# A survey of watermarking security

Teddy Furon

# A survey of watermarking security

Teddy Furon *

TEMICS project, INRIA
Campus de Beaulieu, 35042 Rennes, France,
teddy.furon@irisa.fr

**Abstract.** Digital watermarking studies have always been driven by the improvement of robustness. Most of articles of this field deal with this criterion, presenting more and more impressive experimental assessments. Some key events in this quest are the use of spread spectrum, the invention of resynchronization schemes, the discovery of side information channel, and the formulation of the opponent actions as a game.
On the contrary, security received little attention in the watermarking community. This paper presents a comprehensive overview of this recent topic. We list the typical applications which requires a secure watermarking technique. For each context, a threat analysis is purposed. This presentation allows us to illustrate all the certainties the community has on the subject, browsing all key papers. The end of the paper is devoted to what remains not clear, intuitions and future studies.

## 1 Introduction

Watermarking is the art of hiding metadata in content in a robust manner. 'Hiding' has unfortunately many meanings. Some understand that the embedding of metadata doesn't cause any perceptual distortion. Watermarking is then the art of creating a communication channel inside a piece of content without spoiling its entertainment. Others cast a security requirement in the word 'hiding'. This surprisingly happened at the very beginning of the digital watermarking story.

### 1.1 Historical point of view

This very early relationship between security and watermarking might be explained from a historical perspective. In the analog age, content was protected by copyright laws included in intellectual property treaties dating back from the 50's [1]. There was a balance between conflicting issues like the copyright holders interests and the user-friendly usage of content. The digital age and the merging of formats from the entertainment and computer industries broke this balance in the 90's, spoiling copyright holders. Technical barriers have been created to enforce the copyright laws[1]. As cryptography leaves insecure protected content

* This work is partially funded by the national ACI project FABRIANO.

[1] Technical barriers have been existing for a longer time, but the 90's have seen the generalization of their use, especially with DRM (Digital Right Management).

once decrypted by users, a recent technology named digital watermarking was perceived as the last line of defense. It allows to firmly bound content with metadata such as the copyright holder identity (copyright protection [2]) or the copy status (copy protection [3]). At that time, the naïve rationale was: "If you can't see it, and if it is not removed by common processing, then it must be secure".

Unfortunately, digital watermarking was too young a science to support such an adventurous assertion. The technique was even lacking sufficient robustness to fulfill the requirements of these first applications. Defeats happened very soon [4], so that the watermarking community envisaged applications where security is not an issue (e.g. content enhancement). On the front of copyright and copy protection, new laws has been promoted in the 2000's forbidding the circumvention of DRM system [1]. In a way, this new legal framework patches the security flaws of technical barriers, including digital watermarking. There are now three walls of defense: new laws protect the technical barriers which protect the enforcement of old copyright law which protect content's use and exploitation. On the other hand, absolute security does not exist (not even in cryptography) and a high security level has a cost which nobody wants to pay for (copyright holders, device manufacturers, users?). The goal of the entertainment industry is not to erase piracy but to maximize their incomes. To this end, weak security is better than no security [5], and a slightly secure but cheap protection system is enough to "keep honest people honest".

This historical point of view shows that security of digital watermarking has clearly lost interest in real life applications. However, it becomes a hot issue in the watermarking community [6, 7]. We believe that researchers have stretched the limit of robustness to almost its maximum so that new attacks pertain more to security than classical robustness. Because a secure but non robust watermarking technique would be useless, robustness is the weakest link and it was the priority to be fixed. Huge improvements have been done in this field, and security now appears as the next issue on the list. Even if it is less important for real applications, it is also theoretically challenging because very few certainties are known about watermarking security.

### 1.2 Elements to define security

Does a short and concise definition of watermarking security exist? This question stems from two facts: watermarking security has different implications according the targeted application, and security is too close to robustness to be clearly distinguished [8]. Note that, so far, we have discussed about security understanding it as security of robust watermarking. It is time now to broaden our scope.

In copy protection, copyright protection and fingerprinting, we need to assess that dishonest users cannot remove the watermark signal. However, note that in copy protection, a pirate should not be able to change content status to a less restrictive one (e.g., from 'Copy Never' to 'Copy Once')[3]. In fingerprint, a collusion (group of pirates) should not frame an innocent user, i.e. they should not change their hidden message into the identifier of an honest user [9]. In copyright protection, an author should not copy and paste his watermark (possibly issued

by a trusted third party) in content he didn't create [10]. In authentication, the goal of the pirate is not to remove the authenticating watermark signal but to sign content in place of the secret key holder [11]. In steganography[2], the pirate does not remove watermark signal but detects the presence of hidden data, and the watermarking technique used for it [12].

This suggests criteria to make a clear cut between robustness and security:

**Intention.** In security, there obviously exists a pirate. In robustness, a classical content processing made without any malicious intention, might delude the watermark decoder.

**General.** Robustness usually considers classical content processing. In security, pirates apply malicious attacks dedicated to one watermarking technique.

**Removal.** In robustness, the effect of the attack is to mure the watermarking decoder. The attack succeeds in removing enough watermarking energy or it has desynchronized the embedder and the decoder. In security, we have seen that pirates' goals are different according to the targeted application.

**Number of steps.** In robustness, the pirate applies a processing to the watermarked piece of content. This is a single step process. In security, the pirate observes several watermarked pieces of content. He gains from these observations some information about the watermarking technique and the secret key in use. Then, with this 'stolen' knowledge, he attacks protected content. This is a two-step process. Some say the pirate is not fair, in the sense that he is not contented with the official instruction (e.g. the watermarking technique according to the Kerckhoffs' principle), but he tries to access all the information which may be of any help for his goal (e.g. the secret key) [13, Sect. 2].

