Journal of Machine Learning Research 4 (2003) 1471-1498 Submitted 10/02; Published 12/03

# ICA for Watermarking Digital Images

**Stéphane Bounkong** BOUNKONS@ASTON.AC.UK
**Borémi Toch** TOCHB@ASTON.AC.UK
**David Saad** D.SAAD@ASTON.AC.UK
**David Lowe** D.LOWE@ASTON.AC.UK
*Neural Computing Research Group*
*Aston University*
*Birmingham, B4 7ET, United Kingdom*

**Editors:** Te-Won Lee, Jean-François Cardoso, Erkki Oja and Shun-Ichi Amari

## Abstract

We present a domain-independent ICA-based approach to watermarking. This approach can be used on images, music or video to embed either a robust or fragile watermark. In the case of robust watermarking, the method shows high information rate and robustness against malicious and non-malicious attacks, while keeping a low induced distortion. The fragile watermarking scheme, on the other hand, shows high sensitivity to tampering attempts while keeping the requirement for high information rate and low distortion. The improved performance is achieved by employing a set of statistically independent sources (the independent components) as the feature space and principled statistical decoding methods. The performance of the suggested method is compared to other state of the art approaches. The paper focuses on applying the method to digitized images although the same approach can be used for other media, such as music or video.

**Keywords:** Steganography, Watermarking, Information-Hiding, Authentication, ICA

## 1. Introduction

Modern society increasingly relies on digitized information that can be easily accessed, copied and transmitted. The need for secure authentication and watermarking techniques stimulated research in information hiding over the past three decades. Information hiding, or steganography, has a broad range of applications from copyright protection and transaction tracking, to broadcast monitoring, data integrity, authentication and fingerprinting (Cox et al., 2002). Some applications of steganography, especially in the area of fingerprinting are an essential component in the sale of copyright protected electronic goods, such as software, music and picture files, both off and on-line. Fragile watermarking, on the other hand, plays an important role in authenticating images and audio signals, offering an alternative to traditional methods that often require the transmission of additional metadata; this may be easily identified and removed.

General robust watermarking is based on embedding an imperceptible watermark in the original data (covertext), either for the purpose of identifying the sender/receiver of the data or as a copyright mark. An efficient watermarking technique allows one to embed as much information as possible while minimizing the distortion of the watermarked data with respect to the original and being robust against attacks. A trade off between these conflicting requirements has to be found.

For authentication purposes, in addition to being imperceptible, the watermark has to be sensitive to the slightest modification. This is termed fragile watermarking and allows the detection of tampering attempts. In some cases, the watermark is required to be sensitive only to some attacks while not being affected by others, such as common processing techniques (semi-fragile watermarking).

Current state of the art watermarking methods mostly operate in a feature space, such as the Fourier domain, rather than on the raw data. The allowed level of distortion due to watermarking, based either on conventional measures or on a perceptual model, is limited to a predefined threshold. The embedding process typically relies on modulation or quantization methods, whereas decoding often relies on correlation detection or mapping to the nearest discrete value.

Most modern watermarking techniques have emerged in the last decade. Researchers in this field come from different backgrounds and typically bring with them the knowledge from their previous field; this is reflected in the watermarking techniques devised so far, both in the methods suggested for the watermark embedding process and the feature space chosen for this purpose. The plethora of watermarking methods available and the narrow suitability of each to specific domains make it difficult to provide a principled comprehensive theoretical approach to watermarking; such an approach is a prerequisite to any optimization scheme aimed at maximizing the information embedding rate and the robustness against various attacks, and minimizing the information degradation.

In this paper, we propose a novel approach to watermarking which is independent of the application domain and is supported by existing results from information theory of watermarking. It is based on embedding the message in a set of *statistically independent sources* obtained in our case by independent component analysis (ICA) (Hyvärinen et al., 2001). These sources constitute the spanning of a feature space, and represent the covertext in conjunction with the corresponding set of constant mixing matrices. The distortion measure and the ICA mixing and demixing matrices may differ greatly from one application domain to another, but the watermarking scheme principle remains the same. Indeed, the demixing process gives a set of independent sources, which share similar characteristics and have little correlation with the original application domain. Different generative models may be used for identifying the set of independent sources; ICA, that we will focus on here, is clearly one of the most principled methods to identify statistically independent sources used as the feature space in the suggested scheme.

Recent information theoretical analyses have provided clear upper bounds to the watermark information capacity for a given distortion based on communication channel with side information. Concrete bounds have also been derived in special cases by Cohen and Lapidoth (2002) and Moulin and O'Sullivan (2003). Unfortunately, the analyses do not provide any information on how to design an optimal system and have been carried out under certain assumptions about both source and watermark distributions. However, an important result of this research is that the upper bound, in the case of parallel channels, is reachable only if the sources are statistically independent. Any selection of statistically dependent sources results in theoretically inferior performance; this motivates the use of ICA as our feature space.

In addition, we replace suboptimal threshold-based decoding methods by maximum a posteriori (MAP) decoding using noise and source models derived from experiments.

The resulting domain independent watermarking method was examined against existing techniques and was found to be competitive with, and frequently superior to state of the art approaches.

The paper is organized as follows: in Section 2, we introduce the general watermarking framework and existing techniques. In Section 3 and 4, we present our ICA based approach and details of the method used, respectively. Comparative experiments with other watermarking techniques carried out and their results are analyzed in Section 5. Discussion and conclusions are presented in Section 6.

## 2. Watermarking: General Framework

The problem of robust watermarking can be described as a game between Alice, Bob and Mallory. Alice is the legitimate owner of a digital data **X**. She wants to embed in it some information **m** to protect her intellectual property rights (IPR) and to be able to prove her ownership (or alternatively to embed fingerprinting information about the data transaction, such as details of the buyer). Mallory is the attacker or forger. He wants to counterfeit the watermarked data $\hat{\mathbf{X}}$ by uprooting the embedded information and/or embedding his own personal data. The induced modification is denoted by $n$. Bob is the receiver of the digital data $\tilde{\mathbf{X}}$, he wants to be sure of buying the data from an authorized seller (or legitimate owner, Alice). He therefore investigates the presence of a potentially hidden information $\hat{m}$ from the received data $\tilde{\mathbf{X}}$ (alternatively, Alice may want to find the source of an illegal distribution). This general watermarking problem is also described in Figure 1.

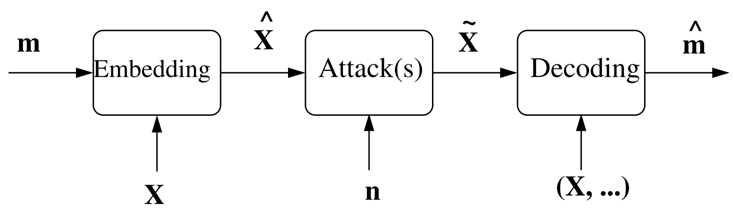


Figure 1: A general watermarking scheme where **m** is the embedded message, **X** the covertext, $\hat{\mathbf{X}}$ the watermarked covertext, $\tilde{\mathbf{X}}$ the attacked covertext and $\hat{m}$ an estimate of the original message **m**.

Fragile watermarking follows roughly along the same lines, except for the fact that Alice embeds a watermark that is highly sensitive to tampering attempts. Mallory would like to modify the data to his advantage (e.g., remove information from CCTV footage), while Bob wants to be sure that the received data $\tilde{\mathbf{X}}$ has not been tampered with.

Clearly, watermarking has more than one objective. It can be used to prove ownership or to fingerprint data (robust watermarking); for this purpose, the embedded information has to be non-removable unless the original data (covertext) is irreversibly damaged and carries no value. Robustness against various attacks is therefore an important requirement for robust watermarking; the watermark acts as a serial number carved on real products. Watermarking can also be used for authentication and verification of data integrity. The watermark proves that the data has not been modified. Fragility is the important feature to focus on. The scheme has to enable the detection of slight changes against which robust watermarking is expected to survive. Authentication should identify both global and local changes, such as the replacement of a character in an image, and eventually locate them. As watermarked images may be compressed or transmitted over a noisy channel, also fragile watermarking may be required to survive some non-malicious attacks, while allowing the detection of deliberate attacks such as feature replacements. This case is usually defined as semi-fragile watermarking.

### 2.1 Watermarking Requirements

Generic watermarking can be seen as a constrained communication problem, of channel with side information, where distortions induced by the sender and the attacker are limited. The three main constraints are imperceptibility, robustness and information rate. A formal model presenting the relation between these three constraints is described by Moulin and O'Sullivan (2003) using an information theoretical framework of information hiding. In this paper, we give an intuitive description of these opposing requirements:

**Imperceptibility** is defined as the similarity between the original covertext and the watermarked version. This can be expressed as $d_1(\mathbf{X}, \hat{\mathbf{X}}) \leq \delta_1$, with $d_1$ being a distortion measure and $\delta_1$ the desired threshold. Typical distortion measures found in the literature are mean square error, signal-to-noise ratio, etc (Cox et al., 2002). More complicated distortion functions, relying on domain specific perceptual models can also be used. Note that the attacker is also limited by a similar distortion constraint $d_2(\hat{\mathbf{X}}, \tilde{\mathbf{X}}) \leq \delta_2$, where the distortion measure $d_2$ and its threshold $\delta_2$ may be different from $d_1$ and $\delta_1$, although the same distortion measure is typically used. In the latter case, the distortion caused by an attack is usually expected to be higher than that of the watermark itself ($\delta_1 \leq \delta_2$).

**Robustness** A watermarking scheme should survive various distortions or attacks. Common attacks may be different depending on the nature of the covertext, for example audio or images. Malicious attacks aim at removing the watermark, while non-malicious attacks are due to common transmission and signal processing practices that might harm the embedded signal. Robust schemes should at least survive non-malicious attacks as well as some malicious attacks. Generally, these attacks are global to the whole covertext. Robustness is often achieved by embedding globally a watermark with little information content compared to the covertext. Robust watermarking aims at maximizing the rate of correctly identifying the original message, $p(\hat{\mathbf{m}} = \mathbf{m}|\tilde{\mathbf{X}})$. The message estimate $\hat{m}$ is obtained by some decoding method.

