceedings of the ACM Multimedia and Security Workshop. (2004) 133–142
17. Voloshynovskiy, S., Herrigel, A., Baumgärtner, N., Pun, T.: A stochastic approach to content adaptive digital image watermarking. In: Proceedings of the Third International Workshop on Information Hiding. Volume 1768 of Lecture Notes in Computer Science. (1999) 211–236
18. Daugman, J.: Complete discrete 2-D Gabor transforms by neural network for image analysis and compression. IEEE Transactions on Acoustics, Speech and Signal Processing **36** (1988) 1169–1179
19. Dunn, D., Higgins, W., Wakeley, J.: Texture segmentation using 2-D Gabor elementary functions. IEEE Transactions on Pattern Analysis and Machine Intelligence **16** (1994) 130–149
20. Duc, B., Fisher, S., Bigün, J.: Face authentication with Gabor information on deformable graphs. IEEE Transactions on Image Processing **8** (1999) 504–516
21. Donato, G., Bartlett, M., Hager, J., Ekman, P., Sejnowski, T.: Classifying facial actions. IEEE Transactions on Pattern Analysis and Machine Intelligence **21** (1999) 974–989
22. Ringach, D.: Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. Journal of Neurophysiology **88** (2002) 455–463
23. Lim, J.: Two-Dimensional Signal and Image Processing. Prentice Hall International Editions (1989)
24. Foley, J., Legge, G.: Contrast masking in human vision. Journal of the Optical Society of America **70** (1980) 1458–1470
25. Doërr, G., Dugelay, J.-L.: A countermeasure to resist block replacement attacks. In: Accepted for publication in the IEEE International Conference on Image Processing. (2005)
26. Eggers, J., Girod, B.: Informed Watermarking. Kluwer Academic Publishers (2002)
27. Cheng, Q., Huang, T.: Robust optimum detection of transform domain multiplicative watermarks. IEEE Transactions on Signal Processing **51** (2003) 906–924