**Probability of success.** In robustness, an attack is usually not always successful, but it leads to a given Bit Error Rate (decoding) or probability of a miss (detection). In security, a successful hack is almost granted when the pirate has an accurate estimation of the secret key (if this is his goal).

However, T. Kalker formulated very elegant definitions of robustness and security [14, Sect. 2]. These may not encompass all cases, but they are the only concise attempts we are aware of. "*Robust watermarking is a mechanism to create a communication channel that is multiplexed into original content [...]. It is required that, firstly, the perceptual degradation of the marked content [...] is minimal and, secondly, that the capacity of the watermark channel degrades as a smooth function of the degradation of the marked content. [...]. Watermarking security refers to the inability by unauthorized users to have access to the raw watermarking channel. [...] to remove, detect and estimate, write or modify the raw watermarking bits.*"

## 2 Theories of watermarking security

This section deals with the recent attempts to fund a theory of watermark security. What is the role of this theory? Before its existence, the security assessment

[2] We only consider passive steganography in this paper.

of a watermarking technique was fuzzy in the sense that the analysts had to think about an attack and to see how dangerous it was. In other words, the security assessment was clearly dependent on the cleverness of the analysts. Maybe, later on, one will discover a more powerful attack which will lower the security level of the watermarking technique. In a way, the role of a watermarking security theory is to assess the security level once for all.

### 2.1 Steganography

Steganography was the first field in data hiding to benefit from a theory of security. This happens very early in 1998-99 compared to robust watermarking.

The first attempt was made by C. Cachin and it is the most famous theory of steganography security [15]. In steganography, the attacker's goal is to detect a hidden communication in content $Z$ that Alice sends to Bob. This job is a hypothesis test: either the piece of content is a stego-content ($\mathcal{H}_1 : Z = Y$), either it is a natural image ($\mathcal{H}_0 : Z = X$). Performances of the test are measured by the probability of false alarm $P_{fa}$ (i.e. probability of wrongly accusing Alice) and the power of the test $P_p$ (i.e. probability of rightly accusing Alice). An efficient test yields $P_p \sim 1$ for $P_{fa} \sim 0$. However, whatever the structure of the test, its performances are limited by the discrimination between the statistics of $Y$ and of $X$. This is stated by the data processing theorem [16, Th. 4.4.1]:

$$D_{KL}(p_X, p_Y) \geq D_{KL}(P_{fa}, P_p) \geq 0 \tag{1}$$

where $D_{KL}(.,.)$ is the Kullback-Leibler distance (aka discrimination or relative entropy). For instance, if Alice succeeds to produce stego-content $Y$ statistically similar to original content $X$ (i.e. pdf $p_X$ and $p_Y$ are identical almost everywhere but possibly on sets of zero $p_X$-probability), then $D_{KL}(p_X, p_Y) = 0$, which implies $P_{fa} = P_p$. The test is null because it is equivalent of a random decision discarding the observation for flipping a coin to take the decision. If the coin is not biased, this yields $P_{fa} = P_p = 0.5$. C. Cachin argues that if Alice is able to show that $D_{KL}(p_X, p_Y) < \epsilon$, then she limits the performances of the attacker, whatever his test. For $\epsilon = 0$, we have unconditional security. This rationale was applied to build stego-systems with $\epsilon$ as small as possible in papers [17, 18].

The second attempt was made by T. Mittelholzer and it surprisingly remains unknown [19]. The security is related to the amount of information about the stego-message $M$ that leaks from the stego-content $Y$ when ignoring the secret key. This amount is measured by the mutual information $I(M;Y)$. Perfect steganography (or unconditional security) is achieved when $I(M;Y) = 0$. Examples of such schemes are given in [19, Sect. 3].

### 2.2 Robust Watermarking

The theory of robust watermarking security, when successfully applied, gives a lower bound on the secret key estimation accuracy depending on the figure of observed contents. Assume that the pirate knows the watermarking technique

except the secret key (i.e., assuming the Kerckhoff's principle). The core idea is that pirates observing watermarked content can derive some knowledge about the secret key. In other words, information about secret parameters leaks from one watermarked content. This amount of leakage is certainly very small. Yet, if pirates observe many pieces of content watermarked with the same key, each of them leaking some information, then their knowledge increase. It reaches a point where an accurate estimation of the secret key allows powerful attacks.

This magnitude of order defines in a way the security level of the watermarking technique as a number of contents: if one watermarks with the same secret key more pieces of content than allowed by the security level, then a pirate can disclose this later one. As the pirate cannot extract more information about the secret than foreseen by the theory, this guarantees a lower bound.

These recent attempts setting a theoretical framework for watermarking security analysis are indeed the adaptation of the fundamental work of C.E. Shannon, which is considered as the theoretical basis of cryptanalysis [20]. There is nothing new here except the adaptation. Ideas of this adaptation work firstly appeared in [21, Sect. 2.6], [22, Sect. 4.1.1]. The key idea of this theory is to measure this amount of information leakage. Shannon mutual information [23, Sect. 3][24, Sect. 3] or Fisher information matrix [23, Sect. 4] are tools used for this purpose. We will not address the differences between this two tools (see [24, Sect. 2]). There are pros and cons, and even other ways to measure information (Kullback Leibler distance or Renyi information [25]), and also relations between them [26]. What is of utmost importance is that the measurement tool provide a physical interpretation.