**Fragility** In fragile watermarking, one replaces the requirement for robustness by fragility, which is the ability to detect an alteration of the watermark due to tampering attempts. Furthermore, it is often needed to localize the distortion. Such a requirement makes it necessary to have a watermark with higher information content than in the robust watermarking framework spread over the entire covertext. Our method has different performance with respect to different attacks. Among them, we focus on fragility with respect to replacement attacks, while maintaining some robustness against non-malicious attacks. Fragile watermarking aims at optimizing the two following constraints:

$$\begin{cases} \max p(\hat{\mathbf{m}} = \mathbf{m}|\tilde{\mathbf{X}}_r)\,, \\ \min p(\hat{\mathbf{m}} = \mathbf{m}|\, \|\tilde{\mathbf{X}}_f - \hat{\mathbf{X}}\| > \delta_f)\,, \end{cases}$$

where $\tilde{\mathbf{X}}_r$ is the watermarked data attacked by a non-malicious attack against which the scheme has to be robust, $\tilde{\mathbf{X}}_f$ is the watermarked data attacked by another type of attack typically localized which needs to be detected by the scheme. The parameter $\delta_f$ sets the fragility of the scheme.

**Information rate** The information capacity of a watermarking system is the supremum of all achievable information hiding rates for a given distortion constraint and a given set of attacks.

A more formal definition, relying on information theory is given by Moulin and O'Sullivan (2003). We are interested in maximizing the information embedding rate for a given distortion and attack constraints, which is clearly upper bounded by the capacity.

There is no general solution to this problem, as it depends on the final purpose of the system. For instance, if the watermark has to survive attacks that induce high distortion, error-correcting codes can be used to improve robustness, but it may be at the expenses of the information rate. From now on, we will focus on digital image watermarking although most of the topics may also be relevant for other domains, such as music or video.

### 2.2 Watermarking Feature Spaces

Early watermarking schemes operated directly in the covertext domain. However, such methods have been shown to be quite poor in their robustness properties. Many feature spaces have been studied in recent years for improving the efficiency of watermarking systems; most of these choices are for reasons of convenience and traditional use in the specific domain, such as Fourier and Cosine transforms in images. A theoretical approach to find an optimal feature space in a particular case, based on an information theoretical approach, has recently been suggested by Ramkumar (2000). However, it is yet to be seen if this approach can be of practical use. Research has focused more on the design of practical systems using well known feature spaces. The three main spaces are:

**Cosine Transform Domain:** is widely used being a good approximation for the Karhunen Loève transform for highly correlated data, such as images. The cosine transform has good variance compaction property (Jain, 1989). Therefore, using this feature space facilitates spreading the watermark across a large part of the image information content. Schemes based on this transformation (Cox et al., 1997) show good robustness against various non-malicious attacks. Window cosine transform (on $8\times8$ pixel windows) is also used (Koch and Zhao, 1995) in conjunction with JPEG quantization table to be robust against the well-known JPEG compression standard (Wallace, 1992), arguably the most common non-malicious attack.

**Fourier Transform Domain:** Watermarking schemes based on this feature space are also usually robust against non-malicious attacks for similar reasons. Moreover, some variants of this transformation, such as Fourier-Mellin transform, inherently include properties such as invariance against affine transformations, thus allowing some watermarking schemes (Ruanaidh and Pun, 1997) to handle geometric attacks, such as rotations. Notice that geometric attacks are so far among the most efficient attacks against general watermarking schemes (Petitcolas and Kuhn, 2002, Stirmark).

**Wavelet Transform Domain:** Motivated by the upcoming JPEG2000 standard (Committee, 2000), this feature space may play an important role in the near future similar to that of the cosine transform space at present. Watermarking in the wavelet domain has recently been the focus of many research projects in this area (Meerwald, 2001). Moreover, this frequency feature space allows the embedding process to be spatially localized.

### 2.3 Embedding, Attacks and Decoding

Various methods have been used to efficiently embed/retrieve information in a medium subject to different attacks. Among them, two main classes have emerged: quantization and modulation methods. Quantization methods have been widely studied in the coding area for decades, and have recently been used in the context of watermarking. Modulation methods are often paired with a correlation detector and a given decision threshold; they show better performance when the original data is available for decoding. The schemes have been studied against common attacks in the literature, among them: noise, lossy compression, band filtering, cropping and collusion.

### 2.3.1 EMBEDDING

Embedding a watermark usually follows one of the two following schemes or can be a variant of one of them.

*Modulation:* Embedding information through modulation is usually carried out using one of the three formulae suggested by Cox et al. (1997):

$$\hat{\mathbf{X}} = \mathbf{X} + \alpha\mathbf{m} \quad , \; \hat{\mathbf{X}} = \mathbf{X}(1 + \alpha\mathbf{m}) \quad , \; \hat{\mathbf{X}} = \mathbf{X}e^{\alpha\mathbf{m}} \; ,$$

where $\mathbf{m}$ is the embedded message, $\mathbf{X}$ the original covertext value, $\alpha$ is a pre-defined strength factor, and $\hat{\mathbf{X}}$ is the watermarked value. In this approach, the value of $\alpha$, common to all components is chosen heuristically.

*Quantization:* Another widely used method to embed information is to quantize the original data $\mathbf{X}$, using a quantization function $q$ that provides different quantization values/grids to different embedded message values $\mathbf{m}$. The embedding strength is determined by the minimal distance $\delta$ between two adjacent quantization values corresponding to two different $m$ symbols.

Other methods, such as modifying a couple of feature space coefficients, while preserving their absolute difference or some other predefined criteria, may be used to embed data. However, in most cases, these are merely variants of modulation and/or quantization.

### 2.3.2 ATTACKS

A watermarked image may undergo some attacks. We distinguish two kinds of attack: non-malicious attacks, which are common signal processing methods, that are not aimed at removing or tampering with the embedded watermark; and malicious attacks, which are deliberate attempts to remove/disable the watermark, possibly using the embedding algorithm itself. It is obvious that some non-malicious attacks can also be used as malicious attacks, especially if the watermarking process is known to be weak against them.

Common non-malicious attacks include:

**Noise** Data may be altered due to transmission through a noisy communication channel.

**Lossy compression** Compression algorithms are often used to transfer image data efficiently. Compression algorithms such as JPEG or JPEG2000 allow excellent compression rates, while they introduce moderate levels of distortion, which depend on the chosen quality level or compression rate. Such attacks are completely deterministic but also difficult to model.

**Enhancement** Very common image processing techniques fall into this category, such as luminosity adjustment, sharpening/blurring, contrast adjustment, edge enhancement.

**De-synchronization** Other common processing techniques do not remove the watermark or affect the quality of the picture, but may disable the watermark detection. These include rescaling, cropping and rotation.

#### 2.3.3 DECODING

Different decoding methods are used for the various embedding techniques. For instance, correlation detection is used when the watermark has been embedded using a modulation scheme. The correlation is computed between the attacked covertext $\tilde{\mathbf{X}}$ (or the difference between the attacked and original data $\tilde{\mathbf{X}} - \mathbf{X}$) and the watermark $\mathbf{m}$. The correlation value is then compared with a predefined detection threshold. A watermark is detected if the correlation is above it. Notice that the original data is required for the decoder to perform efficiently.

Quantization decoding is usually carried out by mapping the attacked value to the nearest quantized value. Knowledge of the quantization process is sufficient for decoding.

A principled decoding method, examined later in this paper, is maximum a posteriori (MAP) decoding. Using probabilistic models of the data, watermark, embedding, noise and corruption process in conjunction with Bayesian statistics, one may obtain posterior mean values of the message as well as error-bars. Drawbacks of this method are: its sensitivity to the accuracy of the probabilistic models used, decoding is carried out in the feature space and at a high computational cost.

## 3. ICA for Watermarking

ICA was introduced several years ago as a blind source separation technique, but since then has been used in a broad range of applications, from sparse coding and denoising to feature extraction (Hyvärinen et al., 2001). The main assumption in ICA is that a given signal can be represented as a linear mixture of statistically independent sources. This property combined with the simplicity of a linear mixture model have made ICA a powerful and useful tool in various research fields.

In the context of watermarking, an ICA based technique has been studied by González-Serrano et al. (2001). The latter, unlike our approach, is related to a least significant bit modification in the ICA domain. Also, the reported results show quite poor performance. In our approach, ICA allows the maximization of the information content and minimization of the induced distortion by decomposing the covertext (in this case the image) into statistically independent sources. Embedding information in one of these independent sources minimizes the emerging cross-channel interference. In fact, for a broad class of attacks and fixed capacity values, one can show that distortion is minimized when the message is embedded in statistically independent sources (Appendix A). Information theoretical analysis also shows that the information hiding capacity of statistically independent sources is maximal (Moulin and O'Sullivan, 2003). Finally, this extremely simple transformation facilitates the use of Bayesian decoding techniques based on statistical models. They can be constructed due to the simple factorized statistics of the sources. Principled Bayesian techniques are expected to improve the decoding performance in real systems.

Another significant advantage of the ICA based approach is its independence with respect to the application domain. The distortion measure and the ICA mixing and demixing matrices may differ from one application domain to another, but the watermarking scheme principle remains the same.

### 3.1 ICA-based Watermarking as a Communication Problem

ICA-based watermarking can be described, from an information theoretical perspective, as a communication channel with side information (Cox et al., 2002). We use the information theoretical description but focus on the application of ICA-based watermarking within it. Exploiting the fact that ICA is a simple linear transformation, we concentrate on the feature space in modeling the sources, attacks and induced distortion.