The physical interpretation in the Shannon paradigm links the mutual information with the equivocation [20, Sect. 12]. This term is a synonym for the uncertainty or ignorance the pirate has about the secret key. It is measured by the entropy of secret key, regarding it as a random variable:

$$h(K|\mathbf{Y}^{N_o}) = h(K) - I(K;\mathbf{Y}^{N_o}). \tag{2}$$

The watermarker has selected a technique and a secret key, the pirate knows the technique but not the key $K$. Entropy $h(K)$ measures the a priori equivocation, which is the amount of uncertainty before the game starts. The watermarker has produced $N_o$ watermarked pieces of content $\mathbf{Y}^{N_o} = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_{N_o}\}$, each of them leaking some information. Entropy $h(K|\mathbf{Y}^{N_o})$ is the a posteriori equivocation, the amount of uncertainty when the game has started: The total information leakage is $I(K;\mathbf{Y}^{N_o}) \geq 0$. We see that the a posteriori equivocation decreases thanks to the total leakage. A watermarking technique is perfectly secure if $I(K;\mathbf{Y}^{N_o}) = 0$: The pirate will never improve his knowledge about $K$, whatever the figure of produced watermarked pieces of content.

The physical interpretation provided by the Fisher information $\mathrm{FIM}(K, \mathbf{Y}^{N_o})$ is less theoretical. It is given by the Cramer-Rao Bound, which states that whatever the unbiased estimator $\hat{K}$ of the secret key, its accuracy is bounded by[3]:

$$\sigma^2_{\hat{K}} \geq \mathrm{FIM}(K, \mathbf{Y}^{N_o})^{-1}. \tag{3}$$

[3] For simplicity reason, we explain here the CRB in the scalar case.

The smaller the leakage $\mathrm{FIM}(K, \mathbf{Y}^{N_o})$ is, the less accurate is the estimator.

### 2.3 Contextual studies

C. Shannon gives a theory of encryption security where the attacker observes cipher texts. This describes what might happen in military communication. However, in the 70's, cryptography broadens its activities to different fields such as the security of financial transactions. In these new applications, the attacker might not only access cipher texts. At that time, Diffie and Hellman suggest to encompass different contexts in the security level assessment [27, Sect. 2]. These contexts are classified according to the type of observations.

The same terminology was applied in watermarking because this tool is used in many different applications leading to different contexts of attack [22, Sect. 3.3]. All we have to do is to replace $\mathbf{Y}^{N_o}$ by $\mathbf{O}^{N_o}$, where $\mathbf{O}$ is an observation: Watermarked Only Attack ($\mathbf{O} = \mathbf{Y}$), Known Original Attack ($\mathbf{O} = \{\mathbf{Y}, \mathbf{X}\}$), Estimated Original Attack ($\mathbf{O} = \{\mathbf{Y}, \hat{\mathbf{X}}\}$), Known Message Attack ($\mathbf{O} = \{\mathbf{Y}, \mathbf{M}\}$). This list is absolutely not closed nor exhaustive. Some other ideas can be:

- Chosen Watermarked Attack: This is another name for the sensitivity attack. The pirate has a sealed watermark detector. His goal is to disclose the secret key inside by feeding it with chosen contents while noting the output. The observations in this attack are pairs of content and detection output.
- Chosen Original Attack: The pirate has a sealed watermark embedder. His goal is to disclose the secret key inside by feeding it with chosen original contents and saving the watermarked version. The observations here are pairs of original and watermarked content (with the hidden messages).
- Single Original, Multiple Watermarked Attack: This tackles the threat in fingerprinting application: collusion. One original piece of content is watermarked several times with different hidden messages. A collusion of $c$ pirates shares their version noticing differences. Observations are $c$-tuples watermarked contents.
- Multiple Original, Multiple Watermarked Attack: This is the same idea, but this time, there are $N_o$ original pieces of content which have been watermarked and distributed to the clients. Observations is a matrix of $c \times N_o$ watermarked contents. Note that a column of this matrix yields a Known Message Attack, and a line a Single Original, Multiple Watermarked Attack.

The security level assessment of a watermarking technique has to be done context by context, decoupled from application considerations. Then, for a given application, the watermark designer selects the most suitable technique depending on which contexts the threats in this scenario relate.

## 3 Practical tools

These theories do not say anything on the way how to extract and exploit the information leakage. Another crucial point is the complexity of such an algorithm.

This notion was already present in [20, Sect. 21] and denoted as *work*. It might be theoretically possible to extract enough information in order to estimate the secret key, while, in practice, demanding too much computing power. Cryptographers distinguish unconditional security (it is proven that no information leaks from the observations) and computational security (the best known algorithm requires an unreasonable amount of computing power) [28, Sect. 1.13.3].

Whereas theory is just the adaptation of Shannon's security model, practice of watermarking security consists in inventing original and efficient algorithms. Their cores are often known signal processing tools from fields which have a priori nothing in common with watermarking. This extremely interesting part of the job proves once again that watermarking is a multi-field science.

### 3.1 Steganalysis

The theory of C. Cachin basically says that an alarm should be raised when a suspicious content does not comply with the statistical model of original images. This raises at least three issues:

**Feature extraction.** First, images gather in the order of one million of pixels. If the random variable $Y$ is an image, then its definition set $\mathcal{Y}$ is extremely big, and indeed, too huge to lie a statistical model on top. Steganalysers usually extract a feature vector $\tilde{Y}$ from image $Y$. This reduces the definition space: $|\tilde{\mathcal{Y}}| < |\mathcal{Y}|$. Yet, this also reduces the discrimination between stego and original content: $D_{KL}(p_{\mathcal{H}_0}(\tilde{Y}), p_{\mathcal{H}_1}(\tilde{Y})) \leq D_{KL}(p_X, p_Y)$. One must take care of extracting the most discriminative features. This strategy is certainly possible when the stego technique is a priori known. However, it is far more difficult when the steganalyser is universal (to spot stego-contents, whatever the stego-system).

The theory provides some clues. If extracted features are modeled as independent under assumption $\mathcal{H}_0$, then $p_{\mathcal{H}_0}(\tilde{Y}) = \prod_i p_{\mathcal{H}_0}(\tilde{Y}_i)$, and [29, Sect. 2.1]:

$$D_{KL}(p_{\mathcal{H}_1}(\tilde{Y}), p_{\mathcal{H}_0}(\tilde{Y})) = I(\tilde{Y}) + \sum_i D_{KL}(p_{\mathcal{H}_1}(\tilde{Y}_i), p_{\mathcal{H}_0}(\tilde{Y}_i)) \tag{4}$$

where $I(\tilde{Y}) = D_{KL}(p_{\mathcal{H}_1}(\tilde{Y}), \prod_i p_{\mathcal{H}_1}(\tilde{Y}_i))$ measures the dependency between the features, and the second term is the KL distance between the pdf of the marginals. This decomposition seems to justify the use of features extracted from wavelet coefficients such as i) prediction errors to measure the dependency between these coefficients and ii) high order statistics (mean, variance, skewness, kurtosis...) to measure the last summand [30, Sect. 2].

**Classifier.** Once the definition space is set, a second problem is the statistical model. Such a model is required because theorems of hypothesis testing state that optimal tests are based on a sufficient statistic $T = p_{\mathcal{H}_1}(\tilde{Y})/p_{\mathcal{H}_0}(\tilde{Y})$ [31, Chap. 2]. A simple statistical model might be derived for a given stego-system. Yet, this strategy is not possible for universal steganalysers. Usually, a SVM

(Support Vector Machine) is trained on databases of features extracted from stego-contents and original contents [30, Sect. 3]. The SVM learns a way to distinguish the two cases, observing samples of $\hat{Y}$ under both hypothesis $\mathcal{H}_1$ or $\mathcal{H}_0$. In a way, it experimentally learns a statistical model for each hypothesis.

**Conditioning.** A third problem stems from the fact that Cachin theorem need a model for $p_{\mathcal{H}_0}(\tilde{Y})$. Yet, natural images are so diverse that $p_{\mathcal{H}_0}(\tilde{Y})$ is a kind of smooth and large function spreading all over the definition set $\tilde{\mathcal{Y}}$. Assuming a parametric model $p(\tilde{Y}|\theta)$, we have $p_{\mathcal{H}_0}(\tilde{Y}) = \int p(\tilde{Y}|\theta)p(\theta)d\theta$. Some universal steganalysers, for instance, are confused with noisy original images or sharp contour original images as this corresponds to unusual parameter $\theta$.

However, for one particular image, the sufficient statistic of the stego-content is usually higher than the one of the original image. In other words, if the steganalyser could observe the original content and then its stego version, it would notice an increase of the sufficient statistic. Or, within this statistical model, if the steganalyser knows $\theta$, it could be based on more accurate statistics. This gives birth to a very general idea [32, Sect. 3]: the suspicious image $Z$ is slightly transformed to yield an estimation $\hat{\theta}$. Features are extracted and used to build the pdf $p_{\mathcal{H}_0}(\tilde{X}|\hat{\theta})$. A mathematical model foresees $p_{\mathcal{H}_1}(\tilde{Y}|\hat{\theta})$, and thus the sufficient statistic can be calculated $T = p_{\mathcal{H}_1}(\tilde{Z}|\hat{\theta})/p_{\mathcal{H}_0}(\tilde{Z}|\hat{\theta})$. If the suspicious image is a natural image, then $T \sim 1$, otherwise $T > 1$. This leads to better steganalysers even in the universal mode. This is not surprising as conditioning always improves discrimination on average [16, Th. 4.3.6].

### 3.2 Watermarking

Here are some useful signal processing tools to hack watermarking schemes.

**Maximum Likelihood Estimator.** Let us consider the watermarking embedder as a system to be identified. Hence, in the Known Message Attack, we observe pairs of input (i.e. message $\mathbf{m}$) and output (watermarked content $\mathbf{y}$). In other words, we have the framework of a input-output identification. If it possible to write the likelihood $p(\mathbf{y}^{N_o}, \mathbf{m}^{N_o}|K)$, the opponent can use the Maximum Likelihood Estimator (MLE), which finds $\hat{K}$ maximizing the likelihood or nullifying its derivative. The MLE is known to be unbiased and consistent, *i.e.* it asymptotically achieves the CRB derived in 2.2. This has been applied to spread spectrum scheme in [33, Sect. 1.1].

**Expectation Maximization Algorithm.** Unknown messages can be considered as hidden data. The MLE based on $p(\mathbf{y}^{N_o}, \mathbf{m}^{N_o}|K)$ is not practical in this case. But, the EM algorithm approaches it by the iteration of a two-step process:

– Expectation. Having an estimation of the key $\hat{K}(i)$, we estimate the messages $\hat{\mathbf{m}}^{N_o}(i)$. This basically corresponds to the decoding algorithm, known by the pirate according to Kerckhoff's principle.

– Maximization. Having an estimate of the messages $\hat{\mathbf{m}}^{N_o}(i)$, we estimate the secret key $\hat{K}(i+1)$. This step uses for instance the MLE seen above.

This has been applied to spread spectrum schemes in [33, Sect. 1.3].