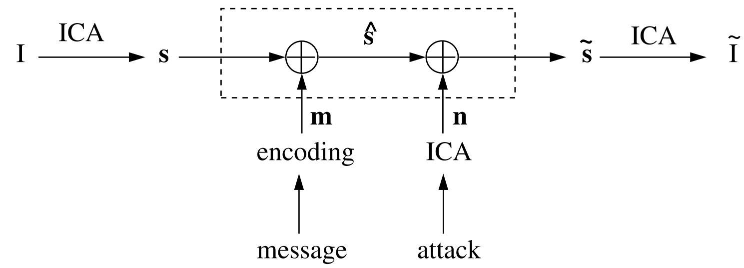


Figure 2: ICA watermarking as a communication problem. In this figure, $I$ is the original image, $\mathbf{s}$ is the demixed signals, $\mathbf{m}$ is the message to embed, $\hat{\mathbf{s}}$ is the watermarked signals, $\mathbf{n}$ is the corruption noise, $\tilde{\mathbf{s}}$ is the corrupted or attacked watermarked signals and $\tilde{I}$ is the attacked watermarked image.

Figure 2 describes the complete watermarking problem, separating the linear mixing/demixing operations from the communication channel itself (within the dashed line). In order to construct a proper statistical model of the process, we need to model the demixed signals $\mathbf{s}$, which are the source realizations, any possible noise $\mathbf{n}$, the attack process $p(\tilde{\mathbf{s}}|\hat{\mathbf{s}},\mathbf{n})$, the message $\mathbf{m}$ and the embedding process $p(\hat{\mathbf{s}}|\mathbf{s},\mathbf{m})$. For convenience, the distribution of $\mathbf{m}$ is set to be uniform on $\{0,1\}$. We generate a statistical model of $\mathbf{s}$ based on real data, in this case, a set of representative images.

The embedding process we use is based on Quantization Index Modulation (QIM) studied by Chen and Wornell (2001). The reasons for selecting this particular technique are:

- its reported high performance (it has been shown to approach optimal performance for some models),
- its inherent non-linearity, which makes it more secure (Craver, 1996),
- the simple statistical model it offers, and
- the simplicity of its implementation.

Finally, we choose to define the attack process as an additive process and fit the noise distribution according to this assumption.

### 3.2 ICA for Images

ICA is a versatile technique used in various applications, including image processing (Hyvärinen et al., 2001). The ICA process derives features that best represent the data via a set of components that are as statistically independent as possible. The main assumption behind ICA is that any typical given signal $\mathbf{X}$ can be represented as a linear combination of statistically independent sources $\mathbf{s}$ using a mixing matrix $A$ such that $\mathbf{X} = A\mathbf{s}$; we also have the inverse (demixing) relation $\mathbf{s} = W\mathbf{X}$, where $W$ denotes the corresponding demixing matrix. Various methods allow ICA basis vector estimation.

In experiments, we used the FastICA algorithm developed by Hyvärinen and Oja (1997), which provides good decomposition results efficiently.

Since it is often impractical to use full size images (bigger than 32×32 pixels) as inputs, we apply the FastICA algorithm to square image patches. Various patch sizes are used in the literature, from 8×8 to 32×32 pixels. Two practical aspects have to be considered: processing time and the size of relevant features. Large patches are theoretically feasible, but their basis estimation is computationally demanding; on the other hand, a small patch size leads to poor performance in the watermarking process. A trade off between these two conflicting constraints has to be found.

Based on this consideration and practical experiments, we constructed our basis (Figure 5) from a training set of 11 natural scene images (Figure 4), from which a set of 11,000 image patches of 16×16 pixels has been randomly sampled. The data obtained have then been centered. To remove noise and improve energy compaction, the data dimensionality was reduced using principal component analysis (PCA). The remaining 60 largest eigenvalues (Figure 3) preserved 98.68% of the data variance. The preprocessed dataset was used as input to the FastICA algorithm.

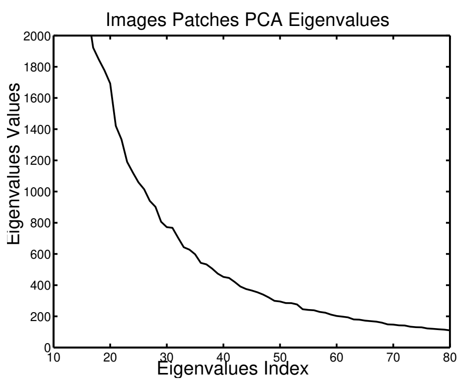


Figure 3: Image patch eigenvalues.


Figure 4: Example of a natural image.

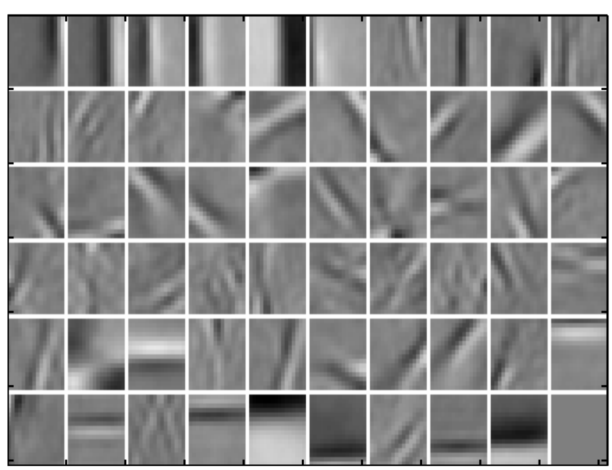

Figure 5: ICA basis obtained from natural images.

## 4. ICA Watermarking Process

In this section, ICA based algorithms for robust and fragile watermarking are proposed. Both algorithms rely on an ICA feature space and QIM. Some differences, which we will highlight later on, make them more suitable to their respective purposes.

The general ICA based watermarking process comprises four stages as described in Figure 6.

1. The image is divided into contiguous image patches giving a set of mixed signals. Each patch is then demixed resulting in **s**, using a predetermined ICA demixing matrix $W$, prepared using an ensemble of typical images.

2. For each patch, a set of coefficients are selected according to the specific task (fragile or robust watermarking) as described below.

3. The selected coefficients are quantized and watermarked. The difference between the watermarked and original values is denoted by $\Delta$.

4. $\Delta$ is multiplied by the mixing matrix $A$ to produce $w$ which is then added to the original picture $I$.

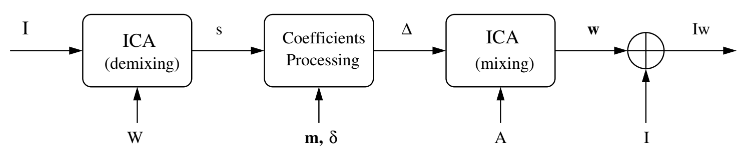


Figure 6: ICA watermarking scheme.

*Robust Watermarking Scheme ICs set selection:* In order to achieve the best robustness, all allowed distortion is concentrated in one IC, using a large quantization step, which has been selected according to some criteria; for instance, robustness against certain types of attacks or suitable statistical properties. Furthermore, to improve imperceptibility we use slight modifications in non-selected ICs to compensate for the distortion induced by the quantization of the selected IC; this is carried out by minimizing $\|A\hat{\mathbf{s}}\|$, where $\hat{s}_i$ is the selected and quantized value and $i$ the index of the quantized source or IC; $A$ is the mixing matrix.

*Fragile Watermarking Scheme ICs set selection:* For authentication purposes, one is required to detect changes in any given patch; therefore, a large set of ICs are selected in each patch to be quantized. The probability for a random *replacement* patch to have the same binary watermark is therefore $p = 2^{-n}$, where $n$ is the number of quantized ICs, chosen to be sufficiently high so as to limit the feasibility of data counterfeiting. In the case of a replacement larger than the patch size, the probability of have the same binary signature can be approximated by $p^m$, where $m$ is here the number of involved patches (Figure 7).

The allowed distortion characterized by $\delta_1$ is distributed across this set. The lower $\delta_1$ is, the more fragile and imperceptible the watermark becomes. However, if $\delta_1$ is too small, the finite resolution of the digital image makes it impossible to embed any watermark.

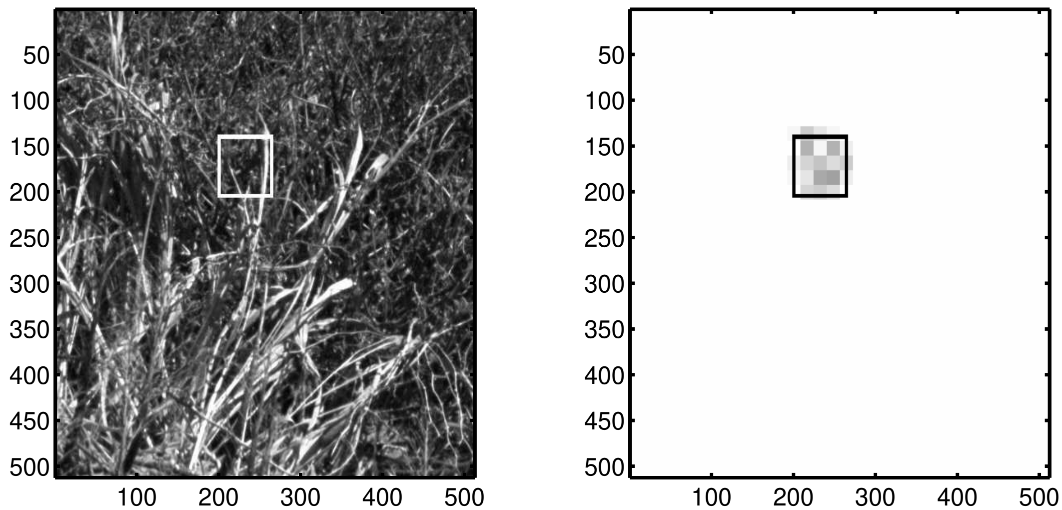


Figure 7: Detection of a tampering attack using ICA-based fragile watermarking. In the left image, a square patch of $64 \times 64$ pixels has been attacked by a Gaussian noise of standard deviation 0.5. The modified area is delimited by a white line. In the right image, the grey regions show the image patches, where a potential tampering attempt has been detected. A black line denotes the area where the modification has been carried out.