**Principal Component Analysis.** Many watermark schemes uses projection onto $N_c$ orthonormal private vectors or carriers $\mathbf{u}_\ell$ in order to increase the SNR at the decoding side. In general, one can write: $\mathbf{y} = \mathbf{x}+\mathbf{w}$, with $\mathbf{x}$ the host signal, and $\mathbf{w} = \sum_{i=1}^{N_c} \gamma_i \mathbf{u}_i$. The coefficients $\gamma_i$ carry the message to be hidden and tackle the perceptual constraint. We asssume they are independent from $\mathbf{x}$, i.i.d and centered. It means that the watermark signal $\mathbf{w}$ lives in a small subspace whose dimension is $N_c$, whereas $\mathbf{x}$ belongs to $\mathbb{R}^{N_v}$. This leaves clues for the pirate as the energy of the watermark is focused on a small subspace. For instance, if $\mathbf{x}$ is a white noise, the covariance matrix of $\mathbf{y}$ is $R_y = \sigma_x^2\mathcal{I} + \sum_{i=1}^{N_c} E\{\gamma_i^2\}\mathbf{u}_i\mathbf{u}_i^T$, whereas $R_x = \sigma_x^2\mathcal{I}$. This means that $R_x$ has one eigenvalue $\sigma_x^2$ with order $N_v$, whereas $R_y$ has one eigenvalue $\sigma_x^2$ with order $N_v - N_c$, and $N_c$ other eigenvalues equaling $\sigma_x^2 + E\{\gamma_i^2\}$. Moreover, these $N_c$ biggest eigenvalues are related to eigenvectors $\mathbf{u}_i$. Consequently, it is very easy for the pirate to estimate these private carriers: i) estimate the covariance matrix $R_y$ with $\hat{R}_y = \sum_{i=1}^{N_o} \mathbf{y}(i)\mathbf{y}(i)^T/N_o$, ii) make an eigen-decomposition of this matrix, iii) isolate the eigen-vectors corresponding to the $N_c$ biggest eigenvalues. Figure 1 illustrates this for toy examples. Ellipses show that the watermarked signals are no more white signals.

This Principal Component Analysis has been applied to spread spectrum based schemes in [33, Sect. 1.3] and also in [34, Sect. V.C].

**Independent Component Analysis.** In the case where $E\{\gamma_i^2\} = cst$, then $R_y$ has one eigenvalue $\sigma_x^2 + E\{\gamma_i^2\}$ with associated subspace of dimension $N_c$. When successful, the PCA reveals this subspace and gives a basis, which is not the one used by the embedder: $\{\mathbf{u}_1, \ldots, \mathbf{u}_{N_c}\}$. The pirate can focus his attack noise on this subspace, or remove the watermark signal nullifying the projection of $\mathbf{y}$ onto this subspace. Yet, he cannot have a read and write access on the watermarking channel.

If symbols $\gamma_i$ are statistically independent, an Independent Component Analysis rotates the PCA basis until the estimated symbols 'look like' independent. When successful, the ICA yields estimated carriers which correspond to the real basis up to permutation $\pi(.)$ and change of sign: $\hat{\mathbf{u}}_i = \pm\mathbf{u}_{\pi(i)}$. This ambiguity prevents the pirate to embed/decode messages, but he can check if two watermarked contents have the same hidden message or he can flip bits of hidden messages. This was applied to spread spectrum based schemes in [33, Sect. 1.3].

**Clustering.** The authors of [34] have tested clustering tools to break a video watermarking technique. This technique randomly embeds one of $n$ watermark signals in one video frame. An average attack does not work as it only estimates a mixture of these $n$ signals. However, if a spatial filter succeeds to isolate

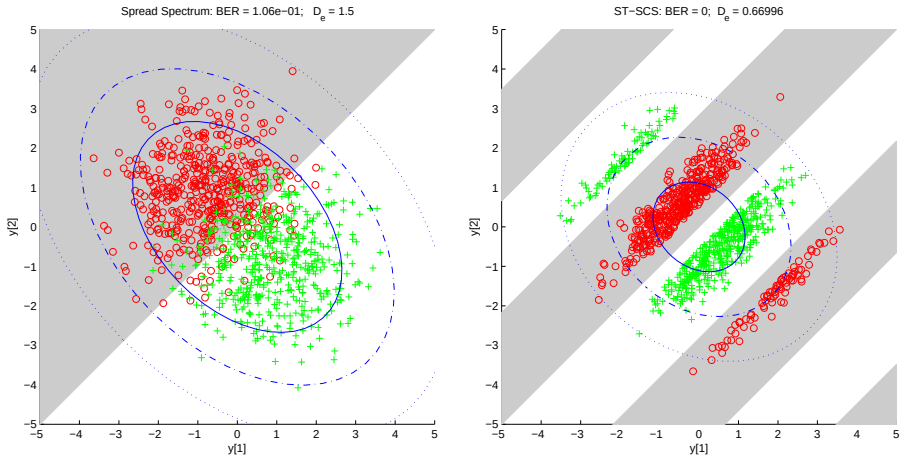


**Fig. 1.** A collection of watermarked signals ($N_v = 2$, $N_c = 1$) with the Spread Spectrum technique (left) and the Spread Transform Scalar Costa Scheme (right). Red circles (green crosses) represent signals hiding symbol '0' (resp., '1'). The grey area (white area) is the decoding region associated to symbol '0' (resp. '1'). Matlab source code available at `www.irisa.fr/temics/Equipe/Furon/iwdw05.m`

enough watermark energy, the pirate obtains noisy estimations of the $n$ watermark signals. His goal is now to split this set of estimations in $n$ clusters of estimations corresponding to one watermark signal, and whose centroids would be good estimates of the $n$ watermark signals. This is a typical task for the $k$-means algorithm (see [34, Sect. IV.D]).

**Vector Quantization.** A closely related tool is the vector quantization, which is used for replacement attack. The pirate as a database of signal blocks and he wishes to replace a block in a watermarked content by a similar block of the database. The word similar is here important. The vector quantization is used to find in the database the most similar 'codeword' (i.e., block), in the sense of the euclidian distance. This tools is used for attacking video watermarking techniques [34, Sect. IV.D] or block based authentication schemes [35].

## 4 Applications

This section gives examples of application where the lack of watermarking security is a threat. We first analyze what the pirate can do with one watermarked content, and then, we see applications where many contents are watermarked.

### 4.1 Watermarking one piece of content

Any application at least discloses one watermarked piece of content.

**Robustness.** The common threat comes from the robustness, except for the scenarios listed below. We especially think here of malicious attack such as the Worst Case Attack or Optimal Attack detailed in recent literature about watermarking robustness [36–38]. Another weakness stems from classical synchronization tricks used in watermarking: templates [39] or geometrically redundant watermarking signals [40, Sect. 6.2] are easy to defeat. The block replacement attack is also a threat: the database is constituted from blocks of non-watermarked contents, or even blocks from the watermarked piece of content [34].

**Deadlock attack and copy-paste attack.** The deadlock attack concerns copyright protection and illustrates the impossibility to prevent somebody to watermark content with his own technique and key (by embedding a watermark signal or by creating a fake original) [41]. This ruins the identification of the owner because two watermarking channels interfere in the same piece of content.

Multiple problems in the field of copyright protection and authentication stems from the copy attack, where the attacker first copies a watermark and then pastes it in a different piece of content [10].

These two last attacks pertain to the protocol layer, in the sense that it questions the link between the presence (or absence) of watermark and the signification at the application layer. We believe that these attacks stem from a misunderstanding of the watermark designers about the targeted application.

In copyright protection, the presence of a watermark has no legal value. The only receivable proof is the belonging of the content to the database of a trusted third party (i.e., an author society). The authors must register their works in this database in order to be protected. It is absolutely useless, from a legal point of view, for the authors to watermark their works on their own. If watermarking is used in copyright protection, it will be embedded by the trusted third party during the registration process.

However, suppose this resort to a trusted third party is not possible (e.g., it is too expensive for the author). At least, the choice of the watermarking technique shall not be given to the author, but somehow imposed by a standard. This standard should select a non-invertible watermarking technique to avoid the deadlock attack. For instance, the secret key should depend on a hash of the original image to prevent the forgery of fake original. Note that this also prevents the copy-paste attack. In the same way, the copy attack is now a nonsense in authentication application. It is true that the very first watermarking authentication schemes were using a constant watermark. But, nowadays, it is well established that the watermark must depend on the original content like a digital signature in cryptography.

### 4.2 Watermarking several pieces of content

Some application discloses several watermarked pieces of content using the same key. This is also the case of video watermarking as the embedder watermarks consecutive blocks of video, whence several pieces.

**Copy protection.** In copy protection, the set of hidden messages is very small (typically 'Copy Never', 'Copy Once', and 'Copy no more'). Moreover, the pirate knows the status of the content. A Known Message Attack is then a real threat.

Another point is that watermark decoders are released in an hostile environment. For instance, they are embedded in consumer electronic devices such as DVD recorders. Pirates can then test watermark decoding as many times as they wish. They do not do this to remove the watermark content by content, but in order to disclose the secret key of the detector [42, 43].

**Authentication.** The assessment of the authentication schemes is sometimes naïve: researchers check that even slight modification of the signed image is indeed detectable. However, in an authentication scenario, it is likely that many images have to be signed. The threat is that a pirate can sign an image without knowing the secret key: he replaces every block of this image by blocks from already signed images. This is indeed a Vector Quantization attack. Counter attacks exist which render the probability of a successful hack extremely small unless the codebooks of replacement blocks are extremely huge [44].

**Fingerprinting.** The typical assessment of fingerprinting schemes is that a collusion of pirates cannot frame an innocent user and that the detector can trace at least one pirate. However, a more complex scenario is the following one: video fingerprinting. In this framework, there are many original contents fingerprinted in different versions, because a watermarking technique embeds hidden messages in a video block by block. For a given user, all these blocks are watermarked with the same secret key and the same hidden message (i.e., the user's codeword). This is a Constant Message Attack (or a Multiple Original, Multiple Watermarked Attack), which is very closed to a Known Message Attack. As watermarking techniques are very weak against it [24, Sect. 4.1],[23, Sect. 4.3], it seems for the moment that secure video fingerprinting is not possible.

## 5 Conclusion: What we do not know yet

**Robust Watermarking.** From a theoretical point of view, the security levels of classical schemes such as spread spectrum and QIM have been well established [45]. However, trellis coding schemes [46] or orthogonal dirty paper codes [47] have not been studied yet. From a practical point of view, algorithms to disclose secret dithering in QIM scheme have not been proposed yet. This might be a complex task especially with a QIM based on VQ. Watermarking

techniques usually used an Error Correction Code, which brings redundancy and thus a security flaw. But, no study has been done on this topic.

**Steganalysis.** Good steganalyzers appeared recently, but they have been tested on simple stego-systems (LSB or +1/-1). No one knows how do they perform against more advanced stego-systems based on QIM.