Such a watermark is easily destroyed as intended, but one may also want to ensure that the watermark survives mild non-malicious attacks that may occur in the recording or transmission process.

### 4.1 Watermark Embedding: Quantization Index Modulation

QIM (Chen and Wornell, 2001) is the embedding method used in our experiments. It can be seen as a quantization process, which uses two grids corresponding to the value of the message bit $m_i$ $\{0,1\}$ (Figure 8). As underlined, the relative simplicity of this embedding process facilitates the use of a statistical model to be used in the decoding process.

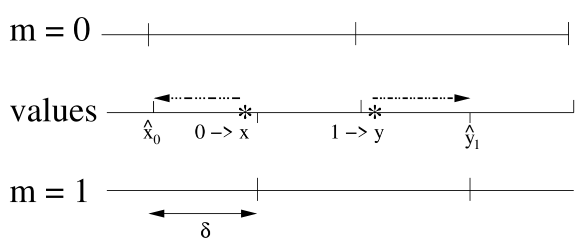


Figure 8: Quantization index modulation (QIM) - the original value is represented by *, with $\delta$ representing the quantization step; the dashed arrows show the quantization index modulation. On the left, we embed the message bit $m = 0$ in the real value $x$ while on the right we embed the message bit $m = 1$ in the original value $y$.

### 4.2 Watermark Decoding

After an attack, the task is to infer the embedded message vector $\mathbf{m}$ from the attacked (corrupted) value of the watermarked covertext $\tilde{\mathbf{s}}$. We will focus on two methods, a simple threshold based approach and on a principled Bayesian decoding method.

*Decoding to the nearest grid point* is probably the simplest decoding for a quantization embedding scheme, described in Figure 9. The robustness of the coding/decoding process is directly linked to the quantization step $\delta$ used. The only requirement of this decoder is knowledge of the quantization grids.

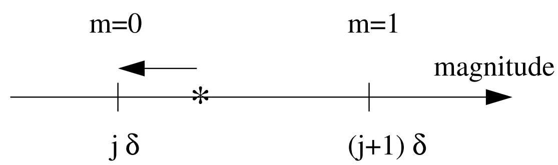


Figure 9: Decoding to the nearest grid point; in this case the retrieved message $m$ is 0.

*MAP decoding* relies on selecting $\mathbf{m}$ values that maximize the posterior probability $p(\mathbf{m}|\tilde{\mathbf{s}})$ as an estimate of the message, $\hat{\mathbf{m}} = \max_{\mathbf{m}} p(\mathbf{m}|\tilde{\mathbf{s}})$. Using Bayes rule, we obtain

$$p(\mathbf{m}|\tilde{\mathbf{s}}) = \frac{p(\tilde{\mathbf{s}}|\mathbf{m})\ p(\mathbf{m})}{p(\tilde{\mathbf{s}})} \ .$$

Exploiting the fact that all components of the source $\mathbf{s}$ and message $\mathbf{m}$ are identically independently distributed and assuming that correlations emerging from the attack process are negligible, one can reduce the multidimensional problem to a factorized single variable inference problem. As $p(\mathbf{m})$ has been chosen uniformly and given the fact that $p(\tilde{\mathbf{s}})$ is a normalization term independent of $\mathbf{m}$, one may reduce the inference problem to $\hat{m}_i \propto \max_{m_i} p(\tilde{s}_i|m_i)$ over the two $m_i$ values, where $i$ is the index in the vector. From now on, the latter index $i$ will be omitted to simplify the notation.

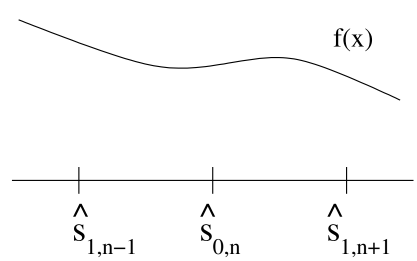


Figure 10: Probability of having $\hat{s}_{0,n}$ given $m = 0$.

The probability of having a given quantized value, say $\hat{s}_{m=0,n}$, which embeds the message $m = 0$ is central to our calculation. The expression provided in Equation 1 is based on the fact that any value between two consecutive grid points $\hat{s}_{m=1,n}$, watermarked by a message $m = 0$ will be quantized to the same value $\hat{s}_{m=0,n}$ as explained graphically in Figure 10, where the first index represents the embedded bit value and the second represents the running quantization number on the combined grid. One straightforwardly derives the expression

$$P(\hat{s}_{0,n}|m=0) = \int_{\hat{s}_{1,n-1}}^{\hat{s}_{1,n+1}} f(x)\,\mathrm{d}x\ , \tag{1}$$

that relies on the source probability density function $f$ which is not known analytically.

From the Markov chain in Figure 2, and using the probabilistic attack model we constructed previously, we derive the following conditional probabilities that facilitate the MAP estimate of $\mathbf{m}$:

$$\begin{aligned}
p(\tilde{s}|m=0) &= \sum_n p_n(\tilde{s}-\hat{s}_{0,n}) \int_{\hat{s}_{1,n-1}}^{\hat{s}_{1,n+1}} f(x)\,\mathrm{d}x\ , \\
&= \sum_{\hat{s}_0} p_n(\tilde{s}-\hat{s}_0) P(\hat{s}_0|m=0)\ , \\
p(\tilde{s}|m=1) &= \sum_n p_n(\tilde{s}-\hat{s}_{1,n}) \int_{\hat{s}_{0,n-1}}^{\hat{s}_{0,n+1}} f(x)\,\mathrm{d}x\ , \\
&= \sum_{\hat{s}_1} p_n(\tilde{s}-\hat{s}_1) P(\hat{s}_1|m=1)\ ,
\end{aligned}$$

where $p_n(\tilde{s}-\hat{s}_{\cdot,n})$ represents a noise model, the argument of which is the difference between the watermarked and attacked IC values.

Clearly the method relies heavily on obtaining reliable probabilistic models for both sources and the attack process. In this study, we constructed a statistical model for the watermarking problem based on an ICA feature space of digital images. A model for $f$, based on the family of Generalized Gaussian Exponential (GGE) has been derived and statistically tested in Appendix B. In order to obtain a better model, more elaborate distributions, such as mixtures of Gaussian, may be used. The disparity between image and noise sources, $\mathbf{s}$ and $\mathbf{n}$, and the constructed probabilistic models $p(s)$ and $p(n)$, respectively, is measured by the $\chi^2$ test (Appendix C and D). Three MAP based decoders are devised for three different attacks: JPEG compression, set partitioning in hierarchical trees (SPIHT) compression and Gaussian noise.

## 5. Experiments and Results

To test the performance of our watermarking scheme against existing state of the art methods, we carried out a set of experiments for each of the watermarking tasks.

### 5.1 Experiments

To ensure the imperceptibility of the watermark for all studied methods, we set the distortion constraint threshold $\delta_1$ related to the embedding process to 43dB using the peak to signal noise ratio (PSNR) measure. All the tests were carried out on a test set (different from the ICA training set) of 11 greyscale natural scene images (Figure 4) of $512 \times 512$ pixels. The embedded messages $\mathbf{m}$ were randomly generated binary sequences on $\{0,1\}$. We tested multiple strengths for each given attack.

*Robust Watermarking Scheme* - To study the performance of our robust watermarking scheme, we used three variants of our method, two DCT based watermarking algorithms and a DWT based algorithm; their performance was tested under Gaussian noise, JPEG compression (Wallace, 1992) and SPIHT compression (Said and Pearlman, 1996) attacks.

For a given image, attack type and attack strength, the test was repeated 100 times with a different embedded message **m** of length 1024. In the next subsection, we give a brief description of the algorithms. The complete settings of experiments can be found in Appendix E.

*Fragile Watermarking* - We tested our fragile watermarking scheme's ability to detect modified patches. In order to simulate a tampering process, a randomly located square patch of $16\times16$ pixels in the watermarked picture was modified by the addition of random noise. Since the size of our image patches is $16\times16$, this attack can affect up to four image patches of our scheme. Each one of the four potentially affected patches has only 128 pixels involved on average. This means that 75% of the patch is unaffected by the tampering. The detection was considered successful if some of the fragile watermark bits could not be correctly retrieved from the corresponding region. For a given image, attack strength and length of message, the test was repeated 100 times with a different message. The embedded messages **m** were of length $1024 \times n$, where $n$ is the number of bits embedded per patch and 1024 is the number of patches per image. In this test, we set $\delta_n$ at 0.1, which gave us a different distortion for each $n$ between 49 and 50 dB PSNR.

We also tested the robustness of our scheme against non-malicious mild attacks such as Gaussian noise or JPEG compression with respect to the number of bits embedded per patch. The trade off between robustness and fragility had to be set according to the final purpose of the watermarking application. In our scheme, the watermark fragility is increased by increasing the number of ICs to modify per patch; doing so also reduces the probability for a random patch to carry the same binary signature. It is also possible to increase the fragility by decreasing the quantization step $\delta_n$, or by considering several adjacent patches together if the relevant feature size in the picture is large. Physical limits of the digital image storage (quantization) set a lower bound to $\delta_n$. For our experiments, we set the watermarking distortion threshold to 43dB PSNR and distributed all this distortion allowance across the set of selected ICs. Details of parameters used can be found in Appendix E. For a given image, attack strength and length of message, the test was repeated 10 times with different messages. The length of the embedded messages **m** was $1024 \times n$, where $n$ is the number of bits embedded per patch.

### 5.2 Algorithm Descriptions

The various algorithms described below are based on quantization of a selected set of coefficients in their respective feature space. The preselection of these sets is also described below.

**ICA Sel** This is an ICA based algorithm, where we preselect a small subset of ICs that are particularly robust against a specific attack; a single IC is then randomly selected from this subset to be watermarked in each patch. Decoding is carried out by mapping to the nearest grid point.