## References


1. Maillard, T., Furon, T.: Towards digital rights and exemptions management systems. Computer Law and Security Report **20** (2004) 281–287
2. Craver, S., Memon, N., Yeo, B.L., Yeung, M.: Resolving rightful ownership with invisible watermarking techniques: limitations, attacks, and implications. IEEE Journal of selected areas in communications **16** (1998) 573–87 Special issue on copyright and privacy protection.
3. Bloom, J., Cox, I., Kalker, T., Linnartz, J.P., Miller, M., Traw, C.: Copy protection for DVD video. Proc. of the IEEE **87** (1999) 1267–1276 Special issue on identification and protection of multimedia information.
4. Stern, J., Craver, S.: Lessons learned from the SDMI. In Dugelay, J.L., Rose, K., eds.: Proc. of the Fourth Workshop on Multimedia Signal Processing (MMSP), Cannes, France, IEEE (2001) 213–218
5. Cox, I., Miller, M.: Electronic watermarking: The first 50 years. In Dugelay, J.L., Rose, K., eds.: Proc. of Fourth Workshop on Multimedia Signal Processing (MMSP), Cannes, France, IEEE (2001) 225–230
6. Bartolini, F., Barni, M., Furon, T.: Special session on watermarking security. In: Proc. of 11th European Signal Processing Conference (EUSIPCO). Volume 1., Toulouse, France (2002) 283–302, 441–461
7. Barni, M., Pérez-González, F.: Special session on watermarking security. In Delp, E.J., Wong, P.W., eds.: Security, Steganography, and Watermarking of Multimedia Contents VII. Volume 5681 of Proceedings of SPIE-IS&T Electronic Imaging., San Jose, CA, USA, SPIE (2005) 685–768
8. Doërr, G., Dugelay, J.L.: Collusion issue in video watermarking. In Delp, E.J., Wong, P.W., eds.: Security, Steganography, and Watermarking of Multimedia Contents. Volume 5681 of Proceedings of SPIE-IS&T Electronic Imaging., San Jose, CA, USA, SPIE (2005) 685–696
9. Trappe, W., Wu, M., Wang, Z., Liu, K.: Anti-collusion fingerprinting for multimedia. IEEE Trans. on Signal Processing **51** (2003) 1069–1087 Special Issue on Signal Processing for Data Hiding in Digital Media and Secure Content Delivery.
10. Kutter, M., Voloshynovskiy, S., Herrigel, A.: Watermark copy attack. In P.W. Wong, E. Delp, eds.: Security and Watermarking of Multimedia Contents II. Volume 3971., San Jose, Cal., USA, SPIE Proceedings (2000)
11. Wong, P.W., Memon, N.: Secret and public key image watermarking schemes for images authentication and ownership verification. IEEE Trans. on Image Processing **10** (2001) 1593–1601
12. Anderson, R., Petitcolas, F.: On the limits of steganography. IEEE Journal of Selected Areas in Communications **16** (1998) 474–481 Special issue on copyright & privacy protection.
13. Barni, M., Bartolini, F., Furon, T.: A general framework for robust watermarking security. Signal Processing **83** (2003) 2069–2084 Special issue on Security of Data Hiding Technologies, invited paper.

14. Kalker, T.: Considerations on watermarking security. In Dugelay, J.L., Rose, K., eds.: Proc of the Fourth Workshop on Multimedia Signal Processing (MMSP), Cannes, France, IEEE (2001) 201–206
15. Cachin, C.: An information-theoretic model for steganography. In Aucsmith, D., ed.: Proc. of the 2nd Int. Workshop on Inf. Hiding. Volume 1525 of LNCS. Portland, Oregon, U.S.A., Springer Verlag (1998) 306–318
16. Blahut, R.: Principes and practice of information theory. Addisson-Wesley (1987)
17. Guillon, P., Furon, T., Duhamel, P.: Applied public-key steganography. In P.W. Wong, E. Delp, eds.: Security and Watermarking of Multimedia Contents IV, San Jose, Cal., USA, SPIE (2002)
18. Wang, Y., Moulin, P.: Steganalysis of block-structured stegotext. In Delp, E., Wong, P.W., eds.: Security, steganography and watermarking of multimedia contents VI. Volume 5306 of Proceedings of SPIE-IS&T Electronic Imaging., San Jose, CA, USA, SPIE (2004) 477–488
19. Mittelholzer, T.: An information-theoritic approach to steganography and watermarking. In Pfitzmann, A., ed.: Proc. of the 3rd Int. Workshop on Inf. Hiding, Dresden, Germany, Springer Verlag (1999) 1–17
20. Shannon, C.: Communication theory of secrecy systems. Bell system technical journal **28** (1949) 656–715
21. Hernandez, J., Pérez-González, F.: Throwing more light on image watermarks. In: Proc. of 2nd Int. Workshop on Inf. Hiding (IH98). Volume 1525 of LNCS, Portland, OR, USA, Springer-Verlag (1998) 191–207
22. Furon, T.: Security analysis. Techn. report, Certimark European Project (2002)
23. Cayre, F., Fontaine, C., Furon, T.: Watermarking security part I: Theory. In Delp, E.J., Wong, P.W., eds.: Proc. SPIE-IS&T Electronic Imaging, SPIE. Volume 5681., San Jose, CA, USA, Security, Steganography, and Watermarking of Multimedia Contents VII (2005) 746–757
24. Comesana, P., Pérez-Freire, L., Pérez-González, F.: Fundamentals of data hiding security and their application to spread-spectrum analysis. In: Proc. of 7th Inf. Hiding Workshop (IH05). LNCS, Barcelona, Spain, Springer Verlag (2005)
25. Cachin, C.: Entropy Measures and Unconditional Security in Cryptography. Volume 1 of ETH Series in Inf. Security and Cryptography. H.-Gorre Verlag (1997)
26. Cedilnik, A., Kosmelj, K.: Relations among Fisher, Shannon-Wiener and Kullback measures of information for continuous variables. In Mrvar, A., Ferligoj, A., eds.: Developments in Statistics. Volume 17 of Metodoloski zvezki (ISSN 1318-1726)., FDV, University of Ljubljana (2002) 55–62
27. Diffie, W., Hellman, M.: New directions in cryptography. IEEE Trans. on Inf. Theory **22** (1976) 644–54
28. Menezes, A., Van Oorschot, P., Vanstone, S.: Handbook of applied cryptography. Discrete mathematics and its applications. CRC Press (1996)
29. Cardoso, J.F.: Dependence, correlation and gaussianity in independent component analysis. Journal of Machine Learning Research **4** (2003) 1177–1203
30. Lyu, S., Farid, H.: Detecting hidden messages using higher-order statistics and support vector machine. In Petitcolas, F., ed.: Proc. of the 5th Int. Work. on Inf. Hiding. Volume 2578 of LNCS., Noordwijkerhout, The Netherlands, Springer Verlag (2002) 340–354
31. Poor, H.V.: An introduction to signal detection and estimation. Springer (1994 (2nd edition))
32. Fridrich, J., Goljan, M., Hogea, D.: New methodology for breaking steganographic techniques for JPEGs. In Delp, E., Wong, P.W., eds.: Security and watermarking of