**ICA Ne** This ICA based algorithm is introduced as a benchmark for the ICA Map algorithm, to show the improvement gained from using a principled decoding method instead of decoding to the nearest grid point. A single IC is selected to be watermarked in all patches for a given attack; the selection criterion is not directly related to its robustness against specific attacks, but having a good agreement with the corresponding source and attack models, according to the $\chi^2$ test. Decoding to the nearest grid point is used.

**ICA Map** This algorithm is similar to ICA Ne, except for the use of MAP decoding instead of decoding to the nearest grid point.

**DCT** A standard, commonly used, DCT based algorithm. It quantizes (QIM) the DCT representation of the entire picture. Among the DCT coefficients which represent a signal with at least one cycle per image patch of $16\times16$, the 1024 lowest frequency ones convey the watermark. The watermarked picture is then obtained by application of an inverse DCT. Decoding to the nearest grid point is used.

**DCTX** A local DCT based algorithm. It relies on a partitioning of the picture into contiguous patches of $16\times16$ pixels. The DCT is applied to each of them. For each patch, a single coefficient is randomly selected among the low frequency ones, and quantized (QIM). An inverse DCT is then applied to obtain each watermarked patch. Decoding to the nearest grid point is used.

**DWT** A multiresolution wavelet transform based watermarking algorithm. The detail of the process can be found in (Kundur and Hatzinakos, 1998), but it basically relies on embedding the message in the third level of the Haar wavelet decomposition with some strength parameter ($Q = 2$ in our case) that determines both the robustness achieved and the level of imperceptibility. A correlation based decoder is used. Full details of the method and its parameters are given in (Kundur and Hatzinakos, 1998).

**ICA Fra** A fragile watermarking scheme using ICA feature space and QIM. The main difference with respect to robust ICA-based watermarking method, is that here one embeds more information per patch by choosing several ICs in each patch but with a lower quantization step for each IC. The number of embedded bits and the quantization steps can be tuned to adapt to fragility/robustness requirements. Here, decoding to the nearest grid point is also used.

### 5.3 Results

*Robust Watermarking Results:* The Gaussian noise attack results in Figure 11 show that ICA Sel, ICA Ne and DCTX perform as well as the global DCT method, while DWT is less robust across the entire range of noise level. ICA Map on the other hand, although based on an IC with sub-optimal robustness properties, outperforms all other schemes due to its good source and noise statistical models that have been exploited in the decoding process.

The JPEG compression attack results in Figure 12 show that ICA Sel performs equally well as the DCTX algorithm. On the other hand, DCT and DWT have quite poor performances in general, while ICA Ne and ICA Map perform as well as DCTX and ICA Sel in the entire range of acceptable compression rates (30-90), but show bad results for high compression. This is due to a breakdown of the statistical models for such high compression rates, indicated by the poor $\chi^2$ test results, and the sub-optimal robustness properties of the selected IC.

The SPIHT compression attack results in Figure 13 show that ICA Sel, ICA Ne, DCT and DCTX have similar performances with slightly better performance showed by ICA Sel and DCTX. DWT performs quite poorly. ICA Map algorithm does not perform very well, presumably due to inaccurate source or/and noise models used in the decoding process.

The results show quite promising results for ICA algorithms in general, which are either competitive with, or outperform other state of the art methods. Improving the source and noise models may facilitate further improvements in performance.

*Fragile Watermarking Results:* An example of detection of attacked patches was given in Figure 7. In this example, a patch was attacked by Gaussian noise above the value that the watermark

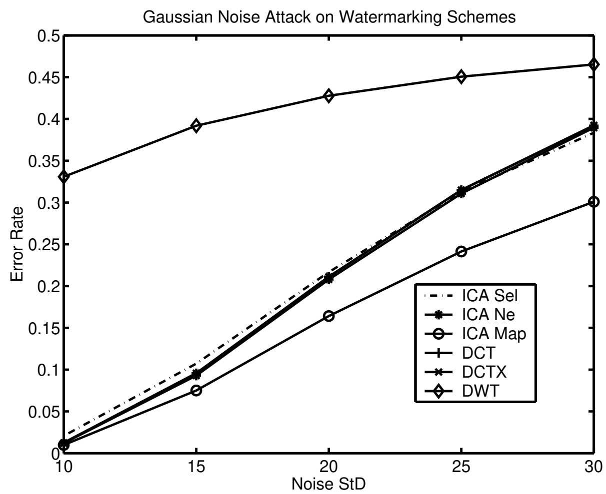


Figure 11: Performance of watermarking schemes against Gaussian noise

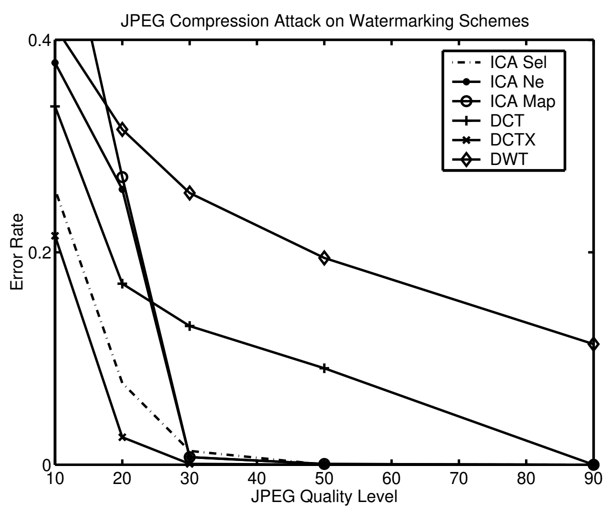


Figure 12: Performance of watermarking schemes against JPEG compression

was designed to tolerate. The probability of identifying and locating the attacked patch then follows directly from the number of bits embedded in each patch.

To study the fragility and robustness properties of our fragile watermarking scheme, ICA Fra, we conducted a set of experiments to determine the probability of identifying an attacked patch, and the percentage of decoding errors observed under mild Gaussian and JPEG attacks. The results of these experiments are shown in Figure 14 and Figure 15 respectively.

In the first set of experiments, we attacked an arbitrary single patch ($16 \times 16$ pixels) in each image using Gaussian noise, and monitored the probability of identifying the attacked patch. The experiment was carried out in low Gaussian noise values of variance smaller than 1 (keeping in mind

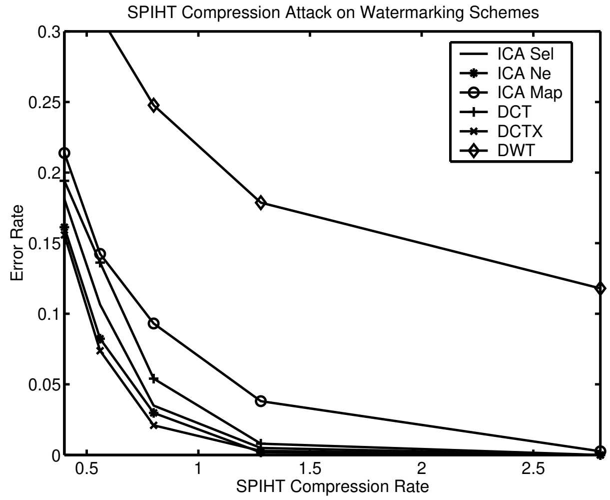


Figure 13: Performance of watermarking schemes against SPIHT compression

that intensity levels are in the range 0-255) and different numbers of marked bits per patch. The probability of detection is marked on the curved lines with respect to the number of bits embedded per patch (horizontal axis) and the noise level (vertical axis). We see that even in relatively low noise levels, it is still possible to identify the attacked patches.

In the second set of experiments we studied the robustness of our method against *global* non-malicious attacks, Gaussian noise and JPEG, as shown in Figure 15. The figures describe the fraction of decoding errors (marked on the curved lines) as a function of the number of bits marked per patch (horizontal axis) and the noise level (vertical axis). Clearly, a higher number of bits marked per patch for a given distortion, will result in smaller quantization steps and lower robustness. The results show that ICA Fra provides an efficient fragile watermarking method even in the presence of mild non-malicious attacks. The embedded fragile watermark, of a given low distortion (43dB PSNR), can easily and reliably be identified using 10 bits per patch (with probability for a random patch to carry the same signature of 0.1%).

Another aspect of our method that one should emphasize is that $A$ and $W$ are unknown to the forger; this, combined with small typical quantization steps will make it very difficult to forge a watermark.

## 6. Conclusion

We have presented a novel approach to both robust and fragile watermarking, using ICA as the feature space in which watermarks are embedded. The new approach, based on embedding information in statistically independent sources, shows high information embedding rate and minimal distortion. We have examined its performance on a set of representative images, random messages and various attacks, and our experiments show promising performance on all the attacks examined.

The main advantage of our approach is that, being based on embedding information using statistically independent sources, the same watermarking method can be easily applied across different media. Based on local information and a linear transform, our method is computationally efficient,

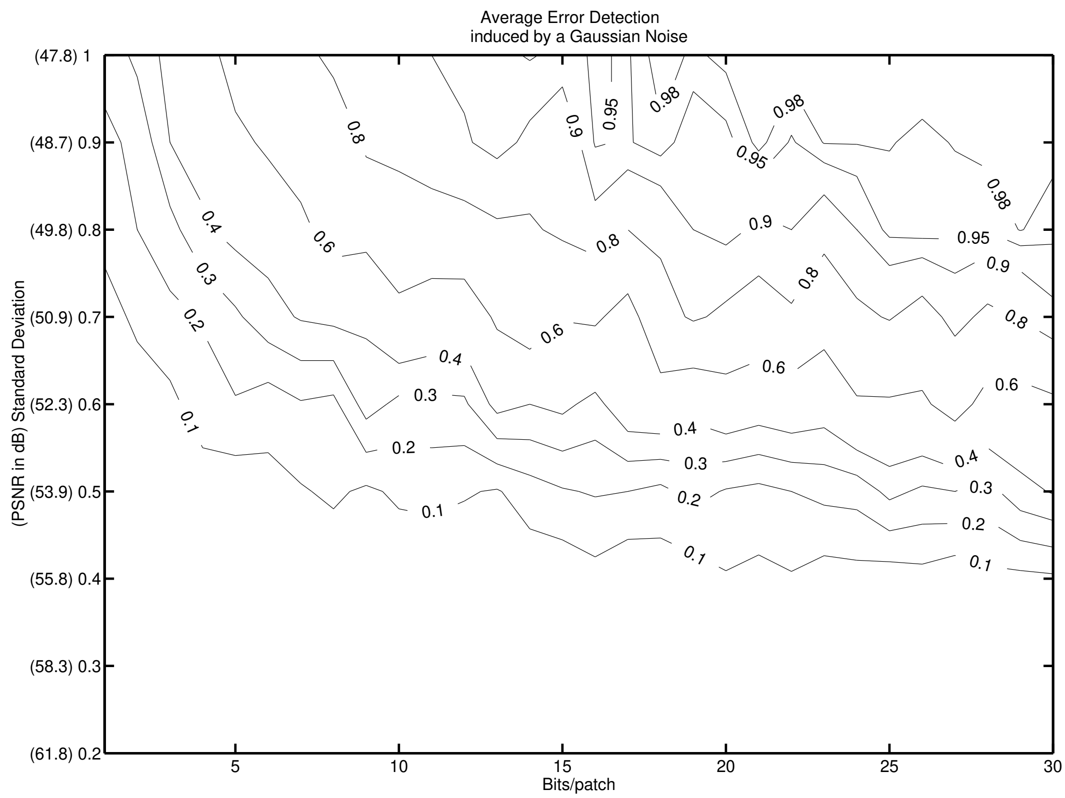


Figure 14: ICA Fra watermarking for local distortion detection.

and offers additional security in the use of *specific* mixing/demixing matrices that are not easy to obtain. The provided statistical models for both sources and attacks facilitates the use of a Bayesian decoding method that has the potential to provide an optimal decoding scheme.

Further research may improve the performance by refining the existing statistical models or identifying a new approach for selecting the IC's to be watermarked or the distortion measures used (for instance a measure based on the human visual system). Applying the same approach to other domains, such as audio signals (Toch et al., 2003), may also require some adaptations due to the different nature of the signals.

## Appendix A. Proof

In this appendix, we will show that a blockwise memoryless watermarking process, for a given power-limited class of blockwise memoryless attacks and a fixed capacity, minimizes the distortion it induces when the sources $\mathbf{s}$ to watermark are independent. This result will be derived from (Moulin and O'Sullivan, 2003), proposition 8.3.

It is assumed that $\mathbf{s}$ is a blockwise memoryless source with block size $L$, that the attack is also blockwise memoryless with blocks of same size $L$ and that the class of attacks $\mathcal{A}$ is limited by a distortion constraint. Let $\overline{p}(\mathbf{s}) = \prod_{i=1}^{L} p(s_i)$ be the product of the marginals $p(s_i)$. Theorem 4.4 in (Moulin and O'Sullivan, 2003) gives us the following expression for the capacity $C$ of the

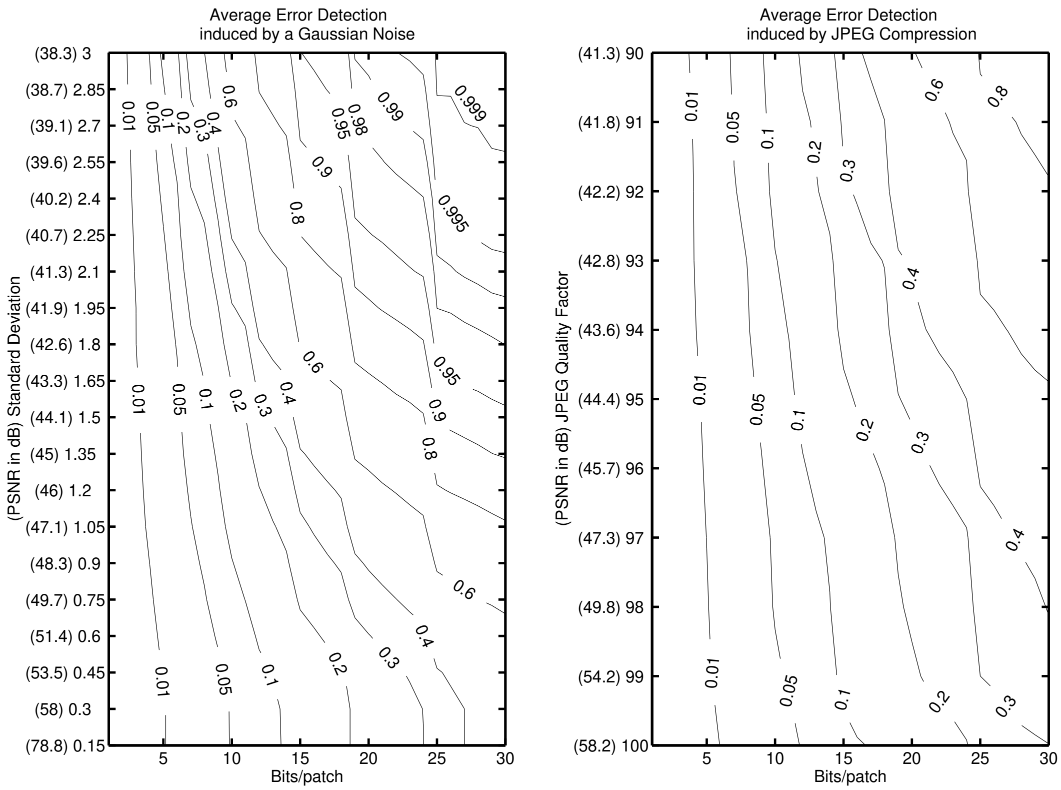


Figure 15: ICA Fra watermarking for global distortion detection.

watermarking game against an attacker subject to a distortion constrain for embedding, assuming the distribution $p(\mathbf{s})$:

$$C = \max_{Q \in \mathcal{Q}} \min_{A \in \mathcal{A}} J(Q, A) \ ,$$

where $Q$ denotes the probability density function of the embedding channel, $A$ the probability distribution function of the attack channel and $J$ represents the cost function (information rate), described in section 4.3 of (Moulin and O'Sullivan, 2003).

Let us first fix the class of attacks $\mathcal{A}$. From proposition 8.3 in (Moulin and O'Sullivan, 2003) and using the same notations, we derive that, subject to a given maximum embedding distortion $D_1$, the capacity $C$ of any distribution $p(\mathbf{s})$ and the capacity $\overline{C}$ of $\overline{p}(\mathbf{s})$ are related by $C \leq \overline{C}$.

For a given distribution $p(\mathbf{s})$, if $D$ and $D'$ are two maximum embedding distortions and $C$ and $C'$ are the respective capacities, then $D \leq D'$ is equivalent to having $C \leq C'$. It can be easily proved, since the capacity is defined as the maximum over a set of probability density functions $\mathcal{Q}$. Increasing the distortion, increases the set $\mathcal{Q}$ thus resulting in a higher capacity.

If there exists a maximum embedding distortion $D_2$ such that the capacity $\overline{C_2}$ of $\overline{p}(\mathbf{s})$ subject to $D_2$, admits $\overline{C_2} = C$, then, according to the result from the previous paragraph applied to the distribution $\overline{p}(\mathbf{s})$, we obtain $D_2 \leq D_1$.

We have proved that, for a given capacity, the lowest maximum distortion induced by the embedding process is achieved when the distribution of the block elements is independent (factorized).

## Appendix B. Generalized Gaussian or Exponential Distributions for $\nu = 1/k$ and $\nu = 2/k$

The general expression of densities belonging to this family is (for zero mean)

$$p_x(x) = C \exp\Big(-\frac{|x|^\nu}{\nu E\{|x|^\nu\}}\Big) .$$

The positive real-valued power $\nu$ determines the type of distribution, and $C$ is a normalizing constant. Two families of probability density functions and probability functions, for $\nu = 1/k$ and $\nu = 2/k$ are given below.

The $GGE(1/k)$ family probability density function and probability function are given below, where $\sigma$ represents the distribution's standard deviation.

$$\begin{aligned}
f_{\frac{1}{k}}(x) &= \frac{1}{2k!\sqrt{t^k}\sigma}\exp\Big\{-\frac{|x|^{\frac{1}{k}}}{\sqrt{t}\sigma^{\frac{1}{k}}}\Big\} , \\
F_{\frac{1}{k}}(x) &= \frac{1}{2}+\frac{1}{2}sign(x)\Bigg\{1-\Big(\sum_{i=1}^{k}\frac{|x|^{\frac{k-i}{k}}}{(k-i)!t^{\frac{k-i}{2}}\sigma^{\frac{k-i}{k}}}\Big)\Bigg\} , \\
t &= \Big(\frac{(k-1)!}{(3k-1)!}\Big)^{\frac{1}{k}} .
\end{aligned}$$

The $GGE(2/(2m+1))$ family expressions are given by the following equations, where $\sigma$ is the standard deviation.

$$\begin{aligned}
f_{\frac{2}{2m+1}}(x) &= \frac{1}{\sqrt{2\pi}\prod_{i=i}^{m}(2i+1)\Sigma^{2m+1}}\exp\Big\{-\frac{|x|^{\frac{2}{2m+1}}}{2\Sigma^2}\Big\} , \\
F_{\frac{2}{2m+1}}(x) &= \frac{1}{2}+sign(x)\Bigg\{\frac{1}{2}+\frac{1}{\sqrt{2\pi}\Sigma^{2m+1}}\Big(\sum_{i=1}^{m}\frac{2^{m-i}(m-i)!}{(2(m-i)+1)!}|x|^{\frac{2(m-i)+1}{2m+1}}\Sigma^{2i}\Big)\cdots \\
&\quad \exp\Big\{-\frac{|x|^{\frac{2}{2m+1}}}{2\Sigma^2}\Big\}-\Phi(-\frac{|x|^{\frac{1}{2m+1}}}{\Sigma})\Bigg\} , \\
\Sigma^2 &= 2\Big(\frac{(2m)!(3m)!}{2m!(6m+1)!}\Big)^{\frac{1}{2m+1}}\sigma^{\frac{2}{2m+1}} , \\
\Phi(u) &= \frac{2}{\sqrt{\pi}}\int_0^u \exp(-t^2)\mathrm{d}t .
\end{aligned}$$