multimedia contents V. Volume 5020 of Proc. of SPIE-IS&T Electronic Imaging., Santa Clara, CA, USA, SPIE (2003) 143–155
33. Cayre, F., Fontaine, C., Furon, T.: Watermarking security part II: Practice. In Delp, E.J., Wong, P.W., eds.: Proc. of SPIE-IS&T Electronic Imaging, SPIE. Volume 5681., San Jose, CA, USA, Security, Steganography, and Watermarking of Multimedia Contents VII (2005) 758–768
34. Doërr, G., Dugelay, J.L.: Security pitfalls of frame-by-frame approaches to video watermarking. IEEE Trans. Sig. Proc. **52** (2004) 2955–2964
35. Holliman, M., N. Memon: Counterfeiting attacks on oblivious block-wise independent invisible watermarking schemes. IEEE Trans. Image Proc. **9** (2000) 432–441
36. Vila-Forcen, J., Voloshynovskiy, S., Koval, O., Pérez-González, F., Pun, T.: Worst case additive attack against quantization-based data-hiding methods. In Delp, E., Wong, P.W., eds.: Security, Steganography, and Watermarking of Multimedia Contents VII. Volume 5681 of Proceedings of SPIE-IS&T Electronic Imaging., San Jose, CA, USA, SPIE (2005) 136–146
37. Pateux, S., Le Guelvouit, G.: Practical watermarking scheme based on wide spread spectrum and game theory. Signal Proc.: Image Communication **18** (2003) 283–296
38. Moulin, P., Ivanovic, A.: The zero-rate spread-spectrum watermarking game. IEEE Trans. on Signal Processing **51** (2003) 1098–1117 Special Issue on Signal Processing for Data Hiding in Digital Media and Secure Content Delivery.
39. Herrigel, A., Voloshynovskiy, S., Rystar, Y.: The watermark template attack. In P.W. Wong, E. Delp, eds.: Security and Watermarking of Multimedia Contents, San Jose, Cal., USA, SPIE Proceedings (2001)
40. Topak, E., Voloshynovskiy, S., Koval, O., Mihcak, M., Pun, T.: Security analysis of robust data-hiding with geometrically structured codebooks. In Delp, E., Wong, P.W., eds.: Security, steganogrpahy, and watermarking of multimedia content VII. Volume 5681 of Proceedings of SPIE-IS&T Electronic Imaging., San jose, CA, USA, SPIE (2005) 709–720
41. Craver, S., N. Memon, B.-L. Yeo, M.M. Yeung: On the invertibility of invisible watermarking technique. In: Proc. of Int. Conf. on Image Processing, Washington, DC, USA, IEEE (1997) 540–543
42. Linnartz, J., van Dijk, M.: Analysis of the sensitivity attack against electronic watermarks in images. In Aucsmith, D., ed.: Proc. of the 2nd Int. Workshop on Inf. Hiding. Volume 1525 of LNCS, Portland, Oregon, USA, Springer Verlag (1998)
43. Choubassi, M.E., Moulin, P.: New sensitivity attack. In Delp, E., Wong, P.W., eds.: Security, steganography, and watermarking of multimedia contents VII. Volume 5681 of Proceedings of SPIE-IS&T Electronic imaging., San Jose, CA, USA, SPIE (2005) 734–745
44. Barreto, P.S., Kim, H.Y., Rijmen, V.: Toward a secure public-key blockwise fragile authentication watermarking. IEE Proc. Vision, Image and Signal Processing **149** (2002) 57–62
45. Pérez-Freire, L., Comesana, P., Pérez-González, F.: Information-theoretic analysis of security in side-informed data hiding. In: Proc. of 7th Inf. Hiding Workshop (IH05). LNCS, Barcelona, Spain, Springer Verlag (2005)
46. Miller, M., Doërr, G., Cox, I.: Applying informed coding and informed embedding to design a robust, high capacity watermark. IEEE Tran. Image Processing **13** (2004) 792–807
47. Abrardo, A., Barni, M.: Orthogonal dirty paper coding for informed data hiding. In Delp, E., Wong, P.W., eds.: Security, steganography, and watermarking of multimedia content VI. Volume 5306 of Proceedings of SPIE-IS&T Electronic Imaging., San Jose, CA, USA, SPIE (2004) 274–286