## Appendix C. The $\chi^2$ Fitting Test

The $\chi^2$ test is a method for testing the relevance of a model against real data. It uses a data distribution model, also called hypothesis, and a set of real data samples. The disparity between model and data is measured by a normalized quadratic difference, Equation 2. Then comparing the $\chi^2$ value obtained to a given threshold, the hypothesis is rejected or accepted. The $\chi^2$ expression is given by the following equation.

$$\chi^2 = \sum_{i=1}^{k}\frac{(M_i-up_i)^2}{up_i} = \sum_{i=1}^{k}M_i^2 up_i - u , \tag{2}$$

where $p_i = F(b_i) - F(a_i)$ is the theoretical probabilities of $x$ falling in $\Delta_i = [a_i, b_i)$, $M_i$ being the number of sample values in $\Delta_i$, with $\sum_{i=1}^{k} M_i = u$. The border bins must satisfy $np_i \geq 1$ and the others $up_i \geq 5$; $m = k - r - 1$ is the degree of freedom of $\chi^2$ where $r$ is the number of parameters.

In our study, we used 26 bins of size 0.3 from $-3.9$ to 3.9, two borders bins are also added from $-\infty$ to -3.9 and from 3.9 to $\infty$, so $m = k - r - 1 = 26 - 1 - 1 = 24$, where $r$ is the number of estimated parameters, the number of samples per signal is $u = 11000$, then the critical value $\chi^2_\alpha$ is 36.4 ($\log_{10}(\chi^2_\alpha) = 1.56$), for a confidence value $\alpha = 0.05$, see Bronshtein and Semendyayev (1997) for further details.

## Appendix D. Images and Attacks Modeling

### D.1 Image Models

Experiments using the $\chi^2$ test and aiming at modeling the image sources with the two GGE families and randomly sampled squared patches show that about 13% of the 60 ICs can be modeled by a GGE distribution with $\nu = 2/3$ or $\nu = 1/2$, as seen in Figure 16. In the case of MAP decoding, these ICs are therefore preferred, as explained in the text. Further research on more complex models may overcome the limitation represented by the restricted choice of ICs to watermark.

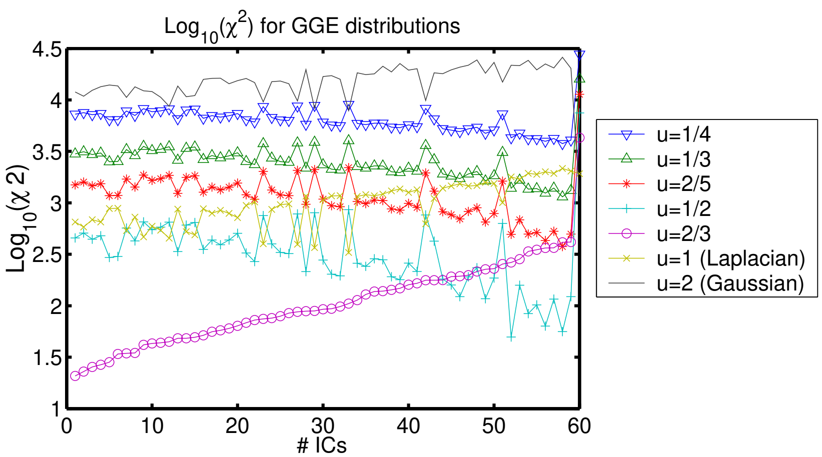


Figure 16: $\chi^2$ fitting test - different GGE ($\nu \in \{\frac{1}{4}, \frac{1}{3}, \frac{2}{5}, \frac{1}{2}, \frac{2}{3}, 1, 2\}$) distributions are tested against some real data. The latter are obtained from a set of 11,000 randomly sampled images patches of 16×16 pixels demixed by a ICA demixing matrix $W$ of 60 ICs.

### D.2 Attacks Models

**JPEG Compression** As shown in Figure 17, JPEG compression with high level quality (low compression), such as JPEG 90 are quite well described by a Gaussian distribution. When the quality level decreases, the Laplacian distribution model becomes more suitable. However, as for the source models, only a few ICs have their model validated and none have their source and noise models validated at the same time. In order to improve the decoding performance, further research is required to refine the models.

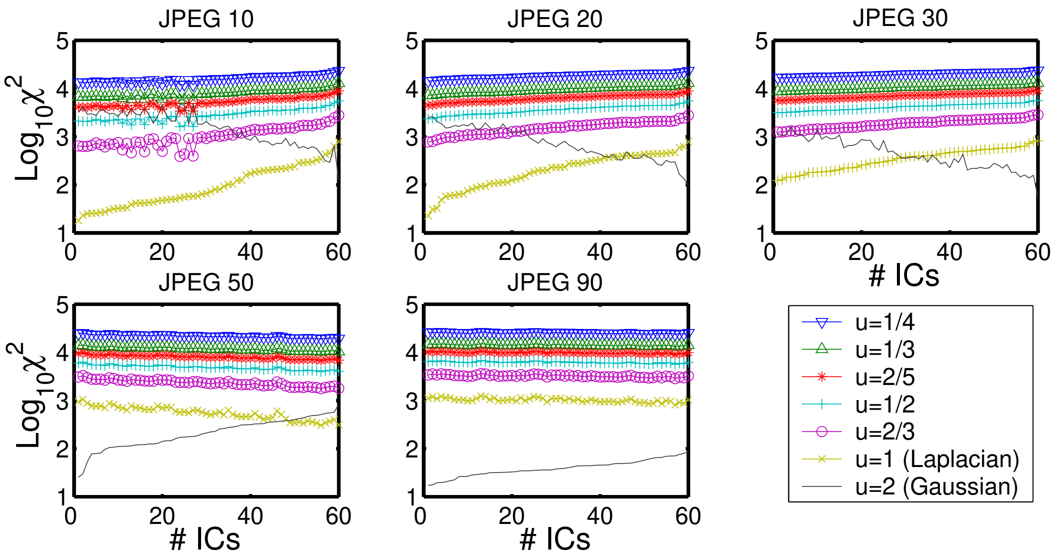


Figure 17: $\chi^2$ fitting test for JPEG compression models.

**SPIHT Compression** As shown in Figure 18, SPIHT compression is best modeled by the Laplacian distribution. However, as for JPEG compression, the $\chi^2$ values also show the need for refinements. As previously, we will use the closest distribution, for instance the Laplacian distribution as a first approximation for our experiments.

**Gaussian Noise** ICA is a linear transform and a linear combination of centered i.i.d. Gaussian variables of the same variance remains Gaussian. Therefore Gaussian noise on images remains Gaussian in the ICA feature space with its standard deviation being a function of the original standard deviation and the demixing matrix $W$. In our study, if $\sigma$ is the standard deviation of the noise attack, and $\sigma_i$ is the standard deviation of the attack on a given IC, the latter can be expressed as $\sigma_i = \|(W^T)_i\|\sigma$, where $\|.\|$ is the 2-norm and $(W^T)_i$ is the $i^{th}$ column index of the matrix $W^T$.

## Appendix E. Settings

In this section, we specify the different experimental settings for each algorithm and each attack; $\delta$ denotes the quantization step of the QIM process.

*ICA Sel/DCT/DCTX/DWT:* For these for algorithms, the same settings are used for all attacks and attack strengths. A set of ICs are first selected, then a proper quantization step is set, such that the distortion constraint requirement is obeyed.

*ICA Ne/Map:* Table 2 summarizes the noise and source models chosen for ICA Map algorithm. The ICs have been selected such that the $\chi^2$ value is minimal for both. This means that for any other IC, either the $\chi^2$ of the noise or source model is higher than both $\chi^2$ of the selected IC. Unfortunately, as shown in the Table 2, all the presented models, but in the Gaussian noise attack case, are rejected.

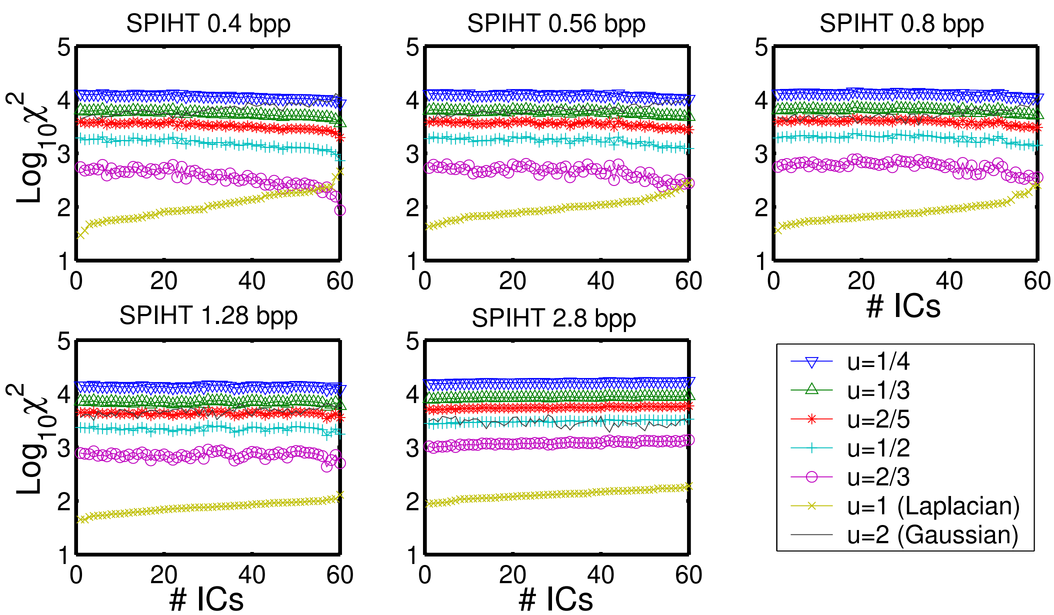


Figure 18: $\chi^2$ Fitting Test for SPIHT Compression Models.

| | ICA Sel | DCT | DCTX | DWT |
|---|---|---|---|---|
| Selected ICs/Coef | $\{5, 55, 59\}$ | low frequency with at least one cycle per $16\times16$ pixels patch | $\{3, 4, 18, 19, 20, 33, 34, 35, 49, 50\}$, see Figure 19 | The coefficients are randomly selected among the 3rd level decomposition |
| $\delta$ (or Q for DWT) | 0.7 | 50 | 50 | 2 |

Table 1: ICA Sel/DCT/DCTX/DWT algorithm experimental settings

This might also explain the lack of improvement shown by the ICA Map algorithm in simulations for this particular attack.

*ICA Fra:* A set of ICs is randomly drawn from a pre-selected set and quantized using a quantization step $\delta_n$, where $n$ is the number of quantized ICs. The pre-selected ICs are the ones within the range 11-40. The different quantization steps are given in Table 3, for a fixed induced distortion of 43dB PSNR.

## Appendix F. Numerical Results

*Robust watermarking results:*

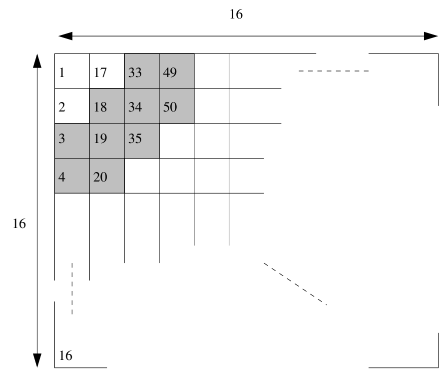


Figure 19: DCTX algorithm coefficient selection.

| | GN | JPEG | | | | | SPIHT | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Strength | 10-30 | 10 | 20 | 30 | 50 | 90 | 0.4 | 0.56 | 0.80 | 1.28 | 2.8 |
| Selected ICs | 19 | 16 | 34 | 43 | 28 | 15 | 20 | 23 | 17 | 23 | 20 |
| $\delta$ | 1.2 | 2.3 | 3 | 1.2 | 2.2 | 1.9 | 2 | 2 | 2 | 2 | 2 |
| Source Model | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| $\chi^2$ | 34.0 | 25.4 | 68.3 | 92.5 | 49.4 | 22.9 | 34.5 | 43.0 | 26.7 | 43.0 | 34.5 |
| Noise Model | 7 | 6 | 6 | 7 | 7 | 7 | 6 | 6 | 6 | 6 | 6 |
| $\chi^2$ | - | 50.6 | 57.1 | 55.3 | 78.5 | 35.5 | 50.7 | 49.3 | 36.6 | 45.1 | 98.8 |

Table 2: ICA Ne/Map algorithm experimental settings, where GN stands for Gaussian Noise, and the noise models correspondences are: 5 for GGE(2/3), 6 for GGE(1) or Laplacian and 7 for GGE(2) or Gaussian.

## References

I.N. Bronshtein and K.A. Semendyayev. *Handbook of Mathematics*. Springer-Verlag, 1997.

B. Chen and G.W. Wornell. Quantization index modulation: a class of provably good methods for digital watermarking and information embedding. *IEEE Transactions on Information Theory*, 47 (4):1423–1443, May 2001.

A.S. Cohen and A. Lapidoth. The gaussian watermarking game. *IEEE Transactions on Information Theory*, 48(6):1639–1667, June 2002.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Number of quantized ICs - $v$ | 1 | 2 | 3 | 4 | 5 | 6 | 7-8 |
| $\delta_v$ | 1.6 | 1.15 | 0.9 | 0.8 | 0.7 | 0.65 | 0.55 |
| Number of quantized ICs - $v$ | 9 | 10-11 | 12-13 | 14-18 | 19-24 | 25-30 | |
| $\delta_v$ | .5 | .45 | .4 | .35 | .3 | .25 | |

Table 3: ICA Fra algorithm quantization step.

| STD | ICA Sel | ICA Ne | ICA Map | DCT | DCTX | DWT |
|---|---|---|---|---|---|---|
| 10 | 0.0206 | 0.0118 | 0.0098 | 0.0125 | 0.0123 | 0.3308 |
| 15 | 0.1064 | 0.0929 | 0.0750 | 0.0955 | 0.0957 | 0.3919 |
| 20 | 0.2165 | 0.2079 | 0.1641 | 0.2112 | 0.2110 | 0.4277 |
| 25 | 0.3130 | 0.3107 | 0.2414 | 0.3149 | 0.3150 | 0.4506 |
| 30 | 0.3832 | 0.3895 | 0.3009 | 0.3919 | 0.3924 | 0.4653 |
| MSE | 3.3714 | 3.1249 | 3.1249 | 2.9285 | 3.1556 | 2.1830 |
| PSNR | 42.5588 | 42.8884 | 42.8884 | 43.1656 | 42.8466 | 45.6137 |

Table 4: Gaussian noise attack on watermarking schemes.

| QL | ICA Sel | ICA Ne | ICA Map | DCT | DCTX | DWT |
|---|---|---|---|---|---|---|
| 10 | 0.2594 | 0.3785 | 0.5548 | 0.3373 | 0.2979 | 0.4125 |
| 20 | 0.0769 | 0.2593 | 0.2708 | 0.1705 | 0.0611 | 0.3157 |
| 30 | 0.0130 | 0.0072 | 0.0071 | 0.1306 | 0.0052 | 0.2559 |
| 50 | 0.0001 | 0.0007 | 0.0007 | 0.0909 | 0.0000 | 0.1947 |
| 90 | 0 | 0 | 0 | 0.0000 | 0 | 0.1136 |
| MSE | 3.3680 | 3.2757 | 3.2757 | 2.9326 | 2.0303 | 2.1941 |
| PSNR | 42.5633 | 42.7187 | 42.7187 | 43.1596 | 44.7611 | 45.6032 |

Table 5: JPEG attack on watermarking schemes.

Final Committee. JPEG2000 part 1 draft version 1.0. tech. rep. FCD15444-1, ISO/IEC, March 2000.

I.J. Cox, J. Kilian, T. Leighton, and T. Shamoon. Secure spread spectrum watermarking for multimedia. *IEEE Transactions on Image Processing*, 6(12):1673–1687, December 1997.

I.J. Cox, M.L. Miller, and J.A. Bloom. *Digital Watermarking*. Morgan Kaufmann Publishers, 2002.

S. Craver. Can invisible watermarks resolve rightful ownership? Technical Report RC 20509, IBM Research Report, July 1996.

| Rate | ICA Sel | ICA Ne | ICA Map | DCT | DCTX | DWT |
|---|---|---|---|---|---|---|
| 0.40 | 0.1812 | 0.1613 | 0.2138 | 0.1942 | 0.1558 | 0.3725 |
| 0.56 | 0.1065 | 0.0823 | 0.1424 | 0.1362 | 0.0738 | 0.3106 |
| 0.80 | 0.0350 | 0.0298 | 0.0931 | 0.0541 | 0.0210 | 0.2477 |
| 1.28 | 0.0049 | 0.0018 | 0.0381 | 0.0080 | 0.0029 | 0.1787 |
| 2.80 | 0 | 0 | 0.0026 | 0 | 0 | 0.1180 |
| MSE | 3.3702 | 3.4640 | 3.4640 | 2.9284 | 2.0279 | 2.1836 |
| PSNR | 42.5602 | 42.4620 | 42.4620 | 43.1658 | 44.7665 | 45.6141 |

Table 6: SPIHT attack on watermarking schemes.

F.J. González-Serrano, H.Y. Molina-Bulla, and J.J. Murillo-Fuentes. Independent component analysis applied to digital watermarking. In *International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, volume 3, pages 1997–2000, May 2001.

A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley-Interscience, 2001.

A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–1492, 1997.

A.K. Jain. *Fundamentals of Digital Image Processing*. Pearson Higher Education, 1989.

E. Koch and J. Zhao. Towards robust and hidden image copyright labelling. *IEEE Workshop on Nonlinear Signal and Image Processing*, pages 452–455, October 1995.

D. Kundur and D. Hatzinakos. Digital watermarking using multiresolution wavelet decomposition. In *International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, volume 5, pages 2969–2972, May 1998.

P. Meerwald. *Digital Image Watermarking in the Wavelet Transform Domain*. PhD thesis, University Salzburg, January 2001.

P. Moulin and J.A. O'Sullivan. Information-theoretic analysis of information hiding. *IEEE Transactions on Information Theory*, 49(3):563–593, March 2003.

F.A.P. Petitcolas and M.G. Kuhn. Stirmark 4.0. Available electronically from `http://www.cl.cam.ac.uk/~fapp2/watermarking/stirmark/`, 2002.

M. Ramkumar. *Data Hiding in Multimedia - Theory and Applications*. PhD thesis, New Jersey Institute of Technology, January 2000.

J.J.K. Ó Ruanaidh and T. Pun. Rotation, scale and translation invariant digital image watermarking. In *Proceedings of the International Conference on Image Processing*, pages 536–539, October 1997.

A. Said and W.A. Pearlman. A new fast and efficient image codec based on set partitioning in hierarchical trees. *IEEE Transactions on Ciruits and Systems for Video Technology*, 6:243–250, June 1996.

B. Toch, D. Lowe, and D. Saad. Watermarking of audio signals using independent component analysis. In *International conference on WEB delivering of music*, pages 71–74, September 2003.

G.K. Wallace. The JPEG still picture compression standard. *IEEE Transactions on Consumer Electronics*, 38(1):18–34, February 1